Published in final edited form as:

Biometrics. 2016 June ; 72(2): 535–545. doi:10.1111/biom.12439.

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

Chi Hyun Lee¹, Xianghua Luo^{1,2}, Chiung-Yu Huang³, Todd E. DeFor², Claudio G. Brunstein^{4,5}, and Daniel J. Weisdorf^{4,5}

Xianghua Luo: luox0054@umn.edu

¹Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, U.S.A

²Biostatistics Core, Masonic Cancer Center, University of Minnesota, Minneapolis, MN 55455, U.S.A

³Division of Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD 21205, U.S.A

⁴Division of Hematology, Oncology and Transplantation, Department of Medicine, University of Minnesota, Minneapolis, MN 55455, U.S.A

⁵Blood and Marrow Transplantation Program, Masonic Cancer Center, University of Minnesota, Minneapolis, MN 55455, U.S.A

Summary

Infection is one of the most common complications after hematopoietic cell transplantation. Many patients experience infectious complications repeatedly after transplant. Existing statistical methods for recurrent gap time data typically assume that patients are enrolled due to the occurrence of an event of interest, and subsequently experience recurrent events of the same type; moreover, for one-sample estimation, the gap times between consecutive events are usually assumed to be identically distributed. Applying these methods to analyze the post-transplant infection data will inevitably lead to incorrect inferential results because the time from transplant to the first infection has a different biological meaning than the gap times between consecutive recurrent infections. Some unbiased yet inefficient methods include univariate survival analysis methods based on data from the first infection or bivariate serial event data methods based on the first and second infections. In this paper, we propose a nonparametric estimator of the joint distribution of time from transplant to the first infection and the gap times between consecutive infections. The proposed estimator takes into account the potentially different distributions of the two types of gap times and better uses the recurrent infection data. Asymptotic properties of the proposed estimators are established.

Correspondence to: Xianghua Luo, luox0054@umn.edu.

Supplementary Materials: Web Appendices referenced in Section 2, a Web Figure and Web Tables referenced in Sections 1 and 4, and the computing code are available at the *Biometrics* website on Wiley Online Library.

Keywords

Bivariate distribution; gap times; nonparametric method; recurrent events

1. Introduction

Serious infection is a major complication after hematopoietic cell transplantation (HCT). It accounts for substantial morbidity and mortality among transplanted patients. Many patients experience infectious events repeatedly. The times between recurrent infections (i.e., gap times) are the natural outcome of interest. Similar to the discussion in Lin, Sun, and Ying (1999), the time from transplant to the first episode of an infectious event and the gap time from one episode to the next episode of infection could both be of interest. The transplant and other adjuvant treatments may function differently on the time to the first infection and gap times between consecutive infections. In this case, the total time from treatment to each episode of the event (i.e., time to events) is of less interest because a treatment which delays the first episode will certainly prolong the total time from the treatment to any subsequent episode even if it is actually ineffective after the first episode. In this paper, we are concerned with gap time from HCT to the first infection and the gap times between recurrent infections. To facilitate the discussion, we first introduce the post-HCT infection data which motivated our work, then we discuss the issues in the existing statistical methods.

1.1 The Post-HCT Infections Data

Infectious events were documented for a total of 1001 patients with malignant disease who received their first HCT between 2000 and 2010 at the University of Minnesota. Demographic and treatment related variables for this population are summarized in Web Table 1. Part of the data have been published in Barker et al. (2005). The database includes various types of infections, while in this paper we restrict our analysis to the two most prevalent types of infections, bacterial and viral infections. As we know, patients have the highest risk of infections prior to the engraftment of donor hematopoietic cells. Among them, neutrophils, a type of white blood cells important for fighting infections, could take as long as 42 days to engraft. Beyond day 42, a patient would be considered an engraftment failure. In this paper, we focus on early infections occurring between day 0 and day 42 post HCT, a short but critical time period for infectious complication management.

Our goal is to describe the nature of early infection process in transplanted patients using nonparametric methods. Specifically, we are interested in estimating the joint distribution of the time from transplant to the first episode of infection and the gap times between consecutive infection episodes. In addition, we are interested in estimating the conditional distribution and conditional quantiles of the gap times between two consecutive infections given the time to the first infection after HCT. It is also of interest to compare different patient subgroups defined by patient- or transplant-related characteristics. For example, patients who received myeloablative and non-myeloablative immunosuppressive regimen could have different risks of bacterial infections (van Burik and Weisdorf, 1999), while patients' cytomegalovirus (CMV) serostatus before transplant could affect CMV infection risk. Knowledge obtained from subgroup analyses using one-sample estimation method can

help to inform the regression analysis used to formally correlate risk factors to the risk of infections.

1.2 Issues in Existing Nonparametric Methods

For univariate survival data (i.e., single event data) such as death, the well-known Kaplan-Meier method (Kaplan and Meier, 1958) provides a useful tool for describing the distribution of the length of the time from the beginning of the follow-up (e.g., the time when a patient receives a treatment) to the time when the event of interest occurs. Corresponding one-sample methods are available for recurrent gap time data such as the estimators considered by Wang and Chang (1999, referred to as the “Wang-Chang estimator” hereinafter) and Peña, Strawderman, and Hollander (2001). However, these existing methods only consider the situation when a patient's enrollment is triggered by an initial event that is of the same type as the recurrent events and assume that all gap times, including the first event time, are identically distributed. Applying these methods to the post-HCT infection data by ignoring the fact that the initial event (i.e., transplant) is not of the same type as the recurrent events (i.e., infections) will inevitably lead to incorrect inferential results. This is because the time from transplant to the first infection has different clinical significance than gap times between recurrent infections after the first infection occurs: The immune response during immune reconstitution after transplant may have different profiles before and after the first infection.

Alternatively, one may analyze data after the first infection to make the existing recurrent gap time methods applicable, but this introduces selection bias because only patients who have experienced at least one infection are included in the analysis. Other naive methods include applying univariate survival data methods on the time to the first infection data only or using bivariate serial event data methods on the data up to the second infection, such as the estimators considered by Visser (1996), Huang and Louis (1998, referred to as the “Huang-Louis estimator” hereinafter), Wang and Wells (1998), and Lin, Sun, and Ying (1999, referred to as the “Lin-Sun-Ying estimator” hereinafter). Thus, all subsequent infection data beyond the first or the second infectious events are ignored. These approaches lead to either loss of information or failure to address appropriate scientific questions. In this paper, we develop a nonparametric estimator that can efficiently use data beyond the first and second infectious episodes for the estimation of the joint distribution.

In Section 2, we first present the bivariate serial event method considered by Huang and Louis (1998). We then discuss how this method can be extended to recurrent event data in a way similar to the weighted risk-set method discussed by Luo and Huang (2011). Weak convergence of the proposed estimator is established by applying the empirical processes theory. In Section 3, we report results of simulation studies. In Section 4, we apply the proposed methods to the recurrent infection data of patients transplanted at the University of Minnesota. Some concluding remarks are provided in Section 5.

2. Nonparametric Estimator of the Bivariate Joint Distribution of Gap Times

2.1 Notation and Assumptions

Consider a study of post transplant infections, where patients are followed from transplant until a censoring time. For subject i , $i = 1, \dots, n$, let W_0 denote the time from transplant to the first infection and W_{ij} , $j = 1, 2, \dots$, the gap times between the following infections. Let $N_i = \{W_{ij}, j = 0, 1, \dots\}$ denote the collection of all gap times since transplant for subject i . We allow the time from transplant to the first infection (i.e., the first gap time) and subsequent gap times between successive infections to have different distributions. To highlight the difference between the two types of gap times, we denote the first gap time by $X_i^0 \equiv W_{i0}$ and the recurrent gap times after the first infection by $Y_{ij}^0 \equiv W_{ij}$, $j = 1, 2, \dots$. Let C_i be the censoring time from transplant, which has a survival function $G(t) = \Pr(C_i > t)$ with a fixed maximum support denoted by $\pi_C = \sup\{t: G(t) > 0\}$. Let m_i denote the number of completely observed infectious episodes for subject i . Figure 1 illustrates a typical patient's recurrent infection process whose first m_i infections are observed without censoring while the m_{i+1}^{th} infection is censored at time C_i (i.e., $\sum_{j=0}^{m_i-1} W_{ij} \leq C_i$ and $\sum_{j=0}^{m_i} W_{ij} > C_i$). The observed data from n subjects are assumed to be independent and identically distributed (i.i.d.). As in existing recurrent gap time methods, such as the ones considered by Wang and Chang (1999) and many others, we assume there exists a subject-specific latent variable or vector (i.e., frailty) γ_i characterizing the within-subject association among the gap times of the same subject, whose distribution is left unspecified. Then, we make the following assumptions:

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

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

Assumption 2: The censoring time C_i is independent of N_i and γ_i .

Under Assumption 2, the first gap time X_i^0 is subject to independent censoring by C_i , whereas the subsequent gap times $(Y_{ij}^0, j = 1, 2, \dots)$ are subject to dependent censoring by $C_i - X_i^0 - \dots - Y_{ij-1}^0$, which is known as “induced dependent censoring” in literature.

2.2 Estimators

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

$$F_{X^0, Y^0}(x, y) = F_{Z^0, \mathbf{V}^0}(x + y, (x, y)), \quad (1)$$

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

$$F_{Z^0, \mathbf{V}^0}(t, \mathbf{u}) = \int_0^t S_{Z^0}(s-) \frac{F_{Z^0, \mathbf{V}^0}(ds, \mathbf{u})}{S_{Z^0}(s-)} = \int_0^t S_{Z^0}(s-) \frac{F_{Z, \mathbf{V}}(ds, \mathbf{u})}{S_Z(s-)}, \quad (2)$$

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

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

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

reconstructed data $\{Z_{ij}, i=1, \dots, n, j=1, \dots, m_i^*\}$, we can estimate $S_{Z_0}(\cdot)$, in the same spirit as Wang and Chang (1999) and Huang and Wang (2005), by

$$\hat{S}_{Z_0}^*(t) = \prod_{t_k^* \leq t} \left(1 - \frac{\hat{H}(t_k^*, \infty)}{\hat{R}(t_k^*)} \right),$$

where $t_1^*, t_2^*, \dots$ are distinct and uncensored recurrence times from $\{Z_{ij}, i=1, \dots, n, j=1, \dots, m_i^*\}$ and $\hat{H}(\cdot)$ and $\hat{R}(\cdot)$ are, respectively, defined as

$$\hat{H}(t, \mathbf{u}) = \frac{1}{n} \sum_{i=1}^n \frac{I(m_i \geq 2)}{m_i^*} \sum_{j=1}^{m_i^*} I(Z_{ij} = t, X_i^0 \leq u_1, Y_{ij}^0 \leq u_2) \text{ and}$$

$$\hat{R}(t) = \frac{1}{n} \sum_{i=1}^n \frac{1}{m_i^*} \sum_{j=1}^{m_i^*} I(Z_{ij} \geq t). \text{ Note that when } j \leq m_i^* \text{ and } m_i \geq 2, \text{ the variables } X_i^0 \text{ and } Y_{ij}^0 \text{ in } \hat{H} \text{ are observed quantities. Let}$$

$$\hat{F}_{Z,V}^*(t, \mathbf{u}) = \frac{1}{n} \sum_{i=1}^n \frac{I(m_i \geq 2)}{m_i^*} \sum_{j=1}^{m_i^*} I(Z_{ij} \leq t, X_i^0 \leq u_1, Y_{ij}^0 \leq u_2). \text{ It is obvious that } \hat{F}_{Z,V}^*(dt, \mathbf{u}) = \hat{H}(t, \mathbf{u}).$$

Now, motivated by Equation (2), we propose to estimate $F_{Z^0, V^0}(t, \mathbf{u})$ by

$$\hat{F}_{Z^0, V^0}^*(t, \mathbf{u}) = \sum_{t_k^* \leq t} \left[\prod_{j < k} \left(1 - \frac{\hat{H}(t_j^*, \infty)}{\hat{R}(t_j^*)} \right) \right] \frac{\hat{H}(t_k^*, \mathbf{u})}{\hat{R}(t_k^*)}. \quad (3)$$

Note that our focus is to estimate the joint distribution, $F_{X^0, Y^0}(x, y)$, for the time from transplant to the first infection (X^0) and the between-infection gap times after the first infection (Y^0). It follows directly from (1) that $F_{X^0, Y^0}(x, y)$ can be estimated by

$$\hat{F}_{X^0, Y^0}^*(x, y) = \hat{F}_{Z^0, V^0}^*(x+y, (x, y)), \text{ which is identifiable for } x+y \leq \pi_C.$$

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

$$\hat{F}_{Y^0|X^0}^*(y|x) = \frac{\hat{F}_{X^0, Y^0}^*(x, y)}{1 - \hat{S}_{X^0}^*(x)}, x+y \leq \pi_C, \quad (4)$$

where $\hat{S}_{X^0}^*(\cdot)$ is the Kaplan-Meier estimator for the time to the first infection. Note that the marginal distribution of gap times between consecutive infections, $F_{Y^0}(y)$ is not generally identifiable for $y > 0$ unless the support for X^0 is less than π_C . Similar discussion for bivariate serial gap time data can be found in Lin, Sun, and Ying (1999), Lin and Ying

(2001), Schaubel and Cai (2004), Cook and Lawless (2007), and Lawless and Yilmaz (2011). As noted by Lin and Ying (2001), when x is large enough, the conditional distribution $F_{Y^0|X^0}(y|x)$ may be a good approximation for the marginal distribution $F_{Y^0}(y)$.

The proposed conditional distribution estimator can be used to estimate quantiles of interest. The p th conditional quantile associated with the conditional distribution function $F_{Y^0|X^0}(y|x)$, is defined as $y_p(x) = \inf\{y: 1 - F_{Y^0|X^0}(y|x) \leq 1 - p\}$. When $p = 0.5$, $y_p(x)$ is the conditional median of the gap times beyond the first infection given that the time from transplant to the first infection is no longer than x . We can estimate $y_p(x)$ by $\hat{y}_p(x) = \inf\{y: 1 - \hat{F}_{Y^0|X^0}(y|x) \leq 1 - p\}$. In practice, the $100(1 - \alpha)\%$ confidence interval (CI) for $\hat{y}_p(x)$ can be constructed using the bootstrap method without the need to estimate the variance of $\hat{F}_{Y^0|X^0}(y|x)$. Otherwise, one can obtain the linear Wald-type CI,

$\{y: -Z_{1-\alpha/2} \leq [\hat{F}_{Y^0|X^0}(y|x) - p] / \widehat{\text{Var}}^{1/2}[\hat{F}_{Y^0|X^0}(y|x)] \leq Z_{1-\alpha/2}\}$ with the variance estimate $\widehat{\text{Var}}[\hat{F}_{Y^0|X^0}(y|x)]$. Other forms such as the log-log transformed and arcsine-square root transformed CIs can also be used (see Klein and Moeschberger, 2003, p. 120).

2.3 Asymptotic Properties

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

$$S_{Z^0}(t) = \prod_{[0, t]} \{1 - \Lambda(ds, \infty)\}. \quad (5)$$

Let $\mathcal{S}(\Omega)$ denote the space of bivariate right-continuous functions on Ω with left-hand limits. Combining (2) and (5), both of which lie in $\mathcal{S}(\Omega)$, we can define a mapping $\Phi: \Lambda \rightarrow F_{Z^0, V^0}$ as follows

$$F_{Z^0, V^0}(t, \mathbf{u}) = \Phi(\Lambda)(t, \mathbf{u}) = \int_0^t \prod_{[0, s]} \{1 - \Lambda(ds, \infty)\} \Lambda(ds, \mathbf{u}). \quad (6)$$

Let the estimator of Λ be

$$\hat{\Lambda}(t, \mathbf{u}) = \int_0^t \frac{\hat{F}_{Z, V}^*(ds, \mathbf{u})}{\hat{R}(s)} \quad (7)$$

by replacing $F_{Z,V}$ and S_Z with $\hat{F}_{Z,V}^*$ and $\hat{R}$, respectively. Plugging (7) into (6), we can derive the proposed estimator of F_{Z^0,V^0} , that is, $\hat{F}_{Z^0,V^0}^* = \Phi(\hat{\Lambda})$. It can be easily verified that this is equivalent to (3).

By the exchangeability of the pairs $(X_i^0, Y_{i1}^0), (X_i^0, Y_{i2}^0), \dots$, we have

$$E \left[\frac{I(m_i \geq 2)}{m_i^*} \sum_{j=1}^{m_i^*} I(Z_{ij} \leq t, \mathbf{V}_{ij}^0 \leq \mathbf{u}) \right] = E[I(Z_{i1}^0 \leq t, \mathbf{V}_{i1}^0 \leq \mathbf{u}, \Delta_{i1}=1)] = F_{Z,V}(t, \mathbf{u})$$

and $E \left[\frac{1}{m_i^*} \sum_{j=1}^{m_i^*} I(Z_{ij} \geq t) \right] = E[I(Z_{i1} \geq t)] = S_Z(t)$. Hence, the moment type estimators $\hat{F}_{Z,V}^*$ and $\hat{R}$ involved in $\hat{\Lambda}$ both converge weakly to the same limit as their counter parts $F_{Z,V}$ and S_Z in the Huang-Louis estimator. The mapping Φ is compactly differentiable at each point of $\mathcal{S}(\Omega)$ with the derivative

$$\{d\Phi(\Lambda) \cdot h\}(t, \mathbf{u}) = \int_0^t \{F_{Z^0,V^0}(s, \mathbf{u}) - F_{Z^0,V^0}(t, \mathbf{u})\} h(ds, \infty) + \int_0^t \{1 - F_{Z^0,V^0}(s, \infty)\} h(ds, \mathbf{u}),$$

where $h \in \mathcal{S}(\Omega)$ (Andersen et al., 1993, Proposition 2.8.7). For the mapping being differentiable, it is sufficient to study the asymptotic properties of $\hat{\Lambda}$ to derive the large samples properties of the proposed method. It is straightforward that the moment type

estimators $\hat{F}_{Z,V}^*$ and $\hat{R}$ included in $\hat{\Lambda}$ both satisfy the weak convergence theorem. The proof is provided in Web Appendix A. By applying the functional delta method, the large sample properties of $\hat{\Lambda}$ can be derived. Define the function

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

Then, $\hat{\Lambda}$ has the following property.

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

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

which converges weakly to a Gaussian process with mean 0 and variance-covariance function $E[\psi_1(t_1, \mathbf{u}_1)\psi_1(t_2, \mathbf{u}_2)]$, where $(t_j, \mathbf{u}_j) \in \Omega, j = 1, 2$.

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

Define the function

$\phi_i(t, \mathbf{u}) = \int_{[0,t]} F_{Z^0, V^0}(s, \mathbf{u}) \psi_i(ds, \infty) + \int_{[0,t]} S_{Z^0}(s) \psi_i(ds, \mathbf{u}) - F_{Z^0, V^0}(t, \mathbf{u}) \psi_i(t, \infty)$. The asymptotic properties of the proposed estimator follows in Theorem 2.

Theorem 2. For any $L < \pi_C$ and $(t, \mathbf{u}) \in \Omega$, the stochastic process

$n^{1/2}\{\hat{F}_{Z^0, V^0}^*(t, \mathbf{u}) - F_{Z^0, V^0}(t, \mathbf{u})\}$ has an asymptotically i.i.d. representation

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

which converges weakly to a Gaussian process with mean 0 and variance-covariance function $E[\phi_1(t_1, \mathbf{u}_1)\phi_1(t_2, \mathbf{u}_2)]$, where $(t_j, \mathbf{u}_j) \in \Omega, j = 1, 2$.

The variance-covariance function of the limiting distribution can be consistently estimated

by $n^{-1} \sum_{i=1}^n \hat{\phi}_i(t_1, \mathbf{u}_1) \hat{\phi}_i(t_2, \mathbf{u}_2)$, where $\hat{\phi}_i$ is the estimator of ϕ_i derived by replacing

F_{Z^0, V^0}, S_{Z^0} , and ψ_i with $\hat{F}_{Z^0, V^0}^*, \hat{S}_{Z^0}^*$, and $\hat{\psi}_i$ respectively. Therefore, for $0 \leq x + y \leq L$, $n^{1/2}\{\hat{F}_{X^0, Y^0}(x, y) - F_{X^0, Y^0}(x, y)\}$ converges weakly to a Gaussian process with mean zero and variance-covariance function $E[\phi_1(x_1 + y_1, (x_1, y_1)) \phi_1(x_2 + y_2, (x_2, y_2))]$, where $0 \leq x_j + y_j \leq L, j = 1, 2$.

The proof of Theorems 1 and 2 follows closely the proof in Huang and Wang (2005) and is provided in Web Appendices B and C.

To establish the asymptotic properties for the conditional distribution $F_{Y^0|X^0}(y|x)$, we need to introduce additional notation for the first gap time. Let $X_i = \min(X_i^0, C_i)$ denote the observed first gap time and $S_X(\cdot)$ and $F_X(\cdot)$ denote its survival function and distribution function, respectively. The corresponding empirical functions for X are

$\hat{S}_X(t) = \frac{1}{n} \sum_{i=1}^n I(X_i \geq t)$ and $\hat{F}_X(t) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq t)$. Define the function

$$\xi_i(t, \mathbf{u}) = \frac{1}{1 - S_{X^0}(u_1)} \phi_i(t, \mathbf{u}) - \frac{F_{X^0, Y^0}(u_1, u_2)}{\{1 - S_{X^0}(u_1)\}^2} \phi_i'(u_1), \text{ where}$$

$$\phi_i'(u_1) = S_{X^0}(u_1) \left\{ \frac{I(X_i \leq C_i) I(X_i \leq u_1)}{S_X(X_i)} - \int_0^{u_1} \frac{I(X_i \geq s)}{S_X^2(s)} dF_X(s) \right\}.$$

The weak convergence of $F_{Y^0|X^0}(y|x) = F_{X^0, Y^0}(x, y) / \{1 - \hat{S}_{X^0}(x)\}$ follows naturally from the weak convergence of the proposed estimator $\hat{F}_{X^0, Y^0}(x, y)$ in the following theorem.

Theorem 3. For any $L < \pi_C$ and $(t, \mathbf{u}) = (u_1 + u_2, (u_1, u_2)) \in \Omega$, the stochastic process

$n^{1/2}\{\hat{F}_{Y^0|X^0}(u_2|u_1) - F_{Y^0|X^0}(u_2|u_1)\}$ has an asymptotically i.i.d. representation

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

which converges weakly to a Gaussian process with mean 0 and variance-covariance function $E[\xi_1(t_1, \mathbf{u}_1) \xi_1(t_2, \mathbf{u}_2)]$, where $(t_j, \mathbf{u}_j) \in \Omega, j = 1, 2$.

The variance-covariance function of the limiting distribution can be consistently estimated by $n^{-1} \sum_{i=1}^n \hat{\xi}_i(t_1, \mathbf{u}_1) \hat{\xi}_i(t_2, \mathbf{u}_2)$, where $\hat{\xi}_i$ is the estimator of ξ_i derived by replacing $S_X^0, S_X, F_X, F_X^0, Y^0, \phi_i$ with $\hat{S}_X^0, \hat{S}_X, \hat{F}_X, \hat{F}_X^0, \hat{Y}^0, \hat{\phi}_i$ respectively. The proof of Theorem 3 can be found in Web Appendix D.

3. Simulation Studies

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

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

Simulation Scenario 1

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

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

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

Simulation Scenario 2

In this scenario, we assume a bivariate frailty (γ_0, γ_1) . The gap times are generated from the model:

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

where the frailty (γ_0, γ_1) follows a bivariate normal distribution with zero mean and variance-covariance matrix $\begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}$. Under this setting, the covariance between the (transformed) first gap time $(\log(X_i^0))$ and a subsequent gap time $(\log(Y_{ij}^0))$ is σ_{01} and that between two transformed subsequent gap times, $\log(Y_{ij}^0)$ and $\log(Y_{ij'}^0)$, $j \neq j'$, is σ_1^2 . We let $\sigma_0^2 = \sigma_1^2 = 0.5$ and consider two levels of σ_{01} ($\sigma_{01} = 0$ and 0.25) in the simulations. The other parameters are the same as in Simulation Scenario 1.

The performance of the proposed estimator of $F_{X^0, Y^0}(x, y)$ is compared with that of the Huang-Louis estimator. Tables 1 and 2 summarize the simulation results for the two simulation scenarios, respectively, at grid points (x, y) , where $x = 15, 20, 30$ and $y = 5, 7, 15$. In each scenario, both the proposed method and the Huang-Louis estimator provide virtually unbiased estimates. The efficiency of the Huang-Louis estimator relative to the proposed estimator, measured by the squared quotient of the respective standard deviation estimates, is less than 1 in all scenarios. The efficiency gain of the proposed estimator is greater when more gap times are observed (U increases from 75 to 150), but diminishes with larger y .

We also conducted a simulation study, using the data from Simulation Scenario 1 with $U = 150$ to investigate the performance of the conditional distribution estimator $\hat{F}_{Y^0|X^0}(y|x)$ for a large x ($x = 100$) on approximating the marginal distribution of the gap times between consecutive infections, $\hat{F}_{Y^0}(y)$. The bias and power for the proposed method, together with two naive estimators, namely the Wang-Chang estimator and the Kaplan-Meier estimator, which use only data beyond the first infection, are presented in Table 3. The results show that the two naive methods have non-ignorable biases, while the conditional distribution based on the proposed method provides satisfactory estimates for the marginal distribution. We further evaluated the conditional median estimator for the between-infection gap times given different length of the time from transplant to the first infection with all simulated data. The detailed result is shown in Web Table 2. The conditional median estimates were virtually unbiased in all scenarios with the coverage probabilities of 95% CIs all close to 0.95.

4. Application

We apply the proposed method to the post-HCT infection data introduced in Section 1 to make inference about the joint distribution of time from transplant to the first infection and the gap times between recurrent infections. By day 42 after transplant, 61 (6%) out of 1001 patients were censored by death, 27 (3%) by relapse, 13 (1%) by a second transplant, and 900 (90%) were alive without a second transplant or relapse. Among the 61 patients who died, 17 (1.7% out of 1001) were found to be infection-related, of whom, 4 (0.4% out of 1001) were due to bacterial infections and 4 (0.4% out of 1001) other due to viral infections. Hence, the noninformative censoring assumption was not expected to be violated in this dataset.

Following Barker et al. (2005), an infectious episode was defined as any infection confirmed by culture, histology, polymerase chain reaction, or antigenemia for which treatment was

initiated and clinically compatible time frames were used to define one infectious episode separated from a second episode with the same organism (see Barker et al., 2005, for details). About 48% of the patients experienced at least one bacterial infection and 34% at least one viral infection. Overall, 752 bacterial and 437 viral infections were observed within 42 days after transplant. Detailed information on number of infections per patient can be found in Web Table 3 and a summary of number of infections for different infection organisms can be found in Web Table 4.

Due to the short follow-up period, we do not expect the exchangeability assumption on the gap times after the first infection would be violated. Nevertheless, we carried out trend tests by Wang and Chen (2000) to assure that the gap times after the first infection are identically distributed. The test results (see p-values in Table 4) show that neither bacterial nor viral infections had significant trend in gap times after the first infection. Table 4 presents the joint distribution estimates of the time from transplant to the first infection and the gap times between two consecutive infections of the same type, for bacterial (upper panel) and viral infections (lower panel) separately. Based on the estimated joint probabilities, bacterial infections had a greater than 2-fold increase than viral infections at all presented time points.

Conditional probability distributions of the gap times between consecutive infections, namely $\Pr(Y^0 = y | X^0 = x_1, X^0 = x_2, x_1 < x_2)$, may also be scientifically interesting as we can identify which patients experienced more frequent recurrent infections after experiencing the first infection within a certain period of time. Similar to (4), this conditional probability can be estimated by $\{\hat{F}_{X^0, Y^0}(x_2, y) - \hat{F}_{X^0, Y^0}(x_1, y)\} / \{\hat{S}_{X^0}(x_1) - \hat{S}_{X^0}(x_2)\}$ and the associated variance can be estimated by the bootstrap method. Figure 2 presents the estimated survival probability of time to the first infection by the Kaplan-Meier method and the estimated conditional probabilities of recurrent gap times for bacterial (left panel) and viral (right panel) infections. It shows that more patients experienced the first bacterial infection than viral infection. Given the first bacterial infection occurred within the first or second or third week, the probability of having another bacterial infection in the following weeks was similar, which may indicate a weak correlation between the first gap time with the following recurrent gap times for the bacterial infections. However, for viral infections, we found that given the first viral infection occurred in the first week, as opposed to the third week after transplant, the probability of having another viral infection within another 3 weeks was twice as high (0.59 [95% CI: 0.37, 0.78] vs. 0.27 [0.20, 0.37]), indicating a positive correlation between time to the first viral infection and the gap times between recurrent viral infections.

The conditional median estimates of the between-infection gap times given different length of time from transplant to the first infection is presented for bacterial and viral infections in Web Figure 1. We found that the median between-infection gap time for bacterial infections was relatively constant regardless of the length of time from transplant to the first bacterial infection, which indicates a weak correlation between the time to the first bacterial infection and the gap times between consecutive bacterial infections, consistent with our previous finding. However, for viral infections, the median between-infection gap time showed a slightly increasing trend when the time from transplant to the first infection was increasing.

In the upper panel of Table 4, we also present the joint distribution estimates for bacterial infections stratified by pre-transplant immunosuppressive regimen. The myeloablative group's probabilities were almost doubled compared with the non-myeloablative group at all time points. We compared the two groups' joint distribution estimates using a nonparametric permutation test. The null hypothesis was that F_{X^0, Y^0} was the same for the two groups and the supremum test statistic had the form $D = \sup_{(x,y)} |\hat{F}_{X^0, Y^0}^{(1)}(x, y) - \hat{F}_{X^0, Y^0}^{(2)}(x, y)|$ with the superscript $g = 1, 2$ indexing the group assignment. Group index was randomly permuted among the patients for 500 times to obtain the sampling distribution of D under the null hypothesis. We found that the difference between the myeloablative and non-myeloablative groups was highly significant ($p < 0.01$). For viral infections, we found that CMV positive serostatus was associated with higher joint probabilities than the CMV negative serostatus (Table 4, lower panel) with a significant p-value based on the permutation test ($p < 0.01$).

5. Discussion

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

Similar to the nonparametric methods by Wang and Chang (1999) and Huang and Wang (2005), our proposed method does not impose distributional assumptions on the frailty and leaves the within-subject correlation between gap times as nuisance. When the between-gap association is of interest, one can adopt the nonparametric method by Lakhal-Chaieb, Cook, and Lin (2010) based on Kendall's τ , by using the first two gap times. Alternatively, one can use semiparametric methods such as the bivariate copula model by Lawless and Yilmaz (2011) or the shared-frailty model by Huang and Liu (2007). However, the validity of the inference from these two methods depends on the correct specification of the form of the copula and the parametric distribution of the frailty, respectively.

Other nonparametric methods for bivariate serial event data such as the one proposed by Lin, Sun, and Ying (1999) can also be extended to handle recurrent infection data by using the weighted risk-set technique. For the bivariate case, it has been demonstrated by de Uña-Álvarez and Meira-Machado (2008), through simulation studies, that the Huang-Louis

estimator could be more efficient than the Lin-Sun-Ying estimator. More theoretical investigation would be needed in order to get general conclusions on their relative efficiency.

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

The primary focus of this paper is on nonparametric estimation of the joint distribution of the gap times. Other nonparametric estimators based on the joint distribution estimator such as the conditional distribution and conditional quantile estimators are also provided. The nonparametric estimators proposed in this article are not only important in their own right, but also have important statistical applications. As demonstrated in Section 4, the proposed estimator F_{X_0, Y_0} enables us to make comparisons between subgroups of patients defined by potential risk factors, which can inform a more formal regression analysis.

To assess covariate effects on gap times, one may use Huang (2002)'s regression model for multistate gap time data, where the number of states to be analyzed is prespecified (e.g., 2 states if only the first two gaps are of interest). This method was originally proposed for sequential events of different types, hence, applying it on recurrent gap times with certain exchangeability property could be inefficient. Alternatively, at the cost of modeling the frailty distribution, one can use the model by Huang and Liu (2007) by including an episode-specific covariate to distinguish the covariate effect on the first gap (transplant to the first infection) from that on the subsequent gap times. A direction of future research could be modeling multivariate recurrent gap time processes, such as the viral and bacterial infection processes jointly. We also note that an infectious event is not a transient event and an infectious episode can last for a certain period of time before being resolved. When both the infection time and infection-free time of an infectious episode are of interest, methods for alternating recurrent event processes by Huang and Wang (2005) and Chang (2004) can be considered.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the Associate Editor and two reviewers for their valuable comments, Dr. James Hodges for the enlightening discussions with the authors, and University of Minnesota Supercomputing Institute for providing computing resources. This research was supported by NIH/NCI 1R03CA187991 to Luo and 1R01CA193888 to Huang.

References

Andersen, PK.; Borgan, O.; Gill, RD.; Keiding, N. Statistical Models Based on Counting Processes. New York: Springer-Verlag; 1993.

- Barker JN, Hough RE, van Burik JA, DeFor TE, MacMillan ML, O'Brien MR, Wagner JE. Serious infections after unrelated donor transplantation in 136 children: impact of stem cell source. *Biology of Blood and Marrow Transplantation*. 2005; 11:362–370. [PubMed: 15846290]
- Billingsley, P. *Convergence of Probability Measures*. New York: Wiley; 1999.
- Breslow N, Crowley J. A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*. 1974; 2:437–453.
- Chang SH. Estimating marginal effects in accelerated failure time models for serial sojourn times among repeated events. *Lifetime Data Analysis*. 2004; 10:175–190. [PubMed: 15293631]
- Cook, RJ.; Lawless, JF. *The Statistical Analysis of Recurrent Events*. Springer; New York, NY: 2007. 2007
- de Uña-Álvarez J, Meira-Machado LF. A simple estimator of the bivariate distribution function for censored gap times. *Statistics and Probability Letters*. 2008; 78:2440–2445.
- Huang CY, Wang MC. Nonparametric estimation of the bivariate recurrence time distribution. *Biometrics*. 2005; 61:392–402. [PubMed: 16011685]
- Huang X, Liu L. A joint frailty model for survival and gap times between recurrent events. *Biometrics*. 2007; 63:389–397. [PubMed: 17688491]
- Huang Y. Censored regression with the multistate accelerated sojourn times model. *Journal of the Royal Statistical Society Series B*. 2002; 64:17–29.
- Huang Y, Louis TA. Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika*. 1998; 85:785–798.
- Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. 1958; 53:457–481.
- Klein, JP.; Moeschberger, ML. *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd. New York: Springer; 2003. 2003
- Lakhal-Chaieb L, Cook RJ, Lin X. Inverse probability of censoring weighted estimates of Kendall's τ for gap time analyses. *Biometrics*. 2010; 66:1145–1152. [PubMed: 20337629]
- Lawless JF, Yilmaz YE. Semiparametric estimation in copula models for bivariate sequential survival times. *Biometrical Journal*. 2011; 53:779–796. [PubMed: 21887793]
- Lin DY, Sun W, Ying Z. Nonparametric estimation of the gap time distribution for serial events with censored data. *Biometrika*. 1999; 86:59–70.
- Lin DY, Ying Z. Nonparametric tests for the gap time distribution of serial events based on censored data. *Biometrics*. 2001; 57:369–375. [PubMed: 11414558]
- Luo X, Huang CY. Analysis of recurrent gap time data using the weighted risk-set method and the modified within-cluster resampling method. *Statistics in Medicine*. 2011; 30:301–311. [PubMed: 20963733]
- Peña EA, Strawderman RL, Hollander M. Nonparametric estimation with recurrent event data. *Journal of American Statistical Association*. 2001; 96:1299–1315.
- Schaubel DE, Cai J. Non-parametric estimation of gap time survival functions for ordered multivariate failure time data. *Statistics in Medicine*. 2004; 23:1885–1900. [PubMed: 15195322]
- van Burik JA, Weisdorf DJ. Infections in recipients of blood and marrow transplantation. *Hematology/Oncology Clinics of North America*. 1999; 13:1065–1089. [PubMed: 10553262]
- van der Vaart, AW. *Asymptotic Statistics*. Cambridge, U.K.: Cambridge University Press; 1998.
- Visser M. Nonparametric estimation of the bivariate survival function with an application to vertically transmitted AIDS. *Biometrika*. 1996; 83:507–518.
- Wang MC, Chang SH. Nonparametric estimation of a recurrent survival function. *Journal of the American Statistical Association*. 1999; 94:146–153. [PubMed: 24244058]
- Wang MC, Chen YQ. Nonparametric and semiparametric trend analysis for stratified recurrence times. *Biometrics*. 2000; 56:789–794. [PubMed: 10985217]
- Wang WJ, Wells MT. Nonparametric estimation of successive duration times under dependent censoring. *Biometrika*. 1998; 85:561–572.

Figure 1. Illustration of time from transplant to first infection and gap times between recurrent infections

Figure 2.

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

Summary of the results for Simulation Scenario 1 with true values of the bivariate joint distribution $F_{X^0, Y^0}(x, y)$ (True); relative bias $\times 10^3$ (Monte-Carlo $SD \times 10^3$ and average asymptotic $SE \times 10^3$) of Huang-Louis estimator (H-L) and the proposed estimator (Proposed); and relative efficiency (re) of H-L vs. Proposed. Results for each setting are based on 1000 simulated datasets, each with $n = 500$ subjects.

Table 1

x	y =	$C_1 \sim Unif(0, 75)$			$C_1 \sim Unif(0, 150)$		
		5	7	15	5	7	15
$\sigma^2 = 0.1$							
		$m^B = 3.20; cr^B = 0.30$			$m = 8.14; cr = 0.15$		
15	True	0.1007	0.1827	0.2551	0.1007	0.1827	0.2551
	H-L	-2.7 (15, 15)	2.9 (20, 20)	-0.8 (23, 22)	2.4 (14, 14)	4.4 (19, 18)	2.1 (21, 21)
	Proposed	-3.2 (12, 12)	0.5 (18, 18)	-0.5 (23, 22)	-0.3 (11, 11)	4.7 (16, 16)	2.0 (21, 21)
	re^C	0.6401	0.8200	0.9952	0.5711	0.7801	0.9936
20	True	0.1497	0.3066	0.4891	0.1497	0.3066	0.4891
	H-L	0.8 (19, 18)	2.5 (24, 24)	-0.2 (28, 27)	2.7 (17, 17)	2.4 (22, 22)	1.2 (24, 24)
	Proposed	2.5 (14, 14)	1.1 (21, 21)	-0.1 (28, 27)	2.0 (12, 12)	3.0 (18, 18)	1.1 (24, 24)
	re	0.5813	0.7381	0.9869	0.4920	0.6713	0.9811
30	True	0.1847	0.4201	0.7898	0.1847	0.4201	0.7898
	H-L	0.3 (21, 20)	1.9 (27, 27)	0.1 (24, 24)	-1.1 (19, 19)	0.0 (24, 24)	0.3 (20, 20)
	Proposed	1.4 (15, 15)	0.6 (21, 21)	0.1 (24, 23)	-0.9 (12, 12)	0.8 (17, 17)	0.1 (20, 20)
	re	0.5372	0.6393	0.9469	0.4304	0.5299	0.9089
$\sigma^2 = 0.5$							
		$m = 4.39; cr = 0.34$			$m = 10.39; cr = 0.18$		
15	True	0.2432	0.3093	0.3514	0.2432	0.3093	0.3514
	H-L	0.8 (21, 21)	2.4 (22, 23)	2.3 (23, 24)	0.0 (20, 20)	1.0 (22, 22)	0.1 (23, 23)
	Proposed	2.1 (19, 19)	2.0 (21, 22)	2.2 (23, 24)	-1.0 (18, 18)	-0.1 (21, 21)	0.2 (23, 22)
	re	0.8212	0.9168	0.9974	0.7920	0.8991	0.9970
20	True	0.2823	0.3911	0.4916	0.2823	0.3911	0.4916
	H-L	-0.9 (22, 22)	1.5 (24, 25)	4.1 (25, 26)	-3.1 (21, 21)	-2.4 (23, 23)	0.1 (24, 24)
	Proposed	1.0 (19, 20)	1.1 (22, 23)	4.1 (25, 26)	-2.7 (18, 18)	-2.3 (21, 21)	0.2 (24, 24)
	re	0.7665	0.8575	0.9911	0.7274	0.8257	0.9875
30	True	0.3035	0.4533	0.6712	0.3035	0.4533	0.6712
	H-L	-0.2 (23, 23)	1.8 (26, 26)	0.5 (26, 27)	-3.0 (22, 22)	-1.3 (24, 24)	-0.7 (24, 23)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

	$C_i \sim Unif(0, 150)$			$C_i \sim Unif(0, 75)$			$\bar{y} = \frac{\sum x}{n}$
	5	7	15	5	7	15	
Proposed	0.7 (19, 20)	0.9 (23, 23)	0.6 (26, 26)	-3.7 (18, 18)	-1.3 (21, 20)	-0.5 (23, 23)	
re	0.7285	0.7907	0.9665	0.6810	0.7358	0.9403	

^a Average number of observed infections per subject

^b Average proportion of subjects without any infections

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

Table 2

Summary of the results for Simulation Scenario 2 with true values of the bivariate joint distribution F_{X^0} , $y^0(x, y)$ (True); relative bias $\times 10^3$ (Monte-Carlo $SD \times 10^3$ and average asymptotic $SE \times 10^3$) of Huang-Louis estimator (H-L) and the proposed estimator (Proposed); and relative efficiency (re) of H-L vs. Proposed. Results for each setting are based on 1000 simulated datasets, each with $n = 500$ subjects.

x	y =	$C_i \sim Unif(0, 75)$				$C_i \sim Unif(0, 150)$			
		5	7	15		5	7	15	
$\sigma_0 = 0$									
		$m^d = 3.44; cv^b = 0.34$				$m = 9.10; cv = 0.18$			
15	True	0.1081	0.1671	0.2901		0.1081	0.1671	0.2901	
	H-L	-2.9 (15, 15)	-5.3 (19, 18)	-1.9 (24, 23)		3.3 (15, 15)	0.3 (18, 18)	-1.5 (21, 22)	
	Proposed	-0.0 (14, 14)	-4.4 (17, 17)	-2.4 (23, 23)		3.0 (13, 13)	-0.9 (16, 16)	-1.8 (21, 21)	
	re^c	0.7890	0.8489	0.9552		0.7492	0.8100	0.9301	
20	True	0.1535	0.2368	0.4097		0.1535	0.2368	0.4097	
	H-L	-9.0 (18, 18)	-9.5 (22, 22)	-1.8 (26, 26)		-5.1 (17, 17)	-6.7 (20, 20)	-3.2 (24, 24)	
	Proposed	-6.2 (16, 16)	-8.3 (20, 20)	-2.2 (25, 25)		-5.0 (15, 15)	-7.5 (18, 18)	-3.7 (23, 23)	
	re	0.7808	0.8383	0.9485		0.7373	0.7943	0.9170	
30	True	0.2147	0.3309	0.5730		0.2147	0.3309	0.5730	
	H-L	-6.3 (22, 21)	-7.0 (26, 25)	-0.4 (28, 28)		2.1 (20, 20)	-1.1 (23, 23)	0.3 (25, 24)	
	Proposed	-5.1 (19, 19)	-6.4 (23, 23)	-0.6 (27, 27)		1.1 (17, 17)	-1.7 (20, 20)	-0.1 (23, 23)	
	re	0.7691	0.8214	0.9357		0.7181	0.7678	0.8863	
$\sigma_0 = 0.25$									
		$m = 3.94; cv = 0.34$				$m = 9.85; cv = 0.18$			
15	True	0.1674	0.2313	0.3295		0.1674	0.2313	0.3295	
	H-L	-11.4 (18, 18)	-5.3 (21, 21)	-9.2 (24, 24)		-7.7 (18, 17)	-4.6 (20, 20)	-9.1 (22, 22)	
	Proposed	-10.5 (17, 16)	-5.4 (20, 20)	-9.5 (24, 23)		-8.7 (16, 15)	-4.9 (18, 18)	-9.6 (22, 22)	
	re	0.8166	0.8790	0.9752		0.7849	0.8515	0.9625	
20	True	0.2136	0.3039	0.4544		0.2136	0.3039	0.4544	
	H-L	-6.3 (21, 20)	-1.2 (24, 23)	-7.2 (26, 26)		-2.4 (19, 19)	0.4 (22, 22)	-5.1 (24, 24)	
	Proposed	-4.8 (18, 18)	-1.3 (22, 22)	-7.4 (26, 26)		-2.8 (17, 17)	-0.0 (20, 20)	-5.6 (23, 23)	
	re	0.7938	0.8547	0.9654		0.7556	0.8193	0.9453	
30	True	0.2640	0.3879	0.6165		0.2640	0.3879	0.6165	
	H-L	-8.1 (23, 23)	-0.3 (26, 26)	-4.3 (27, 27)		-2.7 (21, 21)	1.7 (23, 23)	-4.2 (24, 24)	

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

	$C_i \sim Unif(0, 150)$			$C_i \sim Unif(0, 75)$			$\mathbf{x}$	$\mathbf{y} =$
	5	7	15	5	7	15		
Proposed	-6.3 (20, 20)	-0.0 (23, 23)	-4.6 (27, 27)	-3.1 (18, 18)	0.7 (20, 21)	-4.1 (22, 23)		
re	0.7663	0.8215	0.9465	0.7194	0.7725	0.9056		

^a Average number of observed infections per subject

^b Average proportion of subjects without any infections

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

Table 3

Summary of the simulation results for comparing marginal distribution estimators for gap times beyond the first infection. Results for each setting are based on 1000 simulated datasets, each with $n = 500$ subjects.

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

^aMonte-Carlo relative bias $\times 10^{-3}$

^bBased on 95% CI

^c $F_{Y^0|X^0}(y|100)$ based on the proposed method

^dWang-Chang estimator for the 2nd and higher gaps

^eKaplan-Meier estimator for the 2nd gap time data only

Table 4

The estimates of the bivariate joint distribution $F_{X0}, F_0(x, y)$ ($SE \times 10^3$) of time from transplant to the first infection (x) and gap times between two consecutive infections (y).

x (weeks)	y (weeks)				
	1	2	3	4	5
Bacterial infections					
Overall	$m^a = 0.75$; $cr^b = 0.52$; Trend test $p = 0.09$				
1	0.031 (5)	0.051 (6)	0.067 (8)	0.082 (9)	0.092 (9)
2	0.054 (6)	0.084 (8)	0.112 (10)	0.137 (11)	^c
3	0.065 (7)	0.097 (9)	0.131 (11)	-	-
4	0.069 (7)	0.103 (9)	-	-	-
5	0.073 (8)	-	-	-	-
Myeloablative	$m = 0.87$; $cr = 0.47$; Trend test $p = 0.18$				
1	0.046 (8)	0.074 (10)	0.090 (12)	0.104 (13)	0.112 (13)
2	0.073 (10)	0.111 (12)	0.143 (14)	0.169 (15)	-
3	0.084 (10)	0.124 (13)	0.164 (15)	-	-
4	0.090 (11)	0.133 (13)	-	-	-
5	0.091 (11)	-	-	-	-
Non-myeloablative	$m = 0.59$; $cr = 0.60$; Trend test $p = 0.14$				
1	0.009 (4)	0.018 (6)	0.032 (8)	0.049 (11)	0.064 (12)
2	0.026 (7)	0.046 (10)	0.067 (12)	0.092 (14)	-
3	0.036 (8)	0.059 (11)	0.083 (14)	-	-
4	0.039 (9)	0.061 (11)	-	-	-
5	0.047 (10)	-	-	-	-
Viral infections					
Overall	$m = 0.44$; $cr = 0.66$; Trend test $p = 0.14$				
1	0.005 (2)	0.013 (3)	0.015 (4)	0.015 (4)	0.016 (4)
2	0.008 (3)	0.020 (4)	0.026 (5)	0.030 (5)	-
3	0.019 (4)	0.042 (6)	0.049 (7)	-	-
4	0.027 (5)	0.056 (7)	-	-	-

<i>x</i> (weeks)	<i>y</i> (weeks)				
	1	2	3	4	5
5	0.031 (5)	-	-	-	-
CMV seropositive					
	$m \pm 0.56$; $cr = 0.59$; Trend test $p = 0.20$				
1	0.007 (3)	0.017 (5)	0.021 (6)	0.022 (6)	0.023 (6)
2	0.011 (4)	0.028 (7)	0.037 (8)	0.042 (9)	-
3	0.027 (7)	0.058 (10)	0.071 (11)	-	-
4	0.036 (8)	0.077 (11)	-	-	-
5	0.044 (8)	-	-	-	-
CMV seronegative					
	$m \pm 0.28$; $cr = 0.76$; Trend test $p = 0.16$				
1	0.002 (2)	0.007 (4)	0.007 (4)	0.007 (4)	0.007 (4)
2	0.004 (3)	0.011 (5)	0.011 (5)	0.014 (6)	-
3	0.010 (4)	0.020 (7)	0.020 (7)	-	-
4	0.014 (6)	0.028 (8)	-	-	-
5	0.014 (6)	-	-	-	-

^a Average number of observed infections per subject

^b Proportion of subjects without any infections

^c Non-identifiable