Statistica Sinica Preprint No: SS-2017-0397	
Title	SEMIPARAMETRIC REGRESSION MODEL FOR RECURRENT BACTERIAL INFECTIONS AFTER HEMATOPOIETIC STEM CELL TRANSPLANTATION
Manuscript ID	SS-2017-0397
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0397
Complete List of Authors	Chi Hyun Lee Chiung-Yu Huang Todd E. DeFor Claudio G. Brunstein Daniel J. Weisdorf and Xianghua Luo
Corresponding Author	Chi Hyun Lee
E-mail	clee9@mdanderson.org

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

Statistica Sinica

SEMIPARAMETRIC REGRESSION MODEL FOR RECURRENT BACTERIAL INFECTIONS AFTER HEMATOPOIETIC STEM CELL TRANSPLANTATION

Chi Hyun Lee[1], Chiung-Yu Huang[2], Todd E. DeFor[3], Claudio G. Brunstein[3]

Daniel J. Weisdorf[3] and Xianghua Luo[3]

[1] *University of Massachusetts, Amherst,*

[2] *University of California, San Francisco and* [3] *University of Minnesota*

Abstract: Patients who undergo hematopoietic stem cell transplantation (HSCT) often experience multiple bacterial infections during the early post-transplant period. In this article, we consider a semiparametric regression model that correlates patient- and transplant-related risk factors with inter-infection gap times. Existing regression methods for recurrent gap times are not directly applicable to studies of post-transplant infections because the initiating event (i.e., the transplant) is different to the recurrent events of interest (i.e., post-transplant infections). As a result, the time between a transplant and the first infection and that between consecutive infections have distinct biological meanings and, hence, follow different distributions. Moreover, risk factors may have different effects on these two types of gap times. Therefore, we propose a semiparametric estimation procedure that lets us simultaneously evaluate the covariate effects on the time between a transplant and the first infection and on

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

the gap times between consecutive infections. The proposed estimator accounts for dependent censoring induced by within-subject correlation between recurrent gap times and length bias in the last censored gap time due to intercept sampling. We study the finite sample properties through simulations and apply the proposed method to post-HSCT bacterial infection data collected at the University of Minnesota.

Key words and phrases: Accelerated failure time model, gap times, recurrent events, semiparametric method, weighted risk-set method.

1. Introduction

Infections after hematopoietic stem cell transplantation (HSCT) are often a major source of mortality and morbidity among transplant patients. During the early post-transplant period, bacterial infections are the predominant type of infection. Hence, characterizing the underlying early bacterial infection process and identifying the risk factors are of primary interest in clinical practice. Our motivating data were taken from 516 patients who received their first HSCT using unrelated umbilical cord blood (UCB) as the graft source between 2000 and 2010 at the University of Minnesota. Transplant patients were then followed prospectively, with infectious events recorded until the occurrence of a disease relapse, a second transplant, death, or loss of follow-up. It is well known that the greatest risk of infection in patients who undergo HSCT occurs prior to the engraftment of donor blood cells. Engraftment, especially neutrophil cell engraftment, which is crucial to fighting against bacterial infections, may

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

require as long as 42 days after a transplant. Thus, in our analysis, we focus on bacterial infections observed within 42 days of a transplant. The goal of this research 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 the time from a transplant to the first bacterial infection and on the interoccurrence times (i.e., gap times) from one bacterial infection to the next recurrent infection.

As pointed out by Wang and Chang (1999), analyses of recurrent gap time data can be challenging because of the unique sequential structure of the data. In particular, gap times beyond the first event time are subject to dependent censoring induced by the correlations between gap times of the same subject, even when the overall censoring time is independent of the recurrent event process. Moreover, it is noteworthy that the last censored gap times tend to be longer than the completely observed gap times, owing to intercept sampling. As a result, conventional regression methods for univariate time-to-event data or multivariate clustered survival data are not directly applicable to recurrent gap time data. In the literature, regression methods for recurrent gap time data have been developed by modeling either the hazard functions of gap times (Huang and Chen, 2003; Sun, Park, and Sun, 2006) or the (transformed) gap times directly (Chang, 2004, referred to as "Chang's method" hereinafter; Lu,

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

2005; Strawderman, 2005). More recently, quantile regressions have been studied for recurrent gap time data to account for data heteroscedasticity (Luo, Huang, and Wang, 2013). However, these methods assume that all events, including the initiating event, are of the same type, and that all gap times, including the time to the first occurrence of the recurrent events, have the same marginal distribution. As a result, applying these methods to study post-transplant infections can lead to incorrect inferential results because the time from the initiating event (i.e., the transplant) to the first infection and the gap times between recurrent infections have different clinical implications. Recently, Lee *et al.* (2016) considered a nonparametric estimation of the joint distribution of the time from a transplant to the first infection and the gap times between consecutive infections. To the best of our knowledge, no regression methods have been developed for recurrent gap time data under the setting described above.

In this paper, we propose a semiparametric regression model that allows the time from a transplant to the first infection and the time that elapses between consecutive infections to have distinct baseline distributions and different degrees of association with the covariates. In particular, we assume that the covariate effects are linearly related to the first event time and the gap times on a logarithmic scale, and that the within-subject correlation can be character-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

ized by a subject-specific random variable (i.e., frailty). The proposed model is similar in form to the accelerated failure time (AFT) model for univariate survival data (Kalbfleisch and Prentice, 2002, Chapter 7, and references therein), which is more attractive than the hazard-based regression models owing to its direct interpretation of the covariate effects on survival time. Moreover, the distribution of the frailty is left unspecified, which distinguishes the proposed approach from parametric frailty models (Liu, Wolfe and Huang, 2004; Huang and Liu, 2007; Zeng and Lin, 2008).

The proposed estimation procedure is motivated by the regression method for multistate data developed by Huang (2002, referred to as "Huang's method" hereinafter). Note that by restricting the analysis to data up to the second infection, Huang's method can be applied directly to study the covariate effects on bivariate gap times. However, this approach inevitably leads to a loss of information because patients can experience more than two infections during the course of follow-up. In our data example, some patients experienced as many as six infections. Moreover, the number of infections is informative about the distribution of the gap times. It is likely that patients with a higher risk of infection experience more infections, and thus have shorter gap times. To make better use of the observed data, we extend Huang's method by applying the weighted risk-set method discussed in Luo and Huang (2011) to the gap

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

times beyond the first infection, making use of the exchangeability between the uncensored gap times.

The remainder of this article is organized as follows. In Section 2, we first describe Huang's method, after adapting it for the simplified bivariate gap time data, and then propose an estimation method for the recurrent infection data. In Section 3, we investigate the performance of the proposed method by conducting a series of simulation studies. In Section 4, we apply the proposed method to the post-HSCT bacterial infection data collected at the University of Minnesota. Concluding remarks are presented in Section 5.

2. Methods

2.1. Model Setup

We first introduce the notation required to describe the recurrent infection process after a transplant. Let X^0 denote the time from a transplant to the first infection and let Y_j^0, $j = 1, 2, \ldots$, denote the gap times between two consecutive infections. The collection of all gap times for subject i, $i = 1, \ldots, n$, is denoted as $N_i = \{X_i^0, Y_{ij}^0, j = 1, 2, \ldots\}$, in the absence of censoring. 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 a transplant to the first infection and the log-transformed gap times from one infection to the next are linearly

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

related to the covariates, as follows:

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

respectively, 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 identically and independently distributed (i.i.d.) random errors from an unspecified continuous distribution. The latent vector $(\gamma_{i0}, \gamma_{i1})$, which can be continuous or discrete, is used to account for the heterogeneity among patients and the correlation between gap times within the same subject. The distribution of the latent vector is left unspecified, but is required to have a finite second moment. As a result, the joint distribution of $(X_i^0, Y_{i1}^0, Y_{i2}^0, \ldots)$ is not completely specified. Thus, the proposed model is a semiparametric model rather than a fully parametric model.

Let C_i be the censoring time from a transplant, whose survival function is $G(t) = \Pr(C_i > t)$, with a maximum support $\tau_C < \infty$. We denote the number of observed infections before time C_i by m_i. The random variable m_i is finite and satisfies $\Pr(m_i > 1) > 0$. When $m_i = 0$, $X_i^0 > C_i$; when $m_i = 1$,

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

$X_i^0 \leq C_i$ and $X_i^0 + Y_{i1}^0 > C_i$; and when $m_i > 1$, $X_i^0 + \sum_{j=1}^{m_i-1} Y_{ij}^0 \leq C_i$ 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 $\mathbf{A}_i$. In practice, however, this random censoring condition may be a strong assumption. Extensions of the proposed estimation procedure to handle conditional independent censoring are discussed in Section 2.3.

In our analysis of post-HSCT infections, we focus on early-stage bacterial infections within 42 days of a transplant. We do not expect to see a trend in such a short follow-up period, in general; in other words, we expect the exchangeability condition on the gap times between consecutive infections to hold, approximately. As shown in Section 2.3, the exchangeability condition is crucial to the development of the proposed estimation procedure.

2.2. Existing Method for Bivariate Gap Time Data

To evaluate the covariate effects on the time from a transplant to the first infection and on the gap times from one infection to the next, we can apply Huang's method by fixing the number of states to two. In this section, we adapt Huang's method for bivariate gap time data.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

defined as

$$Z^0_{ii'0}(\mathbf{b}_0) = \exp(\mathbf{A}^T_{ii'}\mathbf{b}_0)X^0_i$$
$$Z^0_{ii'1}(\mathbf{b}) = \exp(\mathbf{A}^T_{ii'}\mathbf{b}_0)X^0_i + \exp(\mathbf{A}^T_{ii'}\mathbf{b}_1)Y^0_{i1},$$

respectively, where $\mathbf{b} = (\mathbf{b}_0^T, \mathbf{b}_1^T)^T$ for $i, i' = 1, \ldots, n$. Given $\mathbf{A}_i$ and $\mathbf{A}_{i'}$, it follows that $Z^0_{ii'0}(\mathbf{b}_0)$ shares the same distribution with $Z^0_{i'0}$, and $Z^0_{ii'1}(\mathbf{b})$ shares the same distribution with $Z^0_{i'1}$ when $\mathbf{b}_0 = \boldsymbol{\beta}_0$ and $\mathbf{b}_1 = \boldsymbol{\beta}_1$ under the model assumption. By constructing the transformed time to the second infection as the sum of two transformed gap times X^0_i and Y^0_{i1}, 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^0_{i0} and Z^0_{i1}. Although the goal is to assess the covariate effects on the lengths of the interoccurrence times between events, the introduction of the time-to-event notation is necessary to properly address the problem of induced dependent censoring on gap times after the first infection. Now, consider the bivariate vectors $\{Z^0_{i0}, Z^0_{ii'0}(\mathbf{b}_0)\}$ and $\{Z^0_{i1}, Z^0_{ii'1}(\mathbf{b})\}$. It is obvious that, given $\mathbf{A}_i$ and $\mathbf{A}_{i'}$, $\{Z^0_{i0}, Z^0_{ii'0}(\boldsymbol{\beta}_0)\}$ has the same distribution as $\{Z^0_{i'i0}(\boldsymbol{\beta}_0), Z^0_{i'0}\}$, denoted by $\{Z^0_{i0}, Z^0_{ii'0}(\boldsymbol{\beta}_0)\} \sim \{Z^0_{i'i0}(\boldsymbol{\beta}_0), Z^0_{i'0}\}$, and that $\{Z^0_{i1}, Z^0_{ii'1}(\boldsymbol{\beta})\} \sim \{Z^0_{i'i1}(\boldsymbol{\beta}), Z^0_{i'1}\}$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T)^T$. Let $O_L(\cdot, \cdot)$ denote a symmetric and continuous scalar function, such that $O_L(t, s) = O_L(s, t)$. We set $O_L(t, s) = 0$ if $t \vee s \geq L$,

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

where $a \vee b = \max(a, b)$, for $L < \tau_C$. Then, it 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 fixed $\mathbf{b}$.

Let the observed times from a 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 $\Delta_{i1} = I(Z^0_{i1} \leq C_i)$, respectively, 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

$$\begin{aligned} Z_{ii'0}(\mathbf{b}_0) &= \exp(\mathbf{A}^T_{ii'}\mathbf{b}_0)X_i, \\ Z_{ii'1}(\mathbf{b}) &= \exp(\mathbf{A}^T_{ii'}\mathbf{b}_0)X_i + \exp(\mathbf{A}^T_{ii'}\mathbf{b}_1)Y_{i1}, \end{aligned} \tag{2.1}$$

respectively. Recall that $G(t)$ is the survival function for the censoring time.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

Then, under the random censoring assumption, we can easily show that

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

Therefore, it follows that

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

Then, we obtain the following estimating functions, which are in the form of U-statistics:

$$\mathbf{D}_0(\mathbf{b}_0) = n^{-2}\sum_{i=1}^{n}\sum_{i'=1}^{n} w(\mathbf{A}_i, \mathbf{A}_{i'}, \mathbf{b}_0)\mathbf{A}_{ii'}\frac{\Delta_{i0}O_{L_0}\{Z_{i0}, Z_{ii'0}(\mathbf{b}_0)\}}{\hat{G}_0(Z_{i0}\wedge L_0)}, \tag{2.2}$$
$$\mathbf{D}_1(\mathbf{b}) = n^{-2}\sum_{i=1}^{n}\sum_{i'=1}^{n} w(\mathbf{A}_i, \mathbf{A}_{i'}, \mathbf{b}_1)\mathbf{A}_{ii'}\frac{\Delta_{i1}O_{L_1}\{Z_{i1}, Z_{ii'1}(\mathbf{b})\}}{\hat{G}_1(Z_{i1}\wedge L_1)}, \tag{2.3}$$

where $\hat{G}_0(t)$ and $\hat{G}_1(t)$ are the Kaplan–Meier estimators of the censoring time survival function $G(t)$ using the data $\{(Z_{i0}, 1-\Delta_{i0}), i = 1, \ldots, n\}$ and $\{(Z_{i1}, 1-\Delta_{i1}), i = 1, \ldots, n\}$, respectively. The artificial limits L_0 and L_1 are imposed to handle the case in which Z^0_{i0} and Z^0_{i1} have larger maximum support than

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

τ_C. Note that subjects whose first or second infection times are censored contribute only to the denominator in functions (2.2) and (2.3) for the estimation of the censoring time survival function. To obtain the estimator of $\boldsymbol{\beta}_0$, we solve $\mathbf{D}_0(\mathbf{b}_0) = 0$, denoting the solution as $\hat{\boldsymbol{\beta}}_0$. Then, we solve $\mathbf{D}_1\{(\hat{\boldsymbol{\beta}}_0^T, \mathbf{b}_1^T)^T\} = 0$ to derive the estimator of $\boldsymbol{\beta}_1$. We denote the resulting estimator of $\boldsymbol{\beta}$, derived using Huang's method, as $\bar{\boldsymbol{\beta}}$.

As discussed in Huang (2002), the log-rank estimation approaches of the univariate AFT model can be applied directly to the data to estimate $\boldsymbol{\beta}_0$. However, such approaches cannot be used to estimate $\boldsymbol{\beta}_1$ when the association between the first event time and the gap time between the first and second infections cannot be completely characterized by the observed covariates. The gap time between consecutive infections is subject to informative censoring induced by within-subject correlation. The estimating equations based on the U-statistic functions in (2.2) and (2.3) properly address this issue.

2.3. Proposed Method for Post-Transplant Recurrent Infection Data

As mentioned earlier, applying Huang's method for multistate data to our recurrent infection data by ignoring the data beyond the second infection will inevitably lead to a loss of information. Moreover, the number of infections, m_i, is informative about the gap time distribution. Thus, we extend Huang's method for bivariate gap time data described in Section 2.2 by applying the weighted

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

risk-set technique discussed in Luo and Huang (2011). Luo and Huang (2011) demonstrated that the weighted risk-set method can be used to pool exchangeable gap times within a subject to improve the efficiency of the estimation. The weighted risk-set technique was used in the one-sample estimation method for post-transplant recurrent infection data by Lee *et al.* (2016). We apply the technique to our proposed regression method in a similar fashion. To proceed, we define $m_i^* = m_i - 1$ for $m_i \geq 2$, and $m_i^* = 1$ for $m_i < 2$. Then, we denote the observed uncensored gap times beyond the second infection by $Y_{ij} = Y_{ij}^0$ for $j = 2, \ldots, m_i^*$, where $m_i > 2$. Obviously, we have $\Delta_{ij} = 1$ for $j = 2, \ldots, m_i^*$. Under the assumptions in Section 2.1, 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 $\mathbf{A}_i$. Therefore, it follows that the observed uncensored gap time pairs, $(X_i, Y_{ij}), j = 1, \ldots, m_i^*$, are also conditionally i.i.d. Thus, the exchangeability between the observed uncensored gap time pairs follows. 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 (2.1) 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 $j = 1, \ldots, m_i^*$. As

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

a result, we can prove that

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

Hence, we propose replacing the estimating equation based on (2.3) with the following estimating function for the estimation of $\boldsymbol{\beta}_1$:

$$\mathbf{D}_1^*(\mathbf{b})=n^{-2}\sum_{i=1}^{n}\sum_{i'=1}^{n}w(\mathbf{A}_i,\mathbf{A}_{i'},\mathbf{b}_1)\mathbf{A}_{ii'}\left[\frac{1}{m_i^*}\sum_{j=1}^{m_i^*}\frac{\Delta_{ij}O_{L_1}\{Z_{ij},Z_{ii'j}(\mathbf{b})\}}{\hat{G}_1(Z_{ij}\wedge L_1)}\right]. \tag{2.4}$$

The estimator $\hat{\boldsymbol{\beta}}_1$ is derived by solving $\mathbf{D}_1^*\{(\hat{\boldsymbol{\beta}}_0^T,\mathbf{b}_1^T)^T\}=0$, where $\hat{\boldsymbol{\beta}}_0$ is derived in the same way as in the existing method discussed in Section 2.2. As discussed earlier, the last censored recurrent gap times are usually longer than the uncensored gap times, owing to intercept sampling. Thus, to avoid bias, the last censored gap times of subjects with $m_i\geq 2$ are not used in Equation (2.4). Under the regularity conditions (C1)–(C3) listed in Web Appendix A.1, $n^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ is asymptotically normal, with mean zero and variance $\Sigma^{-1}\Omega(\Sigma^{-1})^T$, which can be estimated consistently by $\hat{\Sigma}^{-1}\hat{\Omega}(\hat{\Sigma}^{-1})^T$. The definitions of Σ, Ω, $\hat{\Sigma}$, and $\hat{\Omega}$ and the detailed proofs can be found in Web Appendices A.2–A.5.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

Compared with $\mathbf{D}_1(\mathbf{b})$ in (2.3), we use additional uncensored recurrent gap times beyond the second infection in the construction of (2.4). Hence, the proposed estimation method is expected to provide a more efficient estimation on $\boldsymbol{\beta}_1$ than when applying Huang's method to data up to the second infection. We show the efficiency gain of using the proposed estimator over Huang's method in Web Appendix A.6. 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 equations become monotone and a unique solution is attainable. Other choices for O_L and w are discussed in Huang (2002).

Note that the estimating functions (2.2) and (2.3), and the proposed estimating function (2.4) are all constructed based on the random censoring assumption. We can relax this assumption by allowing the censoring to depend on the covariates and to be conditionally independent of the gap time distribution, given $\mathbf{A}$. As pointed out by Huang (2002), the estimators of the censoring time survival function, $\hat{G}_0(t)$ and $\hat{G}_1(t)$, in the estimating functions can be replaced with consistent estimators of the conditional survival function, $G(t \mid \mathbf{A})$. If the covariates have finite numbers of values, such as the treatment arms in randomized trials, $G(t \mid \mathbf{A})$ can be estimated nonparametrically by the covariate-specific Kaplan–Meier estimator $\hat{G}_j(t \mid \mathbf{A})$, using data $(Z_{ij}, 1 - \Delta_{ij})$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

for i such that $\mathbf{A}_i = \mathbf{A}$ and $j = 0, 1$. When $\mathbf{A}$ involves continuous covariates, we can postulate a semiparametric regression model, such as the proportional hazards model, for the censoring distribution. Note that modeling the censoring mechanism may not be robust. As an alternative, we can adopt the local Kaplan–Meier estimator to estimate $G(t \mid \mathbf{A})$ nonparametrically (Dabrowska, 1989; Wang and Wang, 2009). Under the conditional independent censoring assumption, we have

$$\mathrm{E}\left[\frac{\Delta_{i0} O_{L_0}\{Z_{i0}, Z_{ii'0}(\boldsymbol{\beta}_0)\}}{G(Z_{i0} \wedge L_0 \mid \mathbf{A}_i)}\middle| \mathbf{A}_i, \mathbf{A}_{i'}\right] = \mathrm{E}\left[O_{L_0}\{Z^0_{i0}, Z^0_{ii'0}(\boldsymbol{\beta}_0)\}\middle| \mathbf{A}_i, \mathbf{A}_{i'}\right]$$

and

$$\mathrm{E}\left[\frac{1}{m_i^*}\sum_{j=1}^{m_i^*} \mathrm{E}\left[\frac{\Delta_{ij} O_{L_1}\{Z_{ij}, Z_{ii'j}(\boldsymbol{\beta})\}}{G(Z_{ij} \wedge L_1 \mid \mathbf{A}_i)}\middle| m_i, (\gamma_{i0}, \gamma_{i1}), \mathbf{A}_i, \mathbf{A}_{i'}\right]\right]$$
$$= \mathrm{E}\left[O_{L_1}\{Z^0_{i1}, Z^0_{ii'1}(\boldsymbol{\beta})\}\middle| \mathbf{A}_i, \mathbf{A}_{i'}\right].$$

Thus, following the same spirit of (2.2) and (2.4), we obtain the following

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

estimating functions:

$$\mathbf{D}_0^c(\mathbf{b}_0) = n^{-2}\sum_{i=1}^{n}\sum_{i'=1}^{n} w(\mathbf{A}_i, \mathbf{A}_{i'}, \mathbf{b}_0)\mathbf{A}_{ii'}\frac{\Delta_{i0}O_{L_0}\{Z_{i0}, Z_{ii'0}(\mathbf{b}_0)\}}{\hat{G}_0(Z_{i0}\wedge L_0 \mid \mathbf{A}_i)}, \tag{2.5}$$

$$\mathbf{D}_1^{c*}(\mathbf{b}) = n^{-2}\sum_{i=1}^{n}\sum_{i'=1}^{n} w(\mathbf{A}_i, \mathbf{A}_{i'}, \mathbf{b}_1)\mathbf{A}_{ii'}\left[\frac{1}{m_i^*}\sum_{j=1}^{m_i^*}\frac{\Delta_{ij}O_{L_1}\{Z_{ij}, Z_{ii'j}(\mathbf{b})\}}{\hat{G}_1(Z_{ij}\wedge L_1 \mid \mathbf{A}_i)}\right], \tag{2.6}$$

where $\hat{G}_j(t \mid \mathbf{A}), j = 0, 1$, are consistent estimators of $G(t \mid \mathbf{A})$. We denote the solution to $\mathbf{D}_0^c(\mathbf{b}_0) = \mathbf{0}$ as $\tilde{\boldsymbol{\beta}}_0$. The estimator $\tilde{\boldsymbol{\beta}}_1$ is obtained by solving $\mathbf{D}_1^{c*}\{(\tilde{\boldsymbol{\beta}}_0^T, \mathbf{b}_1^T)^T\} = \mathbf{0}$. Let $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_0^T, \tilde{\boldsymbol{\beta}}_1^T)^T$. In Web Appendix B, we provide proofs of the asymptotic properties of $\tilde{\boldsymbol{\beta}}$ under the conditional independent censoring assumption when the covariate-specific Kaplan–Meier estimator is used for the estimation of $G(t \mid \mathbf{A})$. Similar techniques can be used to establish the asymptotic properties when a semiparametric regression model is used to estimate $G(t \mid \mathbf{A})$.

3. Simulation Studies

We conducted a series of simulation studies to evaluate the performance of the proposed method, each with 1000 data sets and $n = 150$ and 300 subjects per data set. We generated the time to the first infection and the gap times

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

between two consecutive infections for each subject using the following model:

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

respectively, where $\mathbf{A}_i = (A_{i1}, A_{i2})^T$, with A_{i1} sampled from a Bernoulli distribution with probability 0.5 and A_{i2} sampled from a uniform distribution $(0, 1)$. The true covariate effects are $\boldsymbol{\beta}_0 = (-0.5, 0.5)^T$ and $\boldsymbol{\beta}_1 = (0.5, 0.5)^T$. We generated the mutually independent error terms ϵ_{ij} from a normal distribution with mean zero and variance 0.25, and the subject-specific latent vector $(\gamma_{i0}, \gamma_{i1})$ from a bivariate normal distribution with a unit mean and the variance–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, $\log(Y_{ij}^0)$. In addition, 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 $\sigma_0^2 = 0.5$, and consider $\sigma_1^2 = 0.1$ or 0.5 and $\rho = 0$ or

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

0.5 in different scenarios. The censoring time $C_i, i = 1, \ldots, n$, is sampled independently from a uniform distribution $(0, U)$, where $U = 10, 30$, or 50. The average number of infections observed per subject ($\bar{m}$) increases with U.

We apply the proposed method to the simulated data, and 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 purposes, we also applied Huang's method and Chang's method. The simulation results are summarized in Tables 1, 2, and 3 for varying ranges of censoring times. The proposed method and Huang's method are virtually unbiased across all settings. The empirical standard deviations and the standard errors are close to each other, and the coverage probabilities are reasonably close to the nominal level. Note that the two methods share the same estimator for the covariate effects on the time to the first infection ($\boldsymbol{\beta}_0$). However, the proposed method yields more efficient results than Huang's method does when estimating the covariate effects on the gap times after the first infection ($\boldsymbol{\beta}_1$) in all settings. The efficiency of the proposed method relative to Huang's method increases as additional recurrent infections are observed per subject (i.e., as $\bar{m}$ increases in Tables 1 to 3).

As expected, biased results are obtained from Chang's method, which assumes that all gap times, including the time from a transplant to the first infection, are equally distributed. Specifically, the method fails to capture the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

different effects of covariate A_1 on the two different types of time variables, namely, the time from a transplant to the first infection, and the gap times between recurrent infections (-0.5 and 0.5, respectively), in the simulated data. Under the simulation setting, covariate $A_1 = 1$ is associated with a shorter time from a transplant to the first infection, but a prolonged gap time from one infection to the next. Chang's method ignores this distinction, which diminishes the estimated "overall effect" of A_1. The covariate effect of A_2 is set to be the same for the two types of time variables (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. This suggests that if one of the covariates in Chang's method has differential effects on the two types of gap times, the estimation of the effect of other covariates, which do not have differential effects, would also be affected.

We also conducted simulation studies to assess the performance of the proposed method under conditional independent censoring. Here, we considered a setting where the two covariates A_1 and A_2 are generated independently from a Bernoulli distribution with probability 0.5. The censoring time is generated from a uniform distribution (0, 20) if $A_1 = 1$, and from a uniform distribution (0, 30) if $A_1 = 0$. The results are shown in Table 4, and are similar to those under the random censoring condition.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

Table 1: Summary of simulation results for censoring time $C \sim Unif(0, 10)$: The table displays the true coefficients (True); the mean of the point estimates (Est); the Monte Carlo standard deviation of the point estimates (SD); the mean of the standard error estimates (SE); and the coverage probability (CP) of the 95% confidence intervals for the proposed method (Proposed), Huang's method (Huang), and Chang's method (Chang).

				$\boldsymbol{\beta}_0$		$\boldsymbol{\beta}_1$					
				Proposed/Huang		Proposed		Huang		Chang	
				A_1	A_2	A_1	A_2	A_1	A_2	A_1	A_2
n	ρ	σ_1^2	True	-0.5	0.5	0.5	0.5	0.5	0.5	-[a]	-
				$\bar{m}$[b]= 1.31; cr_1[c]= 0.36; cr_2[d]= 0.77							
150	0	0.1	Est	-0.507	0.498	0.494	0.488	0.494	0.491	-0.398	0.415
			SD	0.172	0.315	0.190	0.342	0.191	0.347	0.153	0.269
			SE	0.175	0.303	0.197	0.350	0.200	0.354	0.152	0.263
			CP	0.953	0.930	0.950	0.935	0.952	0.936	-	-
				$\bar{m} = 1.42$; $cr_1 = 0.37$; $cr_2 = 0.76$							
		0.5	Est	-0.502	0.510	0.512	0.495	0.512	0.494	-0.405	0.462
			SD	0.177	0.303	0.270	0.493	0.276	0.500	0.161	0.283
			SE	0.175	0.304	0.266	0.463	0.270	0.470	0.160	0.278
			CP	0.950	0.949	0.943	0.913	0.939	0.920	-	-
				$\bar{m} = 1.35$; $cr_1 = 0.36$; $cr_2 = 0.75$							
	0.5	0.1	Est	-0.500	0.492	0.497	0.496	0.499	0.498	-0.376	0.411
			SD	0.178	0.318	0.194	0.341	0.198	0.345	0.148	0.282
			SE	0.175	0.304	0.194	0.344	0.198	0.350	0.151	0.259
			CP	0.939	0.928	0.950	0.940	0.943	0.945	-	-
				$\bar{m} = 1.52$; $cr_1 = 0.36$; $cr_2 = 0.74$							
		0.5	Est	-0.501	0.493	0.510	0.500	0.512	0.499	-0.387	0.442
			SD	0.173	0.311	0.259	0.477	0.264	0.481	0.157	0.273
			SE	0.175	0.302	0.257	0.451	0.262	0.457	0.158	0.272
			CP	0.950	0.944	0.931	0.920	0.931	0.922	-	-
				$\bar{m} = 1.31$; $cr_1 = 0.36$; $cr_2 = 0.77$							
300	0	0.1	Est	-0.500	0.504	0.500	0.501	0.500	0.501	-0.388	0.418
			SD	0.121	0.213	0.137	0.239	0.139	0.241	0.105	0.180
			SE	0.124	0.216	0.138	0.244	0.139	0.247	0.107	0.183
			CP	0.946	0.958	0.952	0.961	0.948	0.960	-	-
				$\bar{m} = 1.43$; $cr_1 = 0.36$; $cr_2 = 0.76$							
		0.5	Est	-0.494	0.501	0.491	0.500	0.492	0.500	-0.405	0.445
			SD	0.121	0.221	0.186	0.330	0.187	0.331	0.110	0.197
			SE	0.124	0.216	0.184	0.322	0.187	0.326	0.111	0.192
			CP	0.952	0.943	0.944	0.935	0.944	0.942	-	-
				$\bar{m} = 1.35$; $cr_1 = 0.36$; $cr_2 = 0.75$							
	0.5	0.1	Est	-0.501	0.503	0.494	0.501	0.495	0.503	-0.376	0.418
			SD	0.119	0.217	0.135	0.246	0.138	0.250	0.105	0.183
			SE	0.124	0.216	0.137	0.245	0.139	0.249	0.106	0.182
			CP	0.950	0.948	0.955	0.950	0.950	0.951	-	-
				$\bar{m} = 1.53$; $cr_1 = 0.36$; $cr_2 = 0.74$							
		0.5	Est	-0.499	0.490	0.491	0.518	0.492	0.520	-0.386	0.433
			SD	0.129	0.218	0.183	0.321	0.184	0.326	0.111	0.191
			SE	0.125	0.217	0.182	0.320	0.184	0.324	0.110	0.190
			CP	0.940	0.951	0.941	0.939	0.941	0.944	-	-

[a]-: True $\boldsymbol{\beta}$-values do not exist;
[b]$\bar{m}$: average number of observed infections per subject;
[c]cr_1: average proportion of subjects without any infections;
[d]cr_2: average proportion of subjects with the first or the second gap times censored.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

Table 2: Summary of simulation results for censoring time $C \sim Unif(0, 30)$: The table displays the true coefficients (True); the mean of the point estimates (Est); the Monte Carlo standard deviation of the point estimates (SD); the mean of the standard error estimates (SE); and the coverage probability (CP) of the 95% confidence intervals for the proposed method (Proposed), Huang's method (Huang), and Chang's method (Chang).

				$\boldsymbol{\beta}_0$		$\boldsymbol{\beta}_1$					
				Proposed/Huang		Proposed		Huang		Chang	
				A_1	A_2	A_1	A_2	A_1	A_2	A_1	A_2
n	ρ	σ_1^2	True	-0.5	0.5	0.5	0.5	0.5	0.5	-[a]	-
				$\bar{m}$[b]= 3.15; cr_1[c]= 0.14; cr_2[d]= 0.32							
150	0	0.1	Est	-0.504	0.498	0.491	0.484	0.493	0.484	-0.028	0.451
			SD	0.161	0.284	0.162	0.257	0.171	0.267	0.120	0.197
			SE	0.163	0.282	0.156	0.269	0.164	0.283	0.115	0.191
			CP	0.948	0.949	0.935	0.954	0.930	0.959	-	-
				$\bar{m} = 3.69$; $cr_1 = 0.14$; $cr_2 = 0.35$							
		0.5	Est	-0.501	0.510	0.497	0.475	0.498	0.474	-0.100	0.428
			SD	0.167	0.293	0.210	0.360	0.214	0.366	0.134	0.236
			SE	0.164	0.286	0.203	0.351	0.209	0.361	0.130	0.224
			CP	0.950	0.939	0.931	0.936	0.929	0.944	-	-
				$\bar{m} = 3.23$; $cr_1 = 0.14$; $cr_2 = 0.32$							
	0.5	0.1	Est	-0.497	0.504	0.484	0.496	0.487	0.497	-0.030	0.445
			SD	0.162	0.297	0.163	0.275	0.170	0.290	0.124	0.195
			SE	0.163	0.282	0.155	0.267	0.162	0.280	0.117	0.196
			CP	0.955	0.938	0.929	0.951	0.934	0.949	-	-
				$\bar{m} = 3.87$; $cr_1 = 0.14$; $cr_2 = 0.34$							
		0.5	Est	-0.505	0.499	0.490	0.485	0.488	0.484	-0.093	0.420
			SD	0.164	0.291	0.198	0.354	0.204	0.367	0.135	0.239
			SE	0.163	0.283	0.197	0.344	0.203	0.355	0.131	0.224
			CP	0.954	0.942	0.941	0.944	0.941	0.939	-	-
				$\bar{m}$= 3.15; cr_1= 0.14; cr_2= 0.32							
300	0	0.1	Est	-0.501	0.508	0.498	0.490	0.496	0.492	-0.034	0.446
			SD	0.110	0.213	0.115	0.204	0.119	0.210	0.082	0.148
			SE	0.117	0.204	0.115	0.199	0.120	0.207	0.083	0.139
			CP	0.967	0.951	0.946	0.941	0.950	0.936	-	-
				$\bar{m} = 3.71$; $cr_1 = 0.14$; $cr_2 = 0.35$							
		0.5	Est	-0.495	0.496	0.501	0.484	0.500	0.484	-0.105	0.417
			SD	0.117	0.213	0.142	0.260	0.147	0.263	0.092	0.160
			SE	0.117	0.204	0.146	0.253	0.150	0.259	0.092	0.158
			CP	0.947	0.941	0.944	0.936	0.949	0.945	-	-
				$\bar{m} = 3.23$; $cr_1 = 0.14$; $cr_2 = 0.32$							
	0.5	0.1	Est	-0.498	0.498	0.496	0.490	0.496	0.494	-0.032	0.434
			SD	0.115	0.211	0.112	0.197	0.117	0.204	0.085	0.148
			SE	0.117	0.204	0.114	0.198	0.119	0.206	0.084	0.141
			CP	0.953	0.949	0.951	0.949	0.952	0.943	-	-
				$\bar{m} = 3.90$; $cr_1 = 0.14$; $cr_2 = 0.35$							
		0.5	Est	-0.499	0.497	0.499	0.506	0.502	0.509	-0.094	0.412
			SD	0.117	0.205	0.140	0.245	0.144	0.252	0.092	0.167
			SE	0.117	0.204	0.143	0.248	0.147	0.255	0.094	0.160
			CP	0.954	0.945	0.949	0.956	0.954	0.955	-	-

[a]-: True $\boldsymbol{\beta}$-values do not exist;
[b]$\bar{m}$: average number of observed infections per subject;
[c]cr_1: average proportion of subjects without any infections;
[d]cr_2: average proportion of subjects with the first or the second gap times censored.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

Table 3: Summary of simulation results for censoring time $C \sim Unif(0, 50)$: The table displays the true coefficients (True); the mean of the point estimates (Est); the Monte Carlo standard deviation of the point estimates (SD); the mean of the standard error estimates (SE); and the coverage probability (CP) of the 95% confidence intervals for the proposed method (Proposed), Huang's method (Huang), and Chang's method (Chang).

				$\boldsymbol{\beta}_0$		$\boldsymbol{\beta}_1$					
				Proposed/Huang		Proposed		Huang		Chang	
				A_1	A_2	A_1	A_2	A_1	A_2	A_1	A_2
n	ρ	σ_1^2	True	-0.5	0.5	0.5	0.5	0.5	0.5	-[a]	-
				$\bar{m}$[b]= 5.25; cr_1[c]= 0.08; cr_2[d]= 0.19							
150	0	0.1	Est	-0.496	0.503	0.499	0.491	0.497	0.492	0.141	0.472
			SD	0.154	0.278	0.141	0.244	0.154	0.265	0.098	0.167
			SE	0.158	0.275	0.141	0.244	0.154	0.266	0.095	0.161
			CP	0.951	0.945	0.942	0.955	0.934	0.950	-	-
				$\bar{m} = 6.23$; $cr_1 = 0.08$; $cr_2 = 0.22$							
		0.5	Est	-0.500	0.511	0.502	0.475	0.500	0.475	0.067	0.417
			SD	0.155	0.278	0.195	0.333	0.204	0.348	0.124	0.211
			SE	0.158	0.275	0.191	0.331	0.199	0.346	0.120	0.207
			CP	0.951	0.950	0.948	0.938	0.943	0.939	-	-
				$\bar{m} = 5.32$; $cr_1 = 0.08$; $cr_2 = 0.19$							
	0.5	0.1	Est	-0.504	0.508	0.491	0.498	0.490	0.498	0.139	0.466
			SD	0.156	0.285	0.145	0.245	0.159	0.263	0.099	0.175
			SE	0.157	0.273	0.141	0.242	0.153	0.263	0.099	0.167
			CP	0.947	0.936	0.942	0.940	0.931	0.945	-	-
				$\bar{m} = 6.46$; $cr_1 = 0.08$; $cr_2 = 0.22$							
		0.5	Est	-0.498	0.503	0.498	0.470	0.500	0.468	0.068	0.426
			SD	0.153	0.281	0.191	0.332	0.198	0.349	0.125	0.228
			SE	0.158	0.274	0.186	0.323	0.195	0.338	0.124	0.214
			CP	0.951	0.943	0.933	0.933	0.936	0.930	-	-
				$\bar{m}$= 5.24; cr_1= 0.08; cr_2= 0.19							
300	0	0.1	Est	-0.499	0.499	0.502	0.491	0.502	0.491	0.138	0.467
			SD	0.110	0.201	0.107	0.176	0.114	0.189	0.069	0.122
			SE	0.113	0.197	0.106	0.184	0.114	0.197	0.069	0.117
			CP	0.956	0.953	0.942	0.957	0.942	0.964	-	-
				$\bar{m} = 6.26$; $cr_1 = 0.08$; $cr_2 = 0.22$							
		0.5	Est	-0.495	0.507	0.496	0.493	0.497	0.492	0.060	0.429
			SD	0.111	0.203	0.138	0.241	0.143	0.247	0.087	0.151
			SE	0.113	0.197	0.139	0.240	0.145	0.249	0.086	0.147
			CP	0.954	0.941	0.949	0.943	0.952	0.945	-	-
				$\bar{m} = 5.32$; $cr_1 = 0.08$; $cr_2 = 0.19$							
	0.5	0.1	Est	-0.501	0.499	0.501	0.489	0.502	0.494	0.137	0.463
			SD	0.109	0.205	0.105	0.181	0.115	0.196	0.070	0.127
			SE	0.113	0.197	0.106	0.183	0.114	0.196	0.071	0.120
			CP	0.958	0.936	0.951	0.952	0.948	0.950	-	-
				$\bar{m} = 6.46$; $cr_1 = 0.08$; $cr_2 = 0.21$							
		0.5	Est	-0.498	0.493	0.495	0.498	0.496	0.498	0.064	0.426
			SD	0.115	0.198	0.135	0.246	0.139	0.258	0.087	0.151
			SE	0.113	0.197	0.136	0.236	0.142	0.245	0.088	0.152
			CP	0.949	0.956	0.958	0.944	0.956	0.936	-	-

[a]-: True $\boldsymbol{\beta}$-values do not exist;
[b]$\bar{m}$: average number of observed infections per subject;
[c]cr_1: average proportion of subjects without any infections;
[d]cr_2: average proportion of subjects with the first or the second gap times censored.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

Table 4: Summary of simulation results under the conditional independent censoring assumption: The table displays the true coefficients (True); the mean of the point estimates (Est); the Monte Carlo standard deviation of the point estimates (SD); the mean of the standard error estimates (SE); and the coverage probability (CP) of the 95% confidence intervals for the proposed estimator (Proposed) and Huang's method (Huang).

				$\boldsymbol{\beta}_0$		$\boldsymbol{\beta}_1$			
				Proposed/Huang		Proposed		Huang	
				A_1	A_2	A_1	A_2	A_1	A_2
n	ρ	σ_1^2	True	-0.5	0.5	0.5	0.5	0.5	0.5
				$\bar{m}$[a]= 2.83; cr_1[b]= 0.16; cr_2[c]= 0.40					
150	0	0.1	Est	-0.497	0.504	0.502	0.494	0.504	0.495
			SD	0.155	0.160	0.125	0.127	0.145	0.143
			SE	0.158	0.159	0.139	0.139	0.153	0.154
			CP	0.953	0.951	0.968	0.958	0.959	0.958
				$\bar{m} = 3.32$; $cr_1 = 0.16$; $cr_2 = 0.42$					
		0.5	Est	-0.502	0.500	0.503	0.489	0.505	0.494
			SD	0.149	0.154	0.186	0.193	0.198	0.214
			SE	0.158	0.159	0.199	0.200	0.210	0.212
			CP	0.957	0.964	0.951	0.966	0.954	0.954
				$\bar{m} = 2.91$; $cr_1 = 0.16$; $cr_2 = 0.40$					
	0.5	0.1	Est	-0.495	0.499	0.503	0.493	0.504	0.495
			SD	0.153	0.155	0.123	0.129	0.142	0.147
			SE	0.158	0.159	0.138	0.139	0.152	0.153
			CP	0.952	0.951	0.976	0.962	0.965	0.953
				$\bar{m} = 3.48$; $cr_1 = 0.17$; $cr_2 = 0.42$					
		0.5	Est	-0.502	0.503	0.505	0.496	0.504	0.495
			SD	0.149	0.161	0.179	0.180	0.196	0.195
			SE	0.158	0.158	0.195	0.198	0.207	0.210
			CP	0.956	0.954	0.948	0.970	0.949	0.966
				$\bar{m} = 2.84$; $cr_1 = 0.17$; $cr_2 = 0.40$					
300	0	0.1	Est	-0.501	0.497	0.505	0.502	0.507	0.503
			SD	0.106	0.106	0.087	0.085	0.094	0.096
			SE	0.111	0.112	0.100	0.099	0.108	0.107
			CP	0.966	0.964	0.975	0.967	0.974	0.968
				$\bar{m} = 3.32$; $cr_1 = 0.16$; $cr_2 = 0.42$					
		0.5	Est	-0.497	0.500	0.504	0.497	0.505	0.498
			SD	0.104	0.106	0.128	0.136	0.134	0.147
			SE	0.112	0.112	0.141	0.141	0.148	0.148
			CP	0.957	0.961	0.965	0.957	0.965	0.949
				$\bar{m} = 2.90$; $cr_1 = 0.16$; $cr_2 = 0.40$					
	0.5	0.1	Est	-0.501	0.498	0.503	0.497	0.503	0.496
			SD	0.101	0.107	0.083	0.083	0.094	0.092
			SE	0.112	0.112	0.100	0.099	0.108	0.108
			CP	0.962	0.962	0.983	0.980	0.976	0.977
				$\bar{m} = 3.50$; $cr_1 = 0.16$; $cr_2 = 0.42$					
		0.5	Est	-0.497	0.502	0.497	0.498	0.497	0.498
			SD	0.105	0.110	0.117	0.131	0.126	0.138
			SE	0.111	0.112	0.138	0.139	0.145	0.146
			CP	0.958	0.950	0.977	0.959	0.980	0.958

[a] $\bar{m}$: average number of observed infections per subject;
[b] cr_1: average proportion of subjects without any infections;
[c] cr_2: average proportion of subjects with the first or the second gap times censored.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

4. Application

To illustrate the proposed estimation method, we analyze the post-HSCT bacterial infection data introduced in Section 1. The data were drawn from 516 HSCT recipients who used unrelated UCB as the graft source. Because 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 into two group at the time of transplant: pediatric patients ($<$ 18 years old, $n = 155$) and adult patients ($\geq$ 18 years old, $n = 361$). Patient- and transplant-related characteristics for the overall group, and for the pediatric and adult groups separately, are summarized in Web Table S1.

We focus on early-phase bacterial infections experienced within 42 days of a transplant. The follow-up for the recurrent infection process was terminated by the 42-day cutoff (89%), death (5%), relapse (4%), or a second transplant (2%) before day 42. Among the 25 deaths, only seven were related to infection; of these, three ($<$ 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 are defined 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 adults, during the first 42 days after a

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

transplant. On average, each patient experienced 0.77 infections, with children experiencing fewer infections than adults (0.55 vs. 0.86). A detailed summary of the infections can be found in Table 5. About 59% of pediatric patients and 48% of adult patients experienced no infections. Among all patients, about 81% (88% of child and 78% of adult patients) had the time from their transplants to the second infection censored. To ensure that the gap times after the first infection were similarly distributed, we conducted the trend test of Wang and Chen (2000) for each patient group. We found no evidence of a 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.

Table 5: Summary of number of patients who experienced k number of bacterial infections within 42 days of a transplant, $k = 0, 1, \cdots, 6$.

	No. patients	No. of infections observed for a patient						
Group	(%)	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)
Child	155	92	45	14	3	1	0	0
	(100)	(59.4)	(29.0)	(9.0)	(1.9)	(0.7)	(0.0)	(0.0)
Adult	361	174	107	55	12	9	2	2
	(100)	(48.2)	(29.6)	(15.2)	(3.3)	(2.5)	(0.6)	(0.6)

First, we conducted univariate regressions to identify potential risk factors. The regression parameters were estimated using the proposed method under the random censoring assumption. The estimated regression coefficients and the asymptotic standard error estimates are presented in Table 6 (upper

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

panel). We found that for pediatric patients, younger age, single donor (vs. double donors), and higher total nucleated cell (TNC) dose are significantly associated with a prolonged time to the first bacterial infection. However, a higher CD34 dose level was associated with shorter recurrent gap times between consecutive bacterial infections. For adult patients, older age was significantly associated with a longer time to the first bacterial infection. Furthermore, a non-myeloablative regimen without anti-thymocyte globulin (ATG), as compared to a myeloablative regimen, was significantly associated with both a longer time to the first infection and longer gap times between two consecutive infections. Other factors, including cytomegalovirus (CMV) serostatus, human leukocyte antigen (HLA) matching, and graft-versus-host disease (GVHD) prophylaxis, were not found to be associated with either type of time variable for either patient cohort.

Multivariable regressions were conducted using all covariates considered in the univariate analysis. The results are shown in the lower panel of Table 6. For pediatric patients, the single donor type and higher TNC dose remained significantly associated with the time to the first infection. However, age lost its significance. No factor showed significant association with the gap times after the first infection. The loss of significance of age may be due to the confounding effect of other factors, because we found that age was associated with both the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

Table 6: Summary of regression analysis of risk factors for early bacterial infections for children and adult patients: Estimated regression coefficients (standard error) are presented for the univariate regression in the upper panel and the multivariate regression in the lower panel.

	Children		Adults	
Variables	1st gap	≥ 2nd gap	1st gap	≥ 2nd gap
		Univariate Regression		
Age at Transplant (Years)	-0.09 (0.03*)	-0.05 (0.06)	0.03 (0.01*)	0.01 (0.01)
CMV Serostatus Positive vs. Negative	-0.49 (0.37)	0.67 (0.35)	-0.10 (0.23)	-0.01 (0.27)
Type of Transplant Double vs. Single	-0.81 (0.28*)	-0.08 (0.52)	-0.38 (0.41)	-0.49 (0.38)
Conditioning Regimen (vs. Myeloablative)				
Non-myeloablative w ATG	NI	NI	0.50 (0.32)	0.18 (0.46)
Non-myeloablative wo ATG	NI	NI	1.26 (0.23*)	0.85 (0.36*)
HLA Match Score 5-6/6 vs. 4/6	-0.74 (0.39)	-0.54 (0.78)	0.31 (0.26)	0.19 (0.29)
GVHD Prophylaxis CSA/MMF/MTX vs. Other	-0.74 (0.38)	0.34 (0.37)	-0.29 (0.61)	-1.32 (0.85)
CD34+ Dose Level High vs. Low	-0.01 (0.43)	-1.72 (0.69*)	0.05 (0.27)	-0.44 (0.37)
TNC Dose Level High vs. Low	1.02 (0.32*)	-0.29 (0.49)	-0.13 (0.27)	-0.19 (0.32)
		Multivariable Regression		
Age at Transplant (Years)	-0.03 (0.04)	-0.11 (0.06)	0.03 (0.01*)	-0.02 (0.02)
CMV Serostatus Positive vs. Negative	-0.36 (0.29)	0.77 (0.43)	0.01 (0.22)	-0.12 (0.26)
Type of Transplant Double vs. Single	-0.87 (0.39*)	-0.32 (0.66)	-0.21 (0.49)	-0.46 (0.63)
Conditioning Regimen (vs. Myeloablative)				
Non-myeloablative w ATG	NI	NI	-0.43 (0.45)	0.883 (0.54)
Non-myeloablative wo ATG	NI	NI	0.56 (0.30)	1.43 (0.43*)
HLA Match Score 5-6/6 vs. 4/6	-0.61 (0.38)	-1.06 (0.73)	0.16 (0.22)	0.11 (0.28)
GVHD Prophylaxis CSA/MMF/MTX vs. Other	-0.07 (0.43)	0.84 (0.88)	0.05 (0.82)	1.04 (1.58)
CD34+ Dose Level High vs. Low	-0.70 (0.43)	-1.32 (1.01)	0.14 (0.28)	-0.89 (0.42*)
TNC Dose Level High vs. Low	1.42 (0.45*)	-0.75 (0.76)	-0.16 (0.26)	0.14 (0.45)

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

number of donors and the TNC dose level among pediatric patients. Specifically, double UCB stem cells were used more frequently for older children, and older children tended to require a higher TNC dose than younger children did, based on our data. For adults, age remained a significant factor for the time to the first infection, but the effect of receiving a non-myeloablative regimen without ATG on the time to the first infection, compared with receiving a myeloablative regimen, became nonsignificant in the multivariable regression. In addition, the CD34 dose level became marginally significant for the gap times between infections for adult patients.

5. Concluding Remarks

In this article, we proposed a semiparametric regression model for recurrent gap time data that allows covariates to have different effects on the first event time and on the following gap times. In our data, patients' recurrent infection process was initiated by the event of a transplant, which is a different type of event to recurrent events (i.e., infections). Hence, the first event time (i.e., time from a transplant to the first infection) and the following gap times (i.e., gap times between two consecutive infections) may have different clinical significance and, thus, 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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

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

When the exchangeability condition on recurrent gap times beyond the first infection time is not satisfied, we can apply the multistate gap time model (Huang, 2002) to the data. The covariates' effects on the 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. However, if the number of states is large, the events of higher states may become rare, which could result in an inefficient estimation.

Our study focuses on early-phase infections after transplantation and on the effects of factors that do not vary over time, for example, patient- and transplant-related characteristics. When a longer follow-up period is of interest, the recurrent gap time structure may become more complex and, thus, may be affected by time-varying variables. Therefore, 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 times. In this

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

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).

Supplementary Material

Web Appendices A and B, and Web Table S1 referenced in Sections 2.3 and 4, respectively, are available online as supplementary material.

Acknowledgements

The authors thank the Associate Editor and two reviewers for their valuable comments, and Dr James Hodges for the enlightening discussions with the authors. The authors also gratefully acknowledge the University of Minnesota Supercomputing Institute and the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing computing resources that have contributed to the research results reported within this paper. This research was supported by the National Institutes of Health grants R01CA193888 to Huang and R03CA187991 to Luo.

References

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

children: impact of stem cell source. *Biol. Blood Marrow Transplant.* **11**, 362-370.

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

Dabrowska, D. (1989). Uniform consistency of the kernel conditional kaplan-meier estimate. *Ann. Stat* **17**, 1157-1167.

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

Huang, Y. (2002). Censored regression with the multistate accelerated sojourn times model. *J. Roy. Stat. Soc. Ser. B* **64**, 17-29.

Huang, Y. and Chen, Y.-Q. (2003). Marginal regression of gaps between recurrent events. *Lifetime Data Anal.* **9**, 293-303.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data* (2nd ed., Wiley series in probability and statistics). Hoboken, N.J.: J. Wiley.

Lee, C. H., Luo, X., Huang, C.-Y., DeFor, T. E., Brunstein, C. G. and Weisdorf, D. J. (2016). Nonparametric methods for analyzing recurrent gap time data with application to infections after hematopoietic cell transplant. *Biometrics* **72**, 535-545.

Liu, L., Wolfe, R. A. and Huang, X. (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics* **60**, 747-756.

Lu, W. (2005). Marginal regression of multivariate event times based on linear transformation models.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

Lifetime Data Anal. **11**, 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. *Stat. Med.* **30**, 301-311.

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

Saavedra, S., Sanz, G. F., Jarque, I., Moscardó, F., Jiménez, C., Lorenzo, I., Martín, G., Martínez, J., De La Rubia, J., Andreu, R., Mollá, S., Llopis, I., Fernández, M. J., Salavert, M., Acosta, B., Gobernado, M. and Sanz, M. A. (2002). Early infections in adult patients undergoing unrelated donor cord blood transplantation. *Bone Marrow Transplant.* **30**, 937-943.

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

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

Wang, H. J. and Wang, L. (2009). Locally weighted censored quantile regression. *J. Amer. Statist. Assoc.* **104**, 1117-1128.

Wang, M.-C. and Chang, S.-H. (1999). Nonparametric estimation of a recurrent survival function. *J. Amer. Statist. Assoc.* **94**, 146-153.

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

Yazaki, M., Atsuta, Y., Kato, K., Kato, S., Taniguchi, S., Takahashi, S., Ogawa, H., Kouzai, Y.,

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

Kobayashi, T., Inoue, M., Kobayashi, R., Nagamura-Inoue, T., Azuma, H., Takanashi, M., Kai, S., Nakabayashi, M. and Saito, H. (2009). Incidence and risk factors of early bacterial infections after unrelated cord blood transplantation. *Biol. Blood Marrow Transplant.* **15**, 439-446.

Zeng, D. and Lin, D. Y. (2008). Semiparametric transformation models with random effects for joint analysis of recurrent and terminal events. *Biometrics* **65**, 746-752.

Department of Biostatistics and Epidemiology,

University of Massachusetts, Amherst, MA 01003, U.S.A.

E-mail: chihyunlee@umass.edu (Corresponding Author)

Department of Epidemiology and Biostatistics,

University of California San Francisco, San Francisco, CA 94158, U.S.A.

E-mail: ChiungYu.Huang@ucsf.edu

Biostatistics Core, Masonic Cancer Center,

University of Minnesota, Minneapolis, MN 55455, U.S.A.

E-mail: defor001@umn.edu

Division of Hematology, Oncology and Transplantation, Department of Medicine,

Blood and Marrow Transplantation Program, Masonic Cancer Center,

University of Minnesota, Minneapolis, MN 55455, U.S.A.

E-mail: bruns072@umn.edu

Division of Hematology, Oncology and Transplantation, Department of Medicine,

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0397

Blood and Marrow Transplantation Program, Masonic Cancer Center,

University of Minnesota, Minneapolis, MN 55455, U.S.A.

E-mail: weisd001@umn.edu

Division of Biostatistics, School of Public Health,

Biostatistics Core, Masonic Cancer Center,

University of Minnesota, Minneapolis, MN 55455, U.S.A.

E-mail: luox0054@umn.edu