<table>
<thead>
<tr>
<th><strong>Statistica Sinica Preprint No:</strong> SS-2017-0397</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Manuscript ID</strong></td>
</tr>
<tr>
<td><strong>URL</strong></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
</tr>
</tbody>
</table>
| **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 |
SEMIPARAMETRIC REGRESSION MODEL
FOR RECURRENT BACTERIAL INFECTIONS
AFTER HEMATOPOIETIC STEM CELL TRANSPLANTATION

Chi Hyun Lee¹, Chiung-Yu Huang², Todd E. DeFor³, Claudio G. Brunstein³
Daniel J. Weisdorf³ and Xianghua Luo³

¹University of Massachusetts, Amherst,
²University of California, San Francisco and ³University 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
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
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
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-
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
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 $Y_{ij}, j = 1, 2, \ldots$, denote the gap times between two consecutive infections. The collection of all gap times for subject $i$, $i = 1, \ldots, n$, is denoted as $N_i = \{X_0^i, Y_{ij}^i, j = 1, 2, \ldots\}$, in the absence of censoring. Let $A_i$ 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
related to the covariates, as follows:

\[
\log X_i^0 = \gamma_i^0 + A_i^T \beta_0 + \epsilon_i^0
\]

\[
\log Y_{ij}^0 = \gamma_i^1 + A_i^T \beta_1 + \epsilon_{ij}, \quad j = 1, 2, \ldots,
\]

respectively, where \(\beta_0\) and \(\beta_1\) are \(p \times 1\) vectors of coefficients specific to the first event time and the following gap times, respectively; \((\gamma_i^0, \gamma_i^1)\) 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 \((\gamma_i^0, \gamma_i^1)\), 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 \(m_i = 1, \ldots\)
$X^0_i \leq C_i$ and $X^0_i + Y^0_{ij} > C_i$; and when $m_i > 1$, $X^0_i + \sum_{j=1}^{m_i-1} Y^{0}_{ij} \leq C_i$ and $X^0_i + \sum_{j=1}^{m_i} Y^{0}_{ij} > C_i$. The censoring time $C_i$ is assumed to be independent of $N_i$, $(\gamma_{i0}, \gamma_{i1})$, and $A_i$. 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 $Z^0_{i0} = X^0_i$ and $Z^0_{i1} = X^0_i + Y^0_{ij}$ 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 $A_{ii'} = A_{i'} - A_i$. The transformed times from a transplant to the first and the second infections are
defined as

\[ Z_{i'i'}^0(b_0) = \exp(A_{i'i'}^T b_0) X_i^0 \]
\[ Z_{i'i'}^0(b) = \exp(A_{i'i'}^T b_0) X_i^0 + \exp(A_{i'i'}^T b_1) Y_{i'i'}^0 , \]

respectively, where \( b = (b_0^T, b_1^T)^T \) for \( i, i' = 1, \ldots, n \). Given \( A_i \) and \( A_{i'} \), it follows that \( Z_{i'i'}^0(b_0) \) shares the same distribution with \( Z_{i'i'}^0 \), and \( Z_{i'i'}^0(b) \) shares the same distribution with \( Z_{i'i'}^1 \) when \( b_0 = \beta_0 \) and \( b_1 = \beta_1 \) under the model assumption. By constructing the transformed time to the second infection as the sum of two transformed gap times \( X_i^0 \) and \( Y_{i'i'}^0 \), the covariate effects on each gap time can be evaluated distinctively. Note that when \( A_i = A_{i'} \), the transformed times reduce to \( Z_{i'i'}^0 \) and \( Z_{i'i'}^1 \). 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 \( \{ Z_{i'i'}^0, Z_{i'i'}^0(b_0) \} \) and \( \{ Z_{i'i'}^1, Z_{i'i'}^1(b) \} \). It is obvious that, given \( A_i \) and \( A_{i'} \), \( \{ Z_{i'i'}^0, Z_{i'i'}^0(b_0) \} \) has the same distribution as \( \{ Z_{i'i'}(\beta_0), Z_{i'i'}^0 \} \), denoted by \( \{ Z_{i'i'}^0, Z_{i'i'}(\beta_0) \} \sim \{ Z_{i'i'}(\beta_0), Z_{i'i'}^0 \} \), and that \( \{ Z_{i'i'}^1, Z_{i'i'}(\beta) \} \sim \{ Z_{i'i'}(\beta), Z_{i'i'}^1 \} \), where \( \beta = (\beta_0^T, \beta_1^T)^T \). 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 \( t \vee s \geq L \),
where \( a \vee b = \max(a, b) \), for \( L < \tau_C \). Then, it follows that, conditional on \( A_i \) and \( A_i' \), \( O_{L_0}\{Z_{i0}^0, Z_{ii'0}^0(b_0)\} \sim O_{L_0}\{Z_{i0}^0, Z_{ii'0}^0(b_0)\} \) and \( O_{L_1}\{Z_{i1}^0, Z_{ii'1}^0(b)\} \sim O_{L_1}\{Z_{i1}^0, Z_{ii'1}^0(b)\} \), for constants \( L_0 < \tau_C \) and \( L_1 < \tau_C \) and \( b = \beta \). This implies that, when evaluated under the truth, \( E[w(A_i, A_i', \beta_0)A_{ii'}O_{L_0}\{Z_{i0}^0, Z_{ii'0}^0(\beta_0)\}] = 0 \) and \( E[w(A_i, A_i', \beta_1)A_{ii'}O_{L_1}\{Z_{i1}^0, Z_{ii'1}^0(\beta)\}] = 0 \), where \( w \) is a continuous and symmetric scalar weight function satisfying \( w(a_1, a_2, b) = w(a_2, a_1, b) \) for fixed \( b \).

Let the observed times from a transplant to the first two infections and the corresponding censoring indicators be denoted as \( Z_{i0} = Z_{i0}^0 \land C_i, \Delta_{i0} = I(Z_{i0}^0 \leq C_i), Z_{i1} = Z_{i1}^0 \land C_i, \Delta_{i1} = I(Z_{i1}^0 \leq C_i) \), respectively, where \( a \land b = \min(a, b) \).

The observed gap times are \( X_i = Z_{i0} \) and \( Y_{i1} = Z_{i1} - Z_{i0} \), respectively. The observed analogs of \( Z_{ii'0}^0(b_0) \) and \( Z_{ii'1}^0(b) \) are then defined as

\[
Z_{ii'0}(b_0) = \exp(A_{ii'}^T b_0)X_i,
\]
\[
Z_{ii'1}(b) = \exp(A_{ii'}^T b_0)X_i + \exp(A_{ii'}^T b_1)Y_{i1},
\]

respectively. Recall that \( G(t) \) is the survival function for the censoring time.
Then, under the random censoring assumption, we can easily show that

\[
E \left[ \frac{\Delta_0 O_L \{Z_{i0}, Z_{ii'0}(\beta_0)\}}{G(Z_{i0} \land L_0)} \bigg| A_i, A_{i'} \right] = E \left[ O_L \{Z_{i0}', Z_{ii'0}(\beta_0)\} \right| A_i, A_{i'} \] and

\[
E \left[ \frac{\Delta_1 O_L \{Z_{i1}, Z_{ii'1}(\beta)\}}{G(Z_{i1} \land L_1)} \bigg| A_i, A_{i'} \right] = E \left[ O_L \{Z_{i1}', Z_{ii'1}(\beta)\} \right| A_i, A_{i'} .
\]

Therefore, it follows that

\[
E \left[ E \left[ w(A_i, A_{i'}, \beta_0) A_{ii'} \frac{\Delta_0 O_L \{Z_{i0}, Z_{ii'0}(\beta_0)\}}{G(Z_{i0} \land L_0)} \bigg| A_i, A_{i'} \right] \right] = 0 \quad \text{and}
\]

\[
E \left[ E \left[ w(A_i, A_{i'}, \beta_1) A_{ii'} \frac{\Delta_1 O_L \{Z_{i1}, Z_{ii'1}(\beta)\}}{G(Z_{i1} \land L_1)} \bigg| A_i, A_{i'} \right] \right] = 0.
\]

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

\[
D_0(b_0) = n^{-2} \sum_{i=1}^{n} \sum_{i'=1}^{n} w(A_i, A_{i'}, b_0) A_{ii'} \frac{\Delta_0 O_L \{Z_{i0}, Z_{ii'0}(b_0)\}}{G_0(Z_{i0} \land L_0)} , \quad (2.2)
\]

\[
D_1(b) = n^{-2} \sum_{i=1}^{n} \sum_{i'=1}^{n} w(A_i, A_{i'}, b_1) A_{ii'} \frac{\Delta_1 O_L \{Z_{i1}, Z_{ii'1}(b)\}}{G_1(Z_{i1} \land L_1)} , \quad (2.3)
\]

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 \( Z_{i0}' \) and \( Z_{i1}' \) have larger maximum support than
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 $\beta_0$, we solve $D_0(b_0) = 0$, denoting the solution as $\hat{\beta}_0$. Then, we solve $D_1\{ (\hat{\beta}_0^T, b_1^T)^T \} = 0$ to derive the estimator of $\beta_1$. We denote the resulting estimator of $\beta$, derived using Huang’s method, as $\bar{\beta}$.

As discussed in Huang (2002), the log-rank estimation approaches of the univariate AFT model can be applied directly to the data to estimate $\beta_0$. However, such approaches cannot be used to estimate $\beta_1$ 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
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 $Y_{ij} = Y^0_{ij}$ 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 $m_i$, $(\gamma^0_i, \gamma^1_i)$, and $A_i$. 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, $Z_{i\prime\prime\prime1}(b)$ in (2.1) with $Z_{i\prime\prime\prime j}(b) = \exp(A^T_{i\prime\prime\prime j}b_0)X_i + \exp(A^T_{i\prime\prime\prime j}b_1)Y_{ij}$, for $j = 1, \ldots, m^*_i$. As
a result, we can prove that

$$E \left[ \frac{1}{m_i^*} \sum_{j=1}^{m_i^*} \frac{\Delta_{ij} O_{L_1} \{ Z_{ij}, Z_{ii'}(b) \}}{G(Z_{ij} \wedge L_1)} \right] = E \left[ \frac{\Delta_{11} O_{L_1} \{ Z_{11}, Z_{ii'}(b) \}}{G(Z_{11} \wedge L_1)} \right].$$

Hence, we propose replacing the estimating equation based on (2.3) with the following estimating function for the estimation of $\beta_1$:

$$D_1^*(b) = n^{-2} \sum_{i=1}^{n} \sum_{i' = 1}^{n} w(A_{ij}, A_{ij'}, b_1) A_{ii'} \left[ \frac{1}{m_i^*} \sum_{j=1}^{m_i^*} \frac{\Delta_{ij} O_{L_1} \{ Z_{ij}, Z_{ii'}(b) \}}{G(Z_{ij} \wedge L_1)} \right].$$

(2.4)

The estimator $\hat{\beta}_1$ is derived by solving $D_1^* \{ (\hat{\beta}_0^T, b_1^T) \} = 0$, where $\hat{\beta}_0$ 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, $n^{1/2}(\hat{\beta} - \beta)$ 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 $\Sigma$, $\Omega$, $\hat{\Sigma}$, and $\hat{\Omega}$ and the detailed proofs can be found in Web Appendices A.2–A.5.
Compared with $D_1(b)$ 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 $\beta_1$ 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 $O_L(t, s) = \log \left\{ \left( t \vee s \right) \wedge L \right\} - \log(L)$ 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 $A$. 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, $G(t \mid A)$. If the covariates have finite numbers of values, such as the treatment arms in randomized trials, $G(t \mid A)$ can be estimated nonparametrically by the covariate-specific Kaplan–Meier estimator $\hat{G}_j(t \mid A)$, using data $(Z_{ij}, 1 - \Delta_{ij})$.
for $i$ such that $A_i = A$ and $j = 0, 1$. When $A$ 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 $G(t \mid A)$ nonparametrically (Dabrowska, 1989; Wang and Wang, 2009). Under the conditional independent censoring assumption, we have

$$E \left[ \frac{\Delta_{i0} O_{L0} \{Z_{i0}, Z_{i0}'(\beta_0)\}}{G(Z_{i0} \wedge L_0 \mid A_i)} \bigg| A_i, A_i' \right] = E \left[ O_{L0} \{Z_{i0}, Z_{i0}'(\beta_0)\} \bigg| A_i, A_i' \right]$$

and

$$E \left[ \frac{1}{m_i^*} \sum_{j=1}^{m_i^*} E \left[ \frac{\Delta_{ij} O_{L_{ij}} \{Z_{ij}, Z_{iij}'(\beta)\}}{G(Z_{ij} \wedge L_{ij} \mid A_i)} \bigg| m_i, (\gamma_{i0}, \gamma_{i1}), A_i, A_i' \right] \right]$$

$$= E \left[ O_{L_{ij}} \{Z_{i1}, Z_{i1}'(\beta)\} \bigg| A_i, A_i' \right].$$

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

\[
D_c^0(b_0) = n^{-2} \sum_{i=1}^{n} \sum_{i'=1}^{n} w(A_i, A_{i'}, b_0) A_{ii'} \frac{\Delta_{0i} O_{L_0} \{Z_{i0}, Z_{ii'0}(b_0)\}}{\hat{G}_0(Z_{i0} \wedge L_0 \mid A_i)}, \quad (2.5)
\]

\[
D_c^1(b) = n^{-2} \sum_{i=1}^{n} \sum_{i'=1}^{n} w(A_i, A_{i'}, b_1) A_{ii'} \left[ \frac{1}{m_i^*} \sum_{j=1}^{m_i^*} \Delta_{ij} O_{L_1} \{Z_{ij}, Z_{ii'j}(b)\} \right], \quad (2.6)
\]

where \( \hat{G}_j(t \mid A), j = 0, 1 \), are consistent estimators of \( G(t \mid A) \). We denote the solution to \( D_c^0(b_0) = 0 \) as \( \tilde{\beta}_0 \). The estimator \( \tilde{\beta}_1 \) is obtained by solving \( D_c^1(\tilde{\beta}_0^T, b_1^T)^T = 0 \). Let \( \tilde{\beta} = (\tilde{\beta}_0^T, \tilde{\beta}_1^T)^T \). In Web Appendix B, we provide proofs of the asymptotic properties of \( \tilde{\beta} \) under the conditional independent censoring assumption when the covariate-specific Kaplan–Meier estimator is used for the estimation of \( G(t \mid A) \). Similar techniques can be used to establish the asymptotic properties when a semiparametric regression model is used to estimate \( G(t \mid A) \).

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

\[
\log(X^0_i) = \gamma_{i0} + A_i^T \beta_0 + \epsilon_{i0}
\]
\[
\log(Y^0_{ij}) = \gamma_{i1} + A_i^T \beta_1 + \epsilon_{ij}, j = 1, 2, \ldots,
\]

respectively, where \( A_i = (A_{i1}, A_{i2})^T \), 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 \( \beta_0 = (-0.5, 0.5)^T \) and \( \beta_1 = (0.5, 0.5)^T \). We generated the mutually independent error terms \( \epsilon_{ij} \) 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 \( \rho \) accounts for the degree of association between the (transformed) time to the first infection, \( \log(X^0_i) \), and one of the (transformed) gap times after the first infection, \( \log(Y^0_{ij}) \). In addition, \( \sigma_1 \) indicates the level of correlation between two (transformed) gap times after the first infection, \( \log(Y^0_{ij}) \) and \( \log(Y^0_{ij}) \). We set \( \sigma_0^2 = 0.5 \), and consider \( \sigma_1^2 = 0.1 \) or 0.5 and \( \rho = 0 \) or
0.5 in different scenarios. The censoring time $C_i, i = 1, \ldots, n,$ is sampled independently from a uniform distribution $(0, U)$, where $U = 10, 30, \text{ or } 50$. 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 ($\beta_0$). 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 ($\beta_1$) 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
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.
Table 1: Summary of simulation results for censoring time $C \sim \text{Unif}(0, 10)$: 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).

<table>
<thead>
<tr>
<th>$\beta_0$</th>
<th>Proposed/Huang</th>
<th>Proposed Huang</th>
<th>Chang</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$\sigma^2$</td>
<td>True</td>
<td>$\alpha$</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\rho$</th>
<th>$\sigma^2$</th>
<th>$\beta_0$ Proposed/Huang</th>
<th>$\beta_1$ Proposed</th>
<th>$\beta_1$ Huang</th>
<th>$\beta_1$ Chang</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$A_1$</td>
<td>$A_2$</td>
<td>$A_1$</td>
<td>$A_2$</td>
</tr>
<tr>
<td>150</td>
<td>0.1</td>
<td>0.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Est</td>
<td>-0.504</td>
<td>0.498</td>
<td>0.491</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SD</td>
<td>0.161</td>
<td>0.284</td>
<td>0.162</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SE</td>
<td>0.163</td>
<td>0.282</td>
<td>0.156</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CP</td>
<td>0.948</td>
<td>0.949</td>
<td>0.935</td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td></td>
<td>Est</td>
<td>-0.501</td>
<td>0.510</td>
<td>0.497</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SD</td>
<td>0.167</td>
<td>0.293</td>
<td>0.210</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SE</td>
<td>0.164</td>
<td>0.286</td>
<td>0.203</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CP</td>
<td>0.950</td>
<td>0.939</td>
<td>0.931</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td></td>
<td>Est</td>
<td>-0.497</td>
<td>0.504</td>
<td>0.484</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SD</td>
<td>0.162</td>
<td>0.297</td>
<td>0.163</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SE</td>
<td>0.163</td>
<td>0.282</td>
<td>0.155</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CP</td>
<td>0.955</td>
<td>0.938</td>
<td>0.929</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td></td>
<td>Est</td>
<td>-0.505</td>
<td>0.499</td>
<td>0.490</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SD</td>
<td>0.164</td>
<td>0.291</td>
<td>0.198</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SE</td>
<td>0.163</td>
<td>0.283</td>
<td>0.197</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CP</td>
<td>0.954</td>
<td>0.942</td>
<td>0.941</td>
</tr>
</tbody>
</table>

$\alpha$: True $\beta$-values do not exist;  
$\mu$: average number of observed infections per subject;  
$c_{r1}$: average proportion of subjects without any infections;  
$c_{r2}$: average proportion of subjects with the first or the second gap times censored.
Table 3: Summary of simulation results for censoring time $C \sim \text{Unif}(0, 50)$: 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).

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\rho$</th>
<th>$\sigma^2$</th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Proposed/Huang</td>
<td>Proposed/Huang</td>
<td>Proposed/Huang</td>
<td>Proposed/Huang</td>
<td>Huang</td>
<td>Huang</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\alpha_1$</td>
<td>$\alpha_2$</td>
<td>$\alpha_1$</td>
<td>$\alpha_2$</td>
<td>$\alpha_1$</td>
<td>$\alpha_2$</td>
</tr>
<tr>
<td>150</td>
<td>0.01</td>
<td>0.1</td>
<td>-0.496</td>
<td>0.503</td>
<td>0.499</td>
<td>0.500</td>
<td>0.497</td>
<td>0.498</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.141</td>
<td>0.472</td>
<td>0.140</td>
<td>0.470</td>
<td>0.139</td>
<td>0.468</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.156</td>
<td>0.278</td>
<td>0.154</td>
<td>0.265</td>
<td>0.153</td>
<td>0.254</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.158</td>
<td>0.275</td>
<td>0.154</td>
<td>0.266</td>
<td>0.153</td>
<td>0.255</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.951</td>
<td>0.945</td>
<td>0.942</td>
<td>0.955</td>
<td>0.934</td>
<td>0.950</td>
</tr>
<tr>
<td>300</td>
<td>0.01</td>
<td>0.1</td>
<td>-0.498</td>
<td>0.503</td>
<td>0.498</td>
<td>0.500</td>
<td>0.498</td>
<td>0.498</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.156</td>
<td>0.285</td>
<td>0.154</td>
<td>0.263</td>
<td>0.153</td>
<td>0.252</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.157</td>
<td>0.273</td>
<td>0.152</td>
<td>0.243</td>
<td>0.151</td>
<td>0.232</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.947</td>
<td>0.936</td>
<td>0.942</td>
<td>0.940</td>
<td>0.931</td>
<td>0.949</td>
</tr>
</tbody>
</table>

-: True $\beta$-values do not exist.

$\bar{m}$: average number of observed infections per subject;

$\text{cr}_1$: average proportion of subjects without any infections;

$\text{cr}_2$: average proportion of subjects with the first or the second gap times censored.
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).

<table>
<thead>
<tr>
<th>n</th>
<th>ρ</th>
<th>σ²</th>
<th>True</th>
<th>β₁ Proposed</th>
<th>β₁ Huang</th>
<th>β₂ Proposed</th>
<th>β₂ Huang</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
<td>-0.497</td>
<td>0.504</td>
<td>0.504</td>
<td>0.495</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td>-0.155</td>
<td>0.160</td>
<td>0.125</td>
<td>0.145</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.158</td>
<td>0.159</td>
<td>0.139</td>
<td>0.153</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.953</td>
<td>0.951</td>
<td>0.968</td>
<td>0.959</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.83</td>
<td>0.16; 0.42</td>
<td>0.16; 0.42</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.1</td>
<td>0.5</td>
<td>-0.502</td>
<td>0.500</td>
<td>0.503</td>
<td>0.505</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td>-0.149</td>
<td>0.154</td>
<td>0.186</td>
<td>0.198</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.158</td>
<td>0.159</td>
<td>0.199</td>
<td>0.210</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.957</td>
<td>0.964</td>
<td>0.951</td>
<td>0.954</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.32; 0.16; 0.42</td>
<td>0.16; 0.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
<td>-0.501</td>
<td>0.497</td>
<td>0.505</td>
<td>0.502</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.106</td>
<td>0.106</td>
<td>0.087</td>
<td>0.094</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.111</td>
<td>0.112</td>
<td>0.100</td>
<td>0.108</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.966</td>
<td>0.964</td>
<td>0.975</td>
<td>0.974</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.32; 0.16; 0.42</td>
<td>0.16; 0.42</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( \bar{m} \): average number of observed infections per subject; 
\( b_{cr1} \): average proportion of subjects without any infections; 
\( c_{cr2} \): average proportion of subjects with the first or the second gap times censored.
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 groups at the time of transplant: pediatric patients (< 18 years old, n = 155) and adult patients (≥ 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
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, $k = 0, 1, \cdots, 6$.

<table>
<thead>
<tr>
<th>Group</th>
<th>No. patients (%)</th>
<th>No. of infections observed for a patient</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Patients</td>
<td>516 (100)</td>
<td>0 (51.6) 1 (29.5) 2 (13.4) 3 (2.9) 4 (1.9) 5 (0.4) 6 (0.4)</td>
</tr>
<tr>
<td>Child</td>
<td>155 (100)</td>
<td>0 (59.4) 1 (29.0) 2 (9.0) 3 (1.9) 4 (0.7) 5 (0.0) 6 (0.0)</td>
</tr>
<tr>
<td>Adult</td>
<td>361 (100)</td>
<td>0 (48.2) 1 (29.6) 2 (15.2) 3 (3.3) 4 (2.5) 5 (0.6) 6 (0.6)</td>
</tr>
</tbody>
</table>

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
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
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.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Children</th>
<th>Adults</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at Transplant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Years)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMV Serostatus Positive vs. Negative</td>
<td>-0.49 (0.37)</td>
<td>0.67 (0.35)</td>
</tr>
<tr>
<td>Type of Transplant Double vs. Single</td>
<td>-0.81 (0.28*)</td>
<td>-0.08 (0.52)</td>
</tr>
<tr>
<td>Conditioning Regimen (vs. Myeloablative)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-myeloablative w ATG</td>
<td>NI</td>
<td>NI</td>
</tr>
<tr>
<td>Non-myeloablative wo ATG</td>
<td>NI</td>
<td>NI</td>
</tr>
<tr>
<td>HLA Match Score</td>
<td>-0.74 (0.39)</td>
<td>-0.54 (0.78)</td>
</tr>
<tr>
<td>5-6/6 vs. 4/6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GVHD Prophylaxis CSA/MMF/MTX vs. Other</td>
<td>-0.74 (0.38)</td>
<td>0.34 (0.37)</td>
</tr>
<tr>
<td>CD34+ Dose Level High vs. Low</td>
<td>-0.01 (0.43)</td>
<td>-1.72 (0.69*)</td>
</tr>
<tr>
<td>TNC Dose Level High vs. Low</td>
<td>1.02 (0.32*)</td>
<td>-0.29 (0.49)</td>
</tr>
<tr>
<td>HLA Match Score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-6/6 vs. 4/6</td>
<td>-0.61 (0.38)</td>
<td>-1.06 (0.73)</td>
</tr>
<tr>
<td>GVHD Prophylaxis CSA/MMF/MTX vs. Other</td>
<td>-0.07 (0.43)</td>
<td>0.84 (0.88)</td>
</tr>
<tr>
<td>CD34+ Dose Level High vs. Low</td>
<td>-0.70 (0.43)</td>
<td>-1.32 (1.01)</td>
</tr>
<tr>
<td>TNC Dose Level High vs. Low</td>
<td>1.42 (0.45*)</td>
<td>-0.75 (0.76)</td>
</tr>
</tbody>
</table>

*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.
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
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
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


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,
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