Received: 10 January 2017 | Revised: 30 September 2017 | Accepted: 30 October 2017

DOI: 10.1002/sim.7563

RESEARCH ARTICLE

WILEY Statistics in Medicine

Semiparametric regression analysis for alternating recurrent event data

Chi Hyun Lee[1] | Chiung-Yu Huang[2] | Gongjun Xu[3] | Xianghua Luo[4,5]

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

[2]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA 94158, USA

[3]Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

[4]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, USA

[5]Biostatistics Core, Masonic Cancer Center, University of Minnesota, Minneapolis, MN 55455, USA

Correspondence
Chi Hyun Lee, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler St, Unit 1411, Houston, TX 77030, USA.
Email: clee9@mdanderson.org

Funding information
NCI, Grant/Award Number: R01CA193888 and R03CA187991; NIMH, Grant/Award Number: R03MH112895; NSF, Grant/Award Number: SES-1659328 and DMS-1712717; NSA, Grant/Award Number: H98230-17-1-0308

Alternating recurrent event data arise frequently in clinical and epidemiologic studies, where 2 types of events such as hospital admission and discharge occur alternately over time. The 2 alternating states defined by these recurrent events could each carry important and distinct information about a patient's underlying health condition and/or the quality of care. In this paper, we propose a semiparametric method for evaluating covariate effects on the 2 alternating states jointly. The proposed methodology accounts for the dependence among the alternating states as well as the heterogeneity across patients via a frailty with unspecified distribution. Moreover, the estimation procedure, which is based on smooth estimating equations, not only properly addresses challenges such as induced dependent censoring and intercept sampling bias commonly confronted in serial event gap time data but also is more computationally tractable than the existing rank-based methods. The proposed methods are evaluated by simulation studies and illustrated by analyzing psychiatric contacts from the South Verona Psychiatric Case Register.

KEYWORDS

accelerated failure time model, alternating renewal process, gap times, recurrent events

1 | INTRODUCTION

Recurrent event data analysis focuses on modeling and estimation of the risk of event occurrence over time and has a wide range of applications in a variety of fields including in reliability, medicine, social sciences, economics, and criminology. In many applications, the study endpoints can be characterized by 2 different alternating events. For example, patients with chronic diseases may be repeatedly admitted to and discharged from hospital, thus creating an alternating sequence of care periods and break periods. In studies of depression, participants may cycle back and forth between periods of normal mood and depressive episodes. Another important example is the relapse phase and the remission phase of a reversible disease, where patients may alternate between the 2 disease states. Such data structure is referred to as *alternating recurrent event* data in this paper to distinguish from its *univariate* counterpart where all recurrent events are of the same type. It is important to point out that the duration of the 2 types of time periods can each

carry distinct information about the underlying health condition of patients and/or the quality of care. For example, a shorter hospital stay can indicate better treatment effect or quality of care, while a short break period would suggest ineffective maintenance strategies in chronically ill patients. Therefore, it is of interest to develop efficient statistical methods that can make full use of the observed data to evaluate the effects of treatment and risk factors on the 2 alternating states.

When the gap time, that is, the duration between consecutive events, is the outcome of interest, it is known that the sequential structure of recurrent events generates analytical challenges.[1,2] For example, because the observable region of the jth gap time ($j \geq 2$) is given by the difference between the overall censoring time and the $j-1$th event time, the second and higher-order gap times are subject to induced dependent censoring as recurrent gap times of the same subject are usually correlated. This is the case even when the overall censoring time is independent of the recurrent event process. In addition, because longer gap times are more likely to be censored, the last censored gap times tend to be longer than the observed uncensored gap times; the phenomenon is known as length bias due to intercept sampling. Finally, the number of gap times is informative about the underlying recurrent event process, as high-risk patients tend to have shorter times between consecutive events, thus more gap times. In the literature, various statistical methods have been developed for analyzing gap time data in the setting of univariate recurrent events. In particular, some authors considered nonparametric estimation of the gap time distribution,[1,3,4] while others have studied various semiparametric regression models for evaluating covariate effects on the gap times.[5-14] Note that the aforementioned methodologies for univariate recurrent events are not directly applicable to analyzing the pooled gap times between alternating recurrent events, as the 2 states of an alternating recurrent event process usually have distinct biological meanings and hence different distributions. It is also not appropriate, as discussed in Yan and Fine,[15] to apply these models to the 2 different types of gap times separately due to the induced dependent censoring. It is theoretically justifiable to apply these methods to the sum of the 2 states; for example, one may consider the elapse times from one hospital admission to the next hospital admission of a patient by ignoring the information about the time of discharge. This simplified approach, however, cannot determine if the covariates are associated with the length of the care periods or the break periods, or both, and thus the rich information available from the alternating recurrent gap time data is not fully used. In fact, a treatment that shortens the care periods and at the same time prolongs the break periods could be deemed as ineffective if the treatment effect is evaluated based on the elapse times between hospital admission times using univariate recurrent gap time methods.

The development of statistical methods for alternating recurrent event data has been scarce. Huang and Wang[2] considered nonparametric estimation of the joint distribution of the 2 alternating states. While nonparametric estimation can serve as a basis for exploring the underlying recurrent event process, regression methods would be more attractive to researchers who are interested in identifying risk factors that are related to the duration of each state. In an early work by Xue and Brookmeyer,[16] a semiparametric bivariate frailty model was proposed for the 2 types of gap times, where a parametric assumption for the joint distribution of the frailties is imposed for deriving maximum likelihood estimator. More recently, Yan and Fine[15] proposed a temporal process regression method focusing on the frequency and the cumulative length of one of the 2 alternating states. Chang[6] considered accelerated failure time (AFT) models for both types of alternating gap times and used a rank-based estimating equation approach for model estimation. However, the rank-based estimating equation approach for AFT models is seldom used in applications due to the lack of efficient and reliable computational methods for obtaining parameter estimates and the corresponding variance estimates.[17,18] The main difficulty in the implementation of rank-based estimation procedure lies in the nonsmoothness of the estimating functions. Unfortunately, the same argument applies to the estimation procedure proposed by Chang,[6] making it less attractive for practical use.

In this paper, we propose a semiparametric estimation approach under the AFT model. We adapt the multi-state model studied by Huang[19] to the first pair of gap times from alternating recurrent gap time data and extend it to include the recurrent pairs using a within-subject averaging technique.[11] The proposed methodology is based on U-statistics that are continuous and compactly differentiable, and as a result, is expected to be more computationally tractable than that proposed by Chang.[6] The remainder of the article is organized as follows. In Section 2, we introduce the data structure and assumptions of the proposed model. In Section 3, we briefly review the estimation method developed by Huang[19] for multi-state data and introduce our proposed method for alternating recurrent events with large sample properties being established. In Section 4, we conduct a series of simulation studies to demonstrate the performance of the proposed method and compare it with the rank-based estimation procedures proposed by Chang.[6] Application of our proposed method to a psychiatric case register data is presented in Section 5. Some concluding remarks can be found in Section 6.

2 | THE MODEL

To facilitate our discussion, we take the alternating sequence of care and break periods in hospitalization data as an example. Suppose that a group of patients are recruited to a study when they are admitted to a hospital due to a certain disease and followed up on any recurrent hospitalizations due to the same disease until the end of the study. In the absence of censoring, we denote the duration of the care and break periods due to the jth hospitalization episode of subject i as X_{ij}^0 and Y_{ij}^0, respectively, then the recurrent hospitalization process of subject i's can be denoted by $N_i = \{(X_{i1}^0, Y_{i1}^0), (X_{i2}^0, Y_{i2}^0), \dots\}, i = 1, \dots, n$. Let a $p \times 1$ vector $\mathbf{A}_i$ denote the baseline covariates and $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2})^\intercal$ a subject-specific latent vector. We assume that conditioning on $\mathbf{A}_i$ and $\boldsymbol{\gamma}_i$, the bivariate pairs $(X_{ij}^0, Y_{ij}^0), j = 1, 2, \dots$, are independently and identically distributed (i.i.d.) within subject i. Thus, the pairs of durations of the care and break periods can be viewed as an alternating renewal process[20] given the baseline covariates and the latent random vector.

To assess the association between covariates and the lengths of care and break periods, we assume that each period is linearly related to covariates in the logarithmic scale:

$$\log X_{ij}^0 = \gamma_{i1} + \mathbf{A}_i^\intercal \boldsymbol{\beta}_1 + \epsilon_{ij1}, \tag{1}$$

$$\log Y_{ij}^0 = \gamma_{i2} + \mathbf{A}_i^\intercal \boldsymbol{\beta}_2 + \epsilon_{ij2}, \tag{2}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the regression coefficients for the care and break periods, respectively; and $\epsilon_{ijk}, i = 1, \dots, n, j = 1, 2, \dots$, and $k = 1, 2$, are mutually independent random errors with mean zero. The distributions of $\boldsymbol{\gamma}_i$ and ϵ_{ijk} are left unspecified. The latent vector $\boldsymbol{\gamma}_i$ characterizes the correlation among the gap times within a subject. Specifically, the association between X_{ij}^0 and Y_{ij}^0 is characterized by the correlation of γ_{i1} and γ_{i2}, whereas the variances of γ_{i1} and γ_{i2} account for the degree of association within the same type of gap times, $X_{ij}^{0\prime}$s and $Y_{ij}^{0\prime}$s, respectively.

Let C_i denote the censoring time of the ith subject. Suppose C_i has a survival function $G(\cdot)$ with a maximum support τ_C defined by $\tau_C = \sup\{t : G(t) > 0\}$. We assume that the censoring time C_i is independent of N_i, $\mathbf{A}_i$, and $\boldsymbol{\gamma}_i$. Denote by m_i the number of observed (censored or uncensored) episodes of bivariate pairs, so that m_i satisfies

$$\sum_{j=1}^{m_i-1} (X_{ij}^0 + Y_{ij}^0) \le C_i, \qquad \sum_{j=1}^{m_i} (X_{ij}^0 + Y_{ij}^0) > C_i.$$

By definition, the observation of the m_ith pair of gap times is always incomplete and the gap times of a lower order, that is, (X_{ij}^0, Y_{ij}^0) for $j = 1, \dots, m_i - 1$, are observed completely if $m_i > 1$. Although the duration of the first care period X_{i1}^0 is subject to independent censoring C_i, the second and higher-order gap times, $X_{ij}^0, j > 1$ and $Y_{ij}^0, j \ge 1$ are likely to be dependent on their corresponding censoring times, $\max\{C_i - \sum_{l=1}^{j-1}(X_{il}^0 + Y_{il}^0), 0\}$ and $\max\{C_i - \sum_{l=1}^{j-1}(X_{il}^0 + Y_{il}^0) - X_{ij}^0, 0\}$ where $\sum_1^0 = 0$, respectively. Hence, it is not appropriate to naively apply clustered survival data methods[21] on the pooled recurrent gap times since the clustered survival data methods typically require that the times from the same cluster are all subject to independent censoring. Moreover, m_i is informative of the underlying distribution of the elapse times between 2 adjacent hospital admissions.

A typical recurrent hospitalization process is illustrated in Figure 1, where the censoring time for the care period of the last hospitalization of the ith subject is denoted by $C_i^* = C_i - \sum_{j=1}^{m_i-1}(X_{ij}^0 + Y_{ij}^0)$. Due to right censoring, the observed data of subject i are $\{(X_{ij}, Y_{ij}, \Delta_{ij}^X, \Delta_{ij}^Y), j = 1, \dots, m_i\}$ where $X_{ij} = X_{ij}^0$, $Y_{ij} = Y_{ij}^0$, and $\Delta_{ij}^X = \Delta_{ij}^Y = 1$ for $j < m_i$; and $X_{im_i} = \min(X_{im_i}^0, C_i^*)$, $Y_{im_i} = \min\{Y_{im_i}^0, \max(C_i^* - X_{im_i}^0, 0)\}$, $\Delta_{im_i}^X = I(X_{im_i}^0 < C_i^*)$, and $\Delta_{im_i}^Y = 0$. The break period in the m_ith hospitalization is always censored and can be unobserved if censoring occurs during the care period.

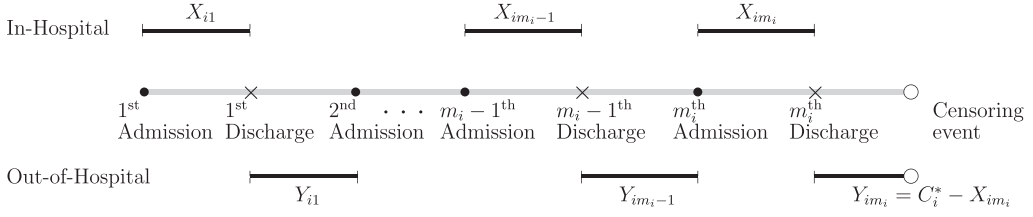

FIGURE 1 An illustration of a typical alternating recurrent event process

3 | ESTIMATION METHODS

3.1 | A brief review of an existing method for bivariate nonrecurrent gap time data

We first consider model estimation based on the first bivariate gap time pairs $\{(X_{i1}, Y_{i1}, \Delta_{i1}^X, \Delta_{i1}^Y); i = 1, \ldots, n\}$ by adapting the methods for multi-state model developed by Huang[19] to our data structure. For the sake of simplicity, we suppress the order of gap time pairs and use $(X_i, Y_i, \Delta_i^X, \Delta_i^Y)$ in notation to denote the first pair throughout this section.

Model (1) implies that, given the covariate values $\mathbf{A}_i$ and $\mathbf{A}_{i'}$ for any 2 subjects i and i', the 2 transformed random variables $\log X_i^0 - \mathbf{A}_i^\intercal \boldsymbol{\beta}_1 + \mathbf{A}_{i'}^\intercal \boldsymbol{\beta}_1$ and $\log X_{i'}^0$ have the same distribution. Define the transformed gap time $X_{ii'}^0(\mathbf{b}_1) = \exp(\mathbf{A}_{ii'}^\intercal \mathbf{b}_1) X_i^0$, where $\mathbf{A}_{ii'} = \mathbf{A}_{i'} - \mathbf{A}_i$ is the contrast between subjects i and i' in terms of baseline covariates. It is easy to see that, given $\mathbf{A}_i$ and $\mathbf{A}_{i'}$, $\{X_i^0, X_{ii'}^0(\boldsymbol{\beta}_1)\}$ and $\{X_{i'i}^0(\boldsymbol{\beta}_1), X_{i'}^0\}$ have the same joint distribution when $\boldsymbol{\beta}_1$ is the true regression parameter. Let $O_L(\cdot, \cdot)$ be a symmetric, continuous function on $\{(t, s) : 0 \le t \le L, 0 \le s \le L\}$, where $O_L(s, t)$ is monotonic in t given s and vice versa. By symmetry of O_L, we have

$$\mathrm{E}\left[\mathbf{A}_{ii'} O_L\{X_i^0, X_{ii'}^0(\boldsymbol{\beta}_1)\}\right] = 0. \tag{3}$$

Next, we define $Z_i^0 = X_i^0 + Y_i^0$, the elapse time between the first 2 consecutive hospital admissions, and the transformed gap time $Z_{ii'}^0(\mathbf{b}) = \exp(\mathbf{A}_{ii'}^\intercal \mathbf{b}_1) X_i^0 + \exp(\mathbf{A}_{ii'}^\intercal \mathbf{b}_2) Y_i^0$, where $\mathbf{b} = (\mathbf{b}_1^\intercal, \mathbf{b}_2^\intercal)^\intercal$. Arguing as before, conditional on $\mathbf{A}_i$ and $\mathbf{A}_{i'}$, $\{Z_i^0, Z_{ii'}^0(\boldsymbol{\beta})\}$ and $\{Z_{i'i}^0(\boldsymbol{\beta}), Z_{i'}^0\}$ share the same joint distribution, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\intercal, \boldsymbol{\beta}_2^\intercal)^\intercal$ are the true regression parameters. Then, it follows that

$$\mathrm{E}\left[\mathbf{A}_{ii'} O_L\{Z_i^0, Z_{ii'}^0(\boldsymbol{\beta})\}\right] = 0. \tag{4}$$

In the absence of censoring, estimating equations using observed data can be constructed directly based on (3) and (4). Under right censoring, we define $Z_i = \min(Z_i^0, C_i)$ the time from the first hospital admission to the next hospital admission or censoring. Analogously, the observed counterparts of $X_{ii'}^0$ and $Z_{ii'}^0$ are defined as $X_{ii'}(\mathbf{b}_1) = \exp(\mathbf{A}_{ii'}^\intercal \mathbf{b}_1) X_i$ and $Z_{ii'}(\mathbf{b}) = \exp(\mathbf{A}_{ii'}^\intercal \mathbf{b}_1) X_i + \exp(\mathbf{A}_{ii'}^\intercal \mathbf{b}_2) Y_i$, respectively. Under the independent censoring assumption, we derive

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

$$\mathrm{E}\left[\frac{\Delta_i^Y O_L\{Z_i, Z_{ii'}(\boldsymbol{\beta})\}}{G(Z_i \wedge L)} | \mathbf{A}_i, \mathbf{A}_{i'}\right] = \mathrm{E}\left[O_L\{Z_i^0, Z_{ii'}^0(\boldsymbol{\beta})\} | \mathbf{A}_i, \mathbf{A}_{i'}\right],$$

by the idea of inverse probability of censoring weights, where $a \wedge b = \min(a, b)$. Following (3) and (4), unconditional on $\mathbf{A}_i$ and $\mathbf{A}_{i'}$, we have

$$\mathrm{E}\left[\mathbf{A}_{ii'} \frac{\Delta_i^X O_L\{X_i, X_{ii'}(\boldsymbol{\beta}_1)\}}{G(X_i \wedge L)}\right] = 0, \tag{5}$$

$$\mathrm{E}\left[\mathbf{A}_{ii'} \frac{\Delta_i^Y O_L\{Z_i, Z_{ii'}(\boldsymbol{\beta})\}}{G(Z_i \wedge L)}\right] = 0. \tag{6}$$

Then, a system of estimating functions can be constructed with the observed bivariate gap times:

$$D_1(\mathbf{b}_1) = n^{-2} \sum_{i=1}^n \left[\sum_{i'=1}^n \mathbf{A}_{ii'} \frac{\Delta_i^X O_{L_1}\{X_i, X_{ii'}(\mathbf{b}_1)\}}{\hat{G}_1(X_i \wedge L_1)}\right], \tag{7}$$

$$D_2(\mathbf{b}) = n^{-2} \sum_{i=1}^n \left[\sum_{i'=1}^n \mathbf{A}_{ii'} \frac{\Delta_i^Y O_{L_2}\{Z_i, Z_{ii'}(\mathbf{b})\}}{\hat{G}_2(Z_i \wedge L_2)}\right], \tag{8}$$

where $\hat{G}_1$ and $\hat{G}_2$ are the Kaplan-Meier estimators of the survival function $G(\cdot)$ based on the data $\{(X_i, 1-\Delta_i^X), i = 1, \ldots, n\}$ and $\{(Z_i, 1-\Delta_i^Y), i = 1, \ldots, n\}$, respectively. As pointed out in Huang,[19] $\hat{G}_2$ can be used in (7) in place of $\hat{G}_1$, but this often

leads to a greater variance of $D_1(\mathbf{b}_1)$. The limits $L_1 < \tau_C$ and $L_2 < \tau_C$ are imposed to address the problem of X_i^0 and Z_i^0 having maximum support greater than τ_C. One can inductively solve the estimating equations $D_1(\mathbf{b}_1) = 0$ and $D_2(\mathbf{b}) = 0$ to obtain the estimates for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Conventional methods for survival analysis under the AFT model (see Kalbfleisch and Prentice[22] and reference therein) are not directly applicable to the estimation of Model (2). In our setting, the break period Y_i^0 is subject to induced dependent censoring because it is censored by $\max(C_i - X_i^0, 0)$, which is informative due to the correlation between X_i^0 and Y_i^0. By considering the elapse time between consecutive admissions Z_i^0 and the sum of transformed care and break periods $Z_{ii'}^0$ instead of the break period Y_i^0 solely, we circumvent the induced dependent censoring issue.

3.2 | The proposed estimation method

We now extend the method in Section 3.1 to deal with alternating recurrent gap time data. As pointed out in Huang and Wang,[2] the m_ith pair of gap times tends to be longer than the uncensored pairs of gap times due to bias induced by intercept sampling (also see Cox[20], p65 for the example of textile fiber sampling). As a result, naively including all observed data in the estimation procedure usually leads to inconsistent estimation. In this section, we extend the method for multi-state models proposed by Huang,[19] which was reviewed in Section 3.1, to the setting of alternating recurrent events.

Define $m_i^* = \max(m_i - 1, 1)$. Thus, for patients with no completely observed bivariate gap time pairs, $m_i^* = 1$; for patients with at least one completely observed gap time pair, m_i^* is the number of complete pairs. Let the elapse time between 2 consecutive hospital admissions be denoted as $Z_{ij}^0 = X_{ij}^0 + Y_{ij}^0$ and its observed counterparts as $Z_{ij} = X_{ij} + Y_{ij}$, for $j = 1, \ldots, m_i^*$. The observed transformed times are defined as

$$X_{ii'j}(\mathbf{b}_1) = \exp(\mathbf{A}_{ii'}^\intercal \mathbf{b}_1) X_{ij},$$
$$Z_{ii'j}(\mathbf{b}) = \exp(\mathbf{A}_{ii'}^\intercal \mathbf{b}_1) X_{ij} + \exp(\mathbf{A}_{ii'}^\intercal \mathbf{b}_2) Y_{ij}, j = 1, \ldots, m_i^*.$$

Under our model assumption, conditioning on m_i, $\boldsymbol{\gamma}_i$, and $\mathbf{A}_i$, the observed bivariate pairs, $(X_{i1}, Y_{i1}), \ldots, (X_{im_i^*}, Y_{im_i^*})$ are i.i.d. when $m_i \geq 2$. Thus, replacing $\{X_i, X_{ii'}(\boldsymbol{\beta}_1), Z_i, Z_{ii'}(\boldsymbol{\beta})\}$ with $\{X_{ij}, X_{ii'j}(\boldsymbol{\beta}_1), Z_{ij}, Z_{ii'j}(\boldsymbol{\beta})\}$ for any $j = 1, \ldots, m_i^*$ in (5) and (6) should give unbiased estimating equations. We propose to apply the idea of weighted risk-set method[11] to assign a weight $1/m_i^*$ to each pair of bivariate gap times and sum over $j = 1, \ldots, m_i^*$ to construct more efficient estimating functions. Specifically, arguing as in Luo and Huang,[11] we can prove that the weighted averages of $O_{L_1}(\cdot,\cdot)$ and $O_{L_2}(\cdot,\cdot)$ over the conditional i.i.d. bivariate pairs have the same expectations as their counterparts for the first bivariate gap time pair only data:

$$\mathrm{E}\left(\frac{1}{m_i^*} \sum_{j=1}^{m_i^*} \mathrm{E}\left[\frac{\Delta_{ij}^X O_L\{X_{ij}, X_{ii'j}(\boldsymbol{\beta}_1)\}}{G(X_{ij} \wedge L)} | m_i^*, \boldsymbol{\gamma}_i, \mathbf{A}_i, \mathbf{A}_{i'} \right] | \mathbf{A}_i, \mathbf{A}_{i'} \right) = \mathrm{E}\left[\frac{\Delta_{i1}^X O_L\{X_{i1}, X_{ii'1}(\boldsymbol{\beta}_1)\}}{G(X_{i1} \wedge L)} | \mathbf{A}_i, \mathbf{A}_{i'} \right],$$
$$\mathrm{E}\left(\frac{1}{m_i^*} \sum_{j=1}^{m_i^*} \mathrm{E}\left[\frac{\Delta_{ij}^Y O_L\{Z_{ij}, Z_{ii'j}(\boldsymbol{\beta})\}}{G(Z_{ij} \wedge L)} | m_i^*, \boldsymbol{\gamma}_i, \mathbf{A}_i, \mathbf{A}_{i'} \right] | \mathbf{A}_i, \mathbf{A}_{i'} \right) = \mathrm{E}\left[\frac{\Delta_{i1}^Y O_L\{Z_{i1}, Z_{ii'1}(\boldsymbol{\beta})\}}{G(Z_{i1} \wedge L)} | \mathbf{A}_i, \mathbf{A}_{i'} \right].$$

It follows directly that

$$\mathrm{E}\left[\mathbf{A}_{ii'} \frac{1}{m_i^*} \sum_{j=1}^{m_i^*} \frac{\Delta_{ij}^X O_L\{X_{ij}, X_{ii'j}(\boldsymbol{\beta}_1)\}}{G(X_{ij} \wedge L)} \right] = 0, \tag{9}$$

$$\mathrm{E}\left[\mathbf{A}_{ii'} \frac{1}{m_i^*} \sum_{j=1}^{m_i^*} \frac{\Delta_{ij}^Y O_L\{Z_{ij}, Z_{ii'j}(\boldsymbol{\beta})\}}{G(Z_{ij} \wedge L)} \right] = 0. \tag{10}$$

By using data only up to the m_i^*th pair in the above formulation for subjects who have at least one completely observed gap time pair, we exclude the potentially longer gap time pairs and avoid intercept sampling bias. Hence, X_{im_i}, either censored or uncensored, and Y_{im_i}, which is always censored, are not used for such subjects in (9) and (10). For subjects who have no completely observed gap time pairs, we only use their data for constructing consistent estimators for $G(\cdot)$ if the first

gap time is censored ($\Delta_{i1}^X = 0$), otherwise ($\Delta_{i1}^X = 1$) the data of such subjects are used in both the estimation of $G(\cdot)$ and the numerator in the expectation in (9). Therefore, we can construct a system of estimating functions as follows:

$$D_1^*(\mathbf{b}_1) = n^{-2}\sum_{i=1}^{n}\left[\sum_{i'=1}^{n}\mathbf{A}_{ii'}\frac{1}{m_i^*}\sum_{j=1}^{m_i^*}\frac{\Delta_{ij}^X O_{L_1}\{X_{ij}, X_{ii'j}(\mathbf{b}_1)\}}{\hat{G}_1(X_{ij}\wedge L_1)}\right], \tag{11}$$

$$D_2^*(\mathbf{b}) = n^{-2}\sum_{i=1}^{n}\left[\sum_{i'=1}^{n}\mathbf{A}_{ii'}\frac{1}{m_i^*}\sum_{j=1}^{m_i^*}\frac{\Delta_{ij}^Y O_{L_2}\{Z_{ij}, Z_{ii'j}(\mathbf{b})\}}{\hat{G}_2(Z_{ij}\wedge L_2)}\right], \tag{12}$$

where $\hat{G}_1$ and $\hat{G}_2$ are Kaplan-Meier estimators based on the first pair of bivariate gap times (censored or uncensored) as in (7) and (8). The proposed estimators $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ can be obtained by inductively solving $D_1^*(\mathbf{b}_1) = 0$ and $D_2^*\{(\hat{\boldsymbol{\beta}}_1^\intercal, \mathbf{b}_2^\intercal)^\intercal\} = 0$. Following Huang,[19] we choose $O_L(t,s) = \log[\min\{\max(t,s), L\}] - \log(L)$ to yield monotonic estimating functions which guarantee a unique solution. Further discussions on the selection of O_L can be found in Huang.[19] Compared with the method for bivariate nonrecurrent gap time data reviewed in Section 3.1, the proposed estimation method is expected to be more efficient because the information beyond the second hospital admission time of each patient is used.

3.3 | Asymptotic properties

In this section, we establish the consistency and the asymptotic normality of the proposed estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^\intercal, \hat{\boldsymbol{\beta}}_2^\intercal)^\intercal$. Following Huang,[19] we begin by rewriting the estimating functions (7) and (8) as

$$D_1(\mathbf{b}_1) = \int_{t,s,\mathbf{a}_1,\mathbf{a}_2}(\mathbf{a}_2 - \mathbf{a}_1)\frac{O_{L_1}(t,s)}{\hat{G}_1(t\wedge L_1)}\hat{F}_1(\mathrm{d}t, \mathrm{d}s, \mathrm{d}\mathbf{a}_1; \mathbf{a}_2, \mathbf{b}_1)\hat{H}(\mathrm{d}\mathbf{a}_2), \tag{13}$$

$$D_2(\mathbf{b}) = \int_{t,s,\mathbf{a}_1,\mathbf{a}_2}(\mathbf{a}_2 - \mathbf{a}_1)\frac{O_{L_2}(t,s)}{\hat{G}_2(t\wedge L_2)}\hat{F}_2(\mathrm{d}t, \mathrm{d}s, \mathrm{d}\mathbf{a}_1; \mathbf{a}_2, \mathbf{b})\hat{H}(\mathrm{d}\mathbf{a}_2), \tag{14}$$

where $\hat{F}_1$, $\hat{F}_2$, and $\hat{H}$ are the empirical estimators of

$$F_1(t,s,\mathbf{a}_1;\mathbf{a}_2,\mathbf{b}_1) = \Pr\left[X_{i1}\le t, \exp\{(\mathbf{a}_2 - \mathbf{A}_i)^\intercal\mathbf{b}_1\}X_{i1}\le s, \mathbf{A}_i\le\mathbf{a}_1, \Delta_{i1}^X = 1\right],$$
$$F_2(t,s,\mathbf{a}_1;\mathbf{a}_2,\mathbf{b}) = \Pr\left[Z_{i1}\le t, \exp\{(\mathbf{a}_2 - \mathbf{A}_i)^\intercal\mathbf{b}_1\}X_{i1} + \exp\{(\mathbf{a}_2 - \mathbf{A}_i)^\intercal\mathbf{b}_2\}Y_{i1}\le s, \mathbf{A}_i\le\mathbf{a}_1, \Delta_{i2}^Y = 1\right],$$

and $H(\mathbf{a}_2) = \Pr(\mathbf{A}_i\le\mathbf{a}_2)$, respectively. Note that $\Pr(\mathbf{A}_i\le\mathbf{a}) = \Pr(A_{i1}\le a_1, \ldots, A_{ip}\le a_p)$, where $\mathbf{A}_i = (A_{i1}, \ldots, A_{ip})^\intercal$ and $\mathbf{a} = (a_1, \ldots, a_p)^\intercal$. Huang[19] showed that D_1 and D_2 are continuous and compactly differentiable functionals through the properties of the components, $\hat{G}_1$, $\hat{G}_2$, $\hat{F}_1$, $\hat{F}_2$, and $\hat{H}$. Based on the re-expression in (13) and (14), both $D_1^\intercal(\mathbf{b}_1)(\mathbf{b}_1 - \boldsymbol{\beta}_1)$ and $D_2^\intercal(\mathbf{b})(\mathbf{b}_2 - \boldsymbol{\beta}_2)$ converge almost surely and uniformly in $\mathbf{b}_1$ and in $\mathbf{b}$ to

$$\mathrm{E}\left[\mathbf{A}_{ii'}^\intercal(\mathbf{b}_1 - \boldsymbol{\beta}_1)O_{L_1}\{X_{i1}^0, X_{ii'1}^0(\mathbf{b}_1)\}\right] \quad \text{and} \tag{15}$$

$$\mathrm{E}\left[\mathbf{A}_{ii'}^\intercal(\mathbf{b}_2 - \boldsymbol{\beta}_2)O_{L_2}\{Z_{i1}^0, Z_{ii'1}^0(\mathbf{b})\}\right], \tag{16}$$

respectively. It can be shown that the estimating functions D_1^* and D_2^* in (11) and (12) converge uniformly to the same limit as D_1 and D_2, respectively. Thus, it follows that $D_1^{*\intercal}(\mathbf{b}_1)(\mathbf{b}_1 - \boldsymbol{\beta}_1)$ and $D_2^{*\intercal}(\mathbf{b})(\mathbf{b}_2 - \boldsymbol{\beta}_2)$ also converge almost surely to (15) and (16). Since (15) equals 0 when $\mathbf{b}_1 = \boldsymbol{\beta}_1$, $\hat{\boldsymbol{\beta}}_1$ is consistent for $\boldsymbol{\beta}_1$. Given the consistency of $\hat{\boldsymbol{\beta}}_1$, the consistency of $\hat{\boldsymbol{\beta}}_2$ follows from the fact that (16) equals 0 when $\mathbf{b}_2 = \boldsymbol{\beta}_2$.

To prove the asymptotic normality of $\hat{\boldsymbol{\beta}}$, it suffices to establish the asymptotic normality and linearity of $D^*(\boldsymbol{\beta}) = \{D_1^{*\intercal}(\boldsymbol{\beta}_1), D_2^{*\intercal}(\boldsymbol{\beta})\}^\intercal$. Huang[19] showed that $n^{1/2}D(\boldsymbol{\beta})$ is asymptotically normal with mean zero and variance Ω using the compact differentiability of (13) and (14), where $D(\boldsymbol{\beta}) = \{D_1^\intercal(\boldsymbol{\beta}_1), D_2^\intercal(\boldsymbol{\beta})\}^\intercal$. For the variance, we define

$$\xi_{i1}(\boldsymbol{\beta}_1) = n^{-3/2}\sum_{i'=1}^{n}\mathbf{A}_{ii'}\left[\frac{\Delta_{i1}^{X}O_{L_1}\{X_{i1},X_{ii'1}(\boldsymbol{\beta}_1)\}}{\hat{G}_1(X_{i1}\wedge L_1)} - \frac{\Delta_{i'1}^{X}O_{L_1}\{X_{i'1},X_{i'i1}(\boldsymbol{\beta}_1)\}}{\hat{G}_1(X_{i'1}\wedge L_1)}\right] + n^{-3/2}\int_0^{L_1}\frac{U_1(t,\boldsymbol{\beta}_1)\hat{G}_1(t-)}{R_1(t)\hat{G}_1(t)}\mathrm{d}\hat{M}_{i1}(t),$$

$$\xi_{i2}(\boldsymbol{\beta}) = n^{-3/2}\sum_{i'=1}^{n}\mathbf{A}_{ii'}\left[\frac{\Delta_{i1}^{Y}O_{L_2}\{Z_{i1},Z_{ii'1}(\boldsymbol{\beta})\}}{\hat{G}_2(Z_{i1}\wedge L_2)} - \frac{\Delta_{i'1}^{Y}O_{L_2}\{Z_{i'1},Z_{i'i1}(\boldsymbol{\beta})\}}{\hat{G}_2(Z_{i'1}\wedge L_2)}\right] + n^{-3/2}\int_0^{L_2}\frac{U_2(t,\boldsymbol{\beta})\hat{G}_2(t-)}{R_2(t)\hat{G}_2(t)}\mathrm{d}\hat{M}_{i2}(t),$$

in which

$$U_1(t,\boldsymbol{\beta}_1) = \sum_{i=1}^{n}\sum_{i'=1}^{n}\mathbf{A}_{ii'}\left[\frac{\Delta_{i1}^{X}O_{L_1}\{X_{i1},X_{ii'1}(\boldsymbol{\beta}_1)\}}{\hat{G}_1(X_{i1}\wedge L_1)}I(X_{i1}>t)\right],$$

$$U_2(t,\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{i'=1}^{n}\mathbf{A}_{ii'}\left[\frac{\Delta_{i1}^{Y}O_{L_2}\{Z_{i1},Z_{ii'1}(\boldsymbol{\beta})\}}{\hat{G}_2(Z_{i1}\wedge L_2)}I(Z_{i1}>t)\right],$$

$R_1(t)=\sum_{i=1}^{n}I(X_{i1}\geq t)$, $R_2(t)=\sum_{i=1}^{n}I(Z_{i1}\geq t)$, $\hat{M}_{i1}(t)=I(X_{i1}\leq t,\Delta_{i1}^{X}=0)-\int_0^t I(X_{i1}\geq s)\mathrm{d}\hat{\Lambda}_1(s)$, $\hat{M}_{i2}(t)=I(Z_{i1}\leq t,\Delta_{i1}^{Y}=0)-\int_0^t I(Z_{i1}\geq s)\mathrm{d}\hat{\Lambda}_2(s)$, and $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$ are the Nelson-Aalen estimator corresponding to $\hat{G}_1$ and $\hat{G}_2$, respectively. The variance Ω is the limit of $\sum_{i=1}^{n}\{\xi_{i1}^{\intercal}(\boldsymbol{\beta}_1),\xi_{i2}^{\intercal}(\boldsymbol{\beta})\}^{\intercal}\{\xi_{i1}^{\intercal}(\boldsymbol{\beta}_1),\xi_{i2}^{\intercal}(\boldsymbol{\beta})\}$. Now, we show the asymptotic normality of $D^*(\boldsymbol{\beta})$ following the approach in Huang.[19] We note that D_1^* and D_2^* are continuous and compactly differentiable. By applying the functional delta method and the influence function approach, $n^{1/2}D^*(\boldsymbol{\beta})$ converges weakly to a normal distribution with mean zero and variance Ω^*. Define

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

$$\xi_{i2}^{*}(\boldsymbol{\beta}) = n^{-3/2}\sum_{i'=1}^{n}\mathbf{A}_{ii'}\left[\frac{1}{m_i^*}\sum_{j=1}^{m_i^*}\frac{\Delta_{ij}^{Y}O_{L_2}\{Z_{ij},Z_{ii'j}(\boldsymbol{\beta})\}}{\hat{G}_2(Z_{ij}\wedge L_2)} - \frac{1}{m_{i'}^*}\sum_{l=1}^{m_{i'}^*}\frac{\Delta_{i'l}^{Y}O_{L_2}\{Z_{i'l},Z_{i'il}(\boldsymbol{\beta})\}}{\hat{G}_2(Z_{i'l}\wedge L_2)}\right] + n^{-3/2}\int_0^{L_2}\frac{U_2^*(t,\boldsymbol{\beta})\hat{G}_2(t-)}{R_2^*(t)\hat{G}_2(t)}\mathrm{d}\hat{M}_{i2}^*(t),$$

in which,

$$U_1^*(t,\beta_1) = \sum_{i=1}^{n}\sum_{i'=1}^{n}\mathbf{A}_{ii'}\left[\frac{1}{m_i^*}\sum_{j=1}^{m_i^*}\frac{\Delta_{ij}^{X}O_{L_1}\{X_{ij},X_{ii'j}(\boldsymbol{\beta}_1)\}}{\hat{G}_1(X_{ij}\wedge L_1)}I(X_{ij}>t)\right],$$

$$U_2^*(t,\beta) = \sum_{i=1}^{n}\sum_{i'=1}^{n}\mathbf{A}_{ii'}\left[\frac{1}{m_i^*}\sum_{j=1}^{m_i^*}\frac{\Delta_{ij}^{Y}O_{L_2}\{Z_{ij},Z_{ii'j}(\boldsymbol{\beta})\}}{\hat{G}_2(Z_{ij}\wedge L_2)}I(Z_{ij}>t)\right],$$

$R_1^*(t)=\sum_{i=1}^{n}\sum_{j=1}^{m_i^*}I(X_{ij}\geq t)/m_i^*$, $R_2^*(t)=\sum_{i=1}^{n}\sum_{j=1}^{m_i^*}I(Z_{ij}\geq t)/m_i^*$, $\hat{M}_{i1}^*(t)=\sum_{j=1}^{m_i^*}I(X_{ij}\leq t,\Delta_{ij}^{X}=0)/m_i^*-\int_0^t\sum_{j=1}^{m_i^*}I(X_{ij}\geq s)/m_i^*\mathrm{d}\hat{\Lambda}_1(s)$, and $\hat{M}_{i2}^*(t)=\sum_{j=1}^{m_i^*}I(Z_{ij}\leq t,\Delta_{ij}^{Y}=0)/m_i^*-\int_0^t\sum_{j=1}^{m_i^*}I(Z_{ij}\geq s)/m_i^*\mathrm{d}\hat{\Lambda}_2(s)$. By exchangeability, the weighted average ξ_{ik}^*, U_k^*, R_k^*, and M_{ik}^* converge uniformly to the same limit as their counterparts, ξ_{ik}, U_k, R_k, and M_{ik}, for $k=1,2$. Thus, the variance Ω^* can be estimated by $\hat{\Omega}^*=\sum_{i=1}^{n}\{\xi_{i1}^{*\intercal}(\hat{\boldsymbol{\beta}}_1),\xi_{i2}^{*\intercal}(\hat{\boldsymbol{\beta}})\}^{\intercal}\{\xi_{i1}^{*\intercal}(\hat{\boldsymbol{\beta}}_1),\xi_{i2}^{*\intercal}(\hat{\boldsymbol{\beta}})\}$. By the Glivenko-Cantelli theorem in Pollard,[23] $\sum_{i=1}^{n}\{\xi_{i1}^{*\intercal}(\mathbf{b}_1),\xi_{i2}^{*\intercal}(\mathbf{b})\}^{\intercal}\{\xi_{i1}^{*\intercal}(\mathbf{b}_1),\xi_{i2}^{*\intercal}(\mathbf{b})\}$ converges uniformly and almost surely in $\mathbf{b}$ to a limiting function continuous at $\mathbf{b}=\boldsymbol{\beta}$. Hence, the variance estimate $\hat{\Omega}^*$ is consistent for Ω^* given the consistency of $\hat{\boldsymbol{\beta}}$.

The estimating functions (11) and (12) can be rewritten as $\bar{D}_1^*(\mathbf{b}_1)$ and $\bar{D}_2^*(\mathbf{b})$ by replacing $\hat{F}_1$ and $\hat{F}_2$ in (13) and (14) with their weighted counterparts

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

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

respectively. We note that $\bar{D}^*(\mathbf{b}) = \{\bar{D}_1^{*\intercal}(\mathbf{b}_1), \bar{D}_2^{*\intercal}(\mathbf{b})\}^\intercal$ is not everywhere differentiable. Thus, the first-order Taylor expansion cannot be directly used. Instead, we use the generalized law of mean, proposed in Huang,[24] to accommodate the nondifferentiable functions. By applying the generalized law of mean, we have

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

for $\mathbf{b}$ converging to $\boldsymbol{\beta}$, where Σ_β is the limit of the left and right partial derivative of $\bar{D}^*(\boldsymbol{\beta})$. It follows that $D^*(\mathbf{b})$ is asymptotically linear at $\mathbf{b} = \boldsymbol{\beta}$.

The asymptotic normality of $\hat{\boldsymbol{\beta}}$ naturally follows from the asymptotic normality and linearity of $D^*(\boldsymbol{\beta})$. Thus, $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges weakly to a normal distribution with mean zero and variance consistently estimated by $\hat{\Sigma}_{\hat{\beta}}^{-1}\hat{\Omega}^*(\hat{\Sigma}_{\hat{\beta}}^{-1})^\intercal$ where $\hat{\Sigma}_{\hat{\beta}}$ is the derivative matrix of $\bar{D}^*(\mathbf{b})$ evaluated at $\mathbf{b} = \hat{\boldsymbol{\beta}}$.

4 | SIMULATION STUDIES

We conducted a series of simulation studies to assess the performance of the proposed method. For each setting, we simulated 1000 datasets with sample sizes of $n = 150$ and 300 from the assumed models (1) and (2). Two covariates $\mathbf{A} = (A_1, A_2)^\intercal$ are generated from a Bernoulli distribution with probability 0.5 and a uniform distribution (0, 1), respectively. We set the true regression parameters as $\boldsymbol{\beta}_1 = (0.5, 0.5)^\intercal$ and $\boldsymbol{\beta}_2 = (0, -0.5)^\intercal$ to account for the distinct covariate effects on the 2 alternating states. We consider 2 scenarios where (1) the subject-specific latent vector $(\gamma_{i1}, \gamma_{i2})$ follows a bivariate normal distribution with varying levels of correlation; and (2) the latent variables γ_{i1} and γ_{i2} are from different distributions. The error terms ϵ_{ij1} and ϵ_{ij2} are simulated from independent normal distributions with mean zero and variance 0.1. Since the recurrent event process is subject to right censoring, we generate the censoring time C_i from a uniform distribution that yields 15% or 30% of subjects to have their first bivariate gap time pairs censored on average. Under each setting, we evaluate the performance of the proposed method relative to the rank-based method by Chang[6] (referred to as Chang's method). For the latter method, we present the perturbation-based variance estimates adopted in the original paper.[6] For both methods, we present the mean of the point estimates (Mean), the empirical standard deviation of the point estimates (SD), the empirical average of the standard error estimates (SE), and the coverage probability based on the 95% confidence intervals (CP).

4.1 | Simulation scenario 1

In the first scenario, we generate the subject-specific latent vector $(\gamma_{i1}, \gamma_{i2})$ from a bivariate normal distribution with unit mean vector and variance-covariance matrix $\begin{pmatrix} 0.5 & 0.5\rho \\ 0.5\rho & 0.5 \end{pmatrix}$. We consider $\rho = 1, 0.5$, and 0. When $\rho = 1$, the 2 latent variables $\gamma_{i1} = \gamma_{i2}$; when $\rho = 0$, γ_{i1} and γ_{i2} are independent. The simulation results are summarized in the upper panels of Tables 1 and 2 for sample sizes of $n = 150$ and 300, respectively. The proposed estimator provides virtually unbiased point estimates, and the SE's are close to the SD's across all settings. The CP's are reasonably close to the nominal level. We observe that the SD's and SE's increase as the censoring rate increases from 15% to 30% because fewer bivariate gap time pairs are observed. We note that the level of association between alternating gap times has little impact on the point estimation or the variance estimation. As expected, the variance decreases with the sample size.

As discussed earlier, the point estimation and the resampling-based variance estimation with rank-based, nonsmooth estimating equations tend to be unstable.[25] Under our simulation settings, the proportion of datasets that converged for

TABLE 1 Summary statistics for the simulation study with $n = 150$

			Proposed method		Chang's method	
			β_1	β_2	β_1	β_2
			A_1, A_2	A_1, A_2	A_1, A_2	A_1, A_2
ρ	cr	True	0.5,0.5	0.0,−0.5	0.5,0.5	0.0,−0.5
Scenario 1						
				$\bar{m} = 6.36$		
1	15%	Mean	0.495,0.494	−0.031, −0.508	0.486,0.521	−0.004, −0.408
		SD	0.138,0.262	0.223,0.367	0.141,0.226	0.152,0.217
		SE	0.140,0.245	0.219,0.363	0.144,0.282	0.170,0.263
		CP	0.952,0.930	0.927,0.924	0.954,0.985	0.956,0.961
				$\bar{m} = 3.31$		
	30%	Mean	0.501,0.490	−0.026, −0.505	0.487,0.518	−0.014, −0.404
		SD	0.154,0.271	0.246,0.403	0.138,0.235	0.177,0.227
		SE	0.150,0.263	0.243,0.405	0.155,0.304	0.207,0.279
		CP	0.942,0.934	0.918,0.915	0.949,0.973	0.967,0.958
				$\bar{m} = 5.76$		
0.5	15%	Mean	0.500,0.501	−0.004, −0.494	0.489,0.511	−0.003, −0.414
		SD	0.139,0.253	0.204,0.368	0.141,0.230	0.160,0.210
		SE	0.140,0.245	0.208,0.352	0.140,0.274	0.176,0.251
		CP	0.958,0.940	0.934,0.925	0.947,0.959	0.966,0.952
				$\bar{m} = 3.05$		
	30%	Mean	0.499,0.504	−0.016, −0.511	0.487,0.514	−0.006, −0.404
		SD	0.148,0.265	0.237,0.395	0.139,0.244	0.181,0.236
		SE	0.150,0.263	0.235,0.393	0.149,0.295	0.215,0.273
		CP	0.941,0.949	0.929,0.927	0.957,0.960	0.981,0.941
				$\bar{m} = 5.28$		
0	15%	Mean	0.498,0.501	−0.010, −0.490	0.487,0.464	0.017, −0.416
		SD	0.145,0.255	0.204,0.369	0.133,0.243	0.154,0.221
		SE	0.140,0.245	0.208,0.352	0.140,0.265	0.179,0.254
		CP	0.938,0.945	0.943,0.928	0.955,0.950	0.975,0.945
				$\bar{m} = 2.81$		
	30%	Mean	0.501,0.493	−0.013, −0.509	0.477,0.479	0.002, −0.398
		SD	0.149,0.274	0.227,0.405	0.144,0.268	0.177,0.254
		SE	0.149,0.262	0.232,0.392	0.151,0.288	0.205,0.278
		CP	0.961,0.937	0.942,0.914	0.961,0.955	0.966,0.930
Scenario 2						
				$\bar{m} = 5.32$		
-	15%	Mean	0.500,0.497	0.000, −0.506	0.498,0.491	0.015, −0.417
		SD	0.139,0.249	0.237,0.404	0.131,0.259	0.182,0.252
		SE	0.140,0.245	0.233,0.393	0.139,0.261	0.208,0.270
		CP	0.954,0.943	0.930,0.911	0.959,0.934	0.964,0.942
				$\bar{m} = 2.90$		
	30%	Mean	0.505,0.495	−0.010, −0.513	0.480,0.490	0.000, −0.412
		SD	0.149,0.271	0.250,0.437	0.140,0.249	0.195,0.248
		SE	0.149,0.261	0.252,0.424	0.153,0.280	0.222,0.282
		CP	0.947,0.933	0.920,0.912	0.961,0.958	0.955,0.935

True, true coefficients; Mean, empirical average of point estimates; SD, empirical standard deviation of point estimates; SE, empirical average of standard error estimates; CP, coverage probability based on the 95% confidence interval; $\bar{m}$, average number of observed pairs of recurrence times per subject; cr, average proportion of subjects having the first pair censored; ρ, correlation coefficient.

Chang's method is as low as one third to almost one half of the simulated datasets, depending on the different simulation parameters. Note that the summary results in the tables are based on converged datasets only for the point estimation, and the variance estimation is based on converged perturbation samples only. For the converged datasets,

TABLE 2 Summary statistics for the simulation study with $n = 300$

			Proposed method		Chang's method	
			β_1	β_2	β_1	β_2
			A_1, A_2	A_1, A_2	A_1, A_2	A_1, A_2
ρ	cr	True	0.5,0.5	0.0,−0.5	0.5,0.5	0.0,−0.5
Scenario 1						
				$\bar{m} = 6.38$		
1	15%	Mean	0.501,0.495	−0.010, −0.504	0.496,0.518	−0.001, −0.457
		SD	0.095,0.181	0.165,0.286	0.093,0.150	0.107,0.150
		SE	0.101,0.177	0.164,0.275	0.099,0.199	0.120,0.201
		CP	0.958,0.958	0.935,0.904	0.950,0.985	0.973,0.990
				$\bar{m} = 3.25$		
	30%	Mean	0.499,0.498	−0.014, −0.500	0.492,0.501	−0.004, −0.443
		SD	0.106,0.189	0.183,0.303	0.092,0.155	0.124,0.172
		SE	0.108,0.190	0.182,0.305	0.104,0.214	0.149,0.226
		CP	0.956,0.945	0.934,0.924	0.976,0.988	0.966,0.983
				$\bar{m} = 5.79$		
0.5	15%	Mean	0.503,0.498	−0.016, −0.511	0.491,0.496	−0.008, −0.448
		SD	0.099,0.182	0.153,0.278	0.095,0.165	0.104,0.159
		SE	0.101,0.177	0.159,0.267	0.099,0.194	0.120,0.196
		CP	0.957,0.946	0.946,0.918	0.948,0.955	0.961,0.976
				$\bar{m} = 3.06$		
	30%	Mean	0.506,0.490	−0.006, −0.498	0.493,0.496	−0.006, −0.452
		SD	0.108,0.193	0.182,0.297	0.099,0.168	0.130,0.173
		SE	0.108,0.190	0.176,0.301	0.106,0.212	0.148,0.227
		CP	0.943,0.944	0.922,0.937	0.946,0.981	0.965,0.979
				$\bar{m} = 5.21$		
0	15%	Mean	0.501,0.500	−0.006, −0.491	0.496,0.501	−0.001, −0.464
		SD	0.098,0.179	0.153,0.277	0.093,0.171	0.109,0.176
		SE	0.101,0.178	0.158,0.271	0.099,0.189	0.122,0.204
		CP	0.957,0.947	0.954,0.937	0.945,0.967	0.956,0.967
				$\bar{m} = 2.82$		
	30%	Mean	0.498,0.504	−0.009, −0.482	0.495,0.485	0.000, −0.450
		SD	0.102,0.191	0.176,0.314	0.096,0.171	0.125,0.181
		SE	0.108,0.190	0.173,0.296	0.104,0.209	0.144,0.226
		CP	0.958,0.954	0.933,0.915	0.952,0.970	0.965,0.967
Scenario 2						
				$\bar{m} = 5.32$		
-	15%	Mean	0.502,0.504	−0.008, −0.480	0.498,0.496	0.003, −0.437
		SD	0.099,0.185	0.178,0.305	0.096,0.180	0.137,0.183
		SE	0.101,0.177	0.176,0.299	0.098,0.187	0.143,0.217
		CP	0.948,0.938	0.933,0.925	0.948,0.940	0.940,0.957
				$\bar{m} = 2.91$		
	30%	Mean	0.497,0.503	−0.009, −0.497	0.493,0.506	−0.006, −0.445
		SD	0.106,0.197	0.188,0.313	0.093,0.184	0.134,0.186
		SE	0.108,0.189	0.187,0.320	0.105,0.205	0.157,0.232
		CP	0.950,0.940	0.934,0.931	0.964,0.961	0.981,0.964

True, true coefficients; Mean, empirical average of point estimates; SD, empirical standard deviation of point estimates; SE, empirical average of standard error estimates; CP, coverage probability based on the 95% confidence interval; $\bar{m}$, average number of observed pairs of recurrence times per subject; cr, average proportion of subjects having the first pair censored; ρ, correlation coefficient.

the point estimates based on Chang's method are biased in the estimation of $\boldsymbol{\beta}_2$ for the covariate A_2, especially when the sample size is small. The inconsistency between the SD's and the SE's for this variable may be due to the bias in its point estimation.

4.2 | Simulation scenario 2

In this scenario, we consider a situation in which the subject-specific latent variables follow different distributions. Specifically, γ_{i1} and γ_{i2} are independently generated from a normal distribution with mean 1 and variance 0.5 and a Gamma distribution $(1/\theta, \theta)$ with the scale parameter $\theta = 0.5$. The results are presented in the lower panels of Tables 1 and 2. Again, the proposed method is virtually unbiased and the SE's are close to their corresponding SD's. As expected, the SD's (and the SE's) increase as the censoring rate increases. We note that whether the latent variables are generated from the normal distribution or the Gamma distribution does not affect the proposed estimation by comparing the results of scenario 2 with the results when $\rho = 0$ under scenario 1. Since we impose no parametric assumption for the subject-specific latent vector in our model assumption, the proposed estimator is robust to the distributions of the latent variables.

Similar to the results in scenario 1, about the same amount of datasets failed to converge based on Chang's method and the summary of the converged datasets shows biased estimates for one covariate. Based on our simulation results from both scenarios, the bias in the point estimation of Chang's method decreases and the number of converged datasets increases as the sample size increases, so we expect Chang's method to be more reliable when the sample size is large.

5 | ANALYSIS OF PSYCHIATRIC CASE REGISTER DATA

In this section, we present the analysis of a subset of the South Verona psychiatric case register data[26] to illustrate the proposed method. We studied a total of 336 patients who were examined with schizophrenia or related disorders and contacted the register for the first time between 1981 and 1995 in South Verona, Italy. Among the patients, 47.9% were male, 59.8% received secondary or higher education, and the age of the patients at onset ranged from 13.7 to 84.0 (median: 37.2). Ten patients who had missing values in education level were excluded from analysis. During the follow-up, patients were in either a care period or a break period, and the 2 states alternated repeatedly over time. According to the definition in Sturt et al[27] and Tansella et al,[28] a break period is when no mental health service is used for over 90 days between consecutive mental health services and a care period begins from the time a psychiatric contact is made until a break occurs. A total number of 1035 bivariate pairs were observed from the 336 patients with the follow-up time ranging from 6 to 5817 days (median: 2406 days). On average, each patient experienced about 3.1 care-and-break episodes (range: $1-18$).

We are interested in evaluating the effects of demographic and socioeconomic factors on the length of care and break periods. Specifically, it is of interest to identify patient characteristics that are associated with longer care period and/or shorter break period because patients with such characteristics may require more medical attention and care. The results of simple regression analyses using the proposed method (Table 3, left panel) show that patients who received secondary or higher education tended to have longer care periods than less educated patients, and patients with an older disease onset age tended to have longer break periods. Multiple regression analyses with all 3 covariates (Table 3, right panel) yield similar results: When holding the other covariates fixed, the length of break periods increased by 28% ($= \exp(0.25) - 1$) when the age of onset was delayed by a decade. Also, patients with a secondary or higher education tended to have 1.78 ($= \exp(0.58)$) times longer duration of care than patients with lower level of education. A previous study conducted on costs of community-based psychiatric care[29] has shown that for patient with schizophrenia, higher education was positively associated with costs of care, which is in line with our finding because extended duration of care would inevitably trigger more costs. When the misspecified rank-based method for univariate recurrent gap time data[6] was implemented, patient's

TABLE 3 Summary of the simple and multiple regression analyses of the South Verona psychiatric case register data with the regression coefficient estimate (Est), standard error estimate (SE), and the 95% confidence interval (95% CI) for each variable

		Simple regression			Multiple regression		
Variables	**Period**	**Est**	**SE**	**95% CI**	**Est**	**SE**	**95% CI**
Gender	Care	0.048	0.194	(−0.331, 0.428)	−0.126	0.204	(−0.526, 0.274)
(male = 1, female = 0)	Break	0.013	0.299	(−0.574, 0.600)	0.264	0.260	(−0.245, 0.772)
Age at onset	Care	−0.085	0.060	(−0.203, 0.033)	−0.001	0.071	(−0.140, 0.138)
(in 10 years)	Break	0.205	0.073*	(0.062, 0.349)	0.252	0.079*	(0.098, 0.406)
Education	Care	0.552	0.199*	(0.162, 0.941)	0.577	0.230*	(0.126, 1.029)
(higher = 1, lower = 0)	Break	−0.207	0.286	(−0.769, 0.354)	0.161	0.278	(−0.383, 0.705)

*P value $< .05$.

education was not significant in either the simple or multiple regression analyses (results not shown). We also tried Chang's method for alternating recurrent gap time data, but it failed to converge.

6 | CONCLUDING REMARKS

In this article, we proposed a semiparametric regression model to make inference about the covariate effects on alternating recurrent event data. The proposed model allows the covariates to have different effects on the 2 alternating states, hence can provide a better understanding of the underlying recurrent event process than methods that do not distinguish the 2 different states within a recurrence episode. In the example of hospitalization data, we could identify which risk factors extend or shorten the actual care periods and what factors prolong or speedup the time from one hospital discharge to the next admission. This can provide useful information for studying patients' quality of life and medical costs, especially when direct measures of these data are not available or difficult to obtain. In either case, hospitalization time data can usually be retrieved relatively easily and economically.

In this article, the dependence structures between the 2 alternating states and among different bivariate gap time pairs within each subject is treated as nuisance. However, when the dependence structure is of interest, estimation methods using copula models may be considered.

In our simulation studies, we compare the performance of the proposed estimator with the rank-based estimator under the same model assumptions considered by Chang.[6] The results show that the proposed estimator is more favorable than the rank-based estimator since the convergence of the rank-based estimator is not guaranteed. In addition to the nonconvergence problem in the point estimation, the variance estimation also suffers from such problem. As discussed in Zeng and Lin,[25] the resampling-based variance estimates for rank-based estimators, such as those in Chang,[6] could be influenced by extreme solutions and become unstable. Unfortunately, this problem would not be resolved by increasing the size of resampling. Tools such as induced smoothing[18,30] and efficient resampling methods[25] may be considered to improve the rank-based method in future research.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the use of the anonymous South Verona, Italy, psychiatric case register data for illustrating the proposed method, provided by Dr Michele Tansella. The authors also thank 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 NCI R01CA193888 to Huang, NSF SES-1659328, DMS-1712717, and NSA H98230-17-1-0308 to Xu, and NCI R03CA187991 and NIMH R03MH112895 to Luo.

ORCID

Chi Hyun Lee http://orcid.org/0000-0001-6340-2718
Chiung-Yu Huang http://orcid.org/0000-0003-2313-3562
Xianghua Luo http://orcid.org/0000-0001-7501-6582

REFERENCES

1. Wang MC, Chang SH. Nonparametric estimation of a recurrent survival function. *J Am Stat Assoc*. 1999;94:146-153.
2. Huang C-Y, Wang MC. Nonparametric estimation of the bivariate recurrence time distribution. *Biometrics*. 2005;61:392-402.
3. Peña EA, Strawderman RL, Hollander M. Nonparametric estimation with recurrent event data. *J Am Stat Assoc*. 2001;96:1299-1315.
4. Du P. Nonparametric modeling of the gap time in recurrent event data. *Lifetime Data Anal*. 2009;15:256-277.
5. Huang Y, Chen YQ. Marginal regression of gap between recurrent events. *Lifetime Data Anal*. 2003;9:293-303.
6. Chang SH. Estimating marginal effects in accelerated failure time models for serial sojourn times among repeated events. *Lifetime Data Anal*. 2004;10:175-190.
7. Schaubel DE, Cai J. Regression methods for gap time hazard functions of sequentially ordered multivariate failure time data. *Biometrika*. 2004;91:291-303.
8. Strawderman RL. The accelerated gap times model. *Biometrika*. 2005;92:647-666.
9. Lu W. Marginal regression of multivariate event times based on linear transformation models. *Lifetime Data Anal*. 2005;11:389-404.

10. Sun LQ, Park DH, Sun JG. The additive hazards model for recurrent gap times. *Stat Sinica*. 2006;16:919-932.
11. Luo X, Huang C-Y. Analysis of recurrent gap time data using the weighted risk set method and the modified within-cluster resampling method. *Stat Med*. 2011;30:301-311.
12. Luo X, Huang C-Y, Wang L. Quantile regression for recurrent gap time data. *Biometrics*. 2013;69:375-385.
13. Darlington GA, Dixon SN. Event-weighted proportional hazards modelling for recurrent gap time data. *Stat Med*. 2013;32:124-130.
14. Kang F, Sun L, Zhao X. A class of transformed hazards models for recurrent gap times. *Comput Stat Data Anal*. 2015;83:151-167.
15. Yan J, Fine JP. Analysis of episodic data with application to recurrent pulmonary exacerbations in cystic fibrosis patients. *J Am Stat Assoc*. 2008;103:498-510.
16. Xue X, Brookmeyer R. Bivariate frailty model for the analysis of multivariate survival time. *Lifetime Data Anal*. 1996;2:277-289.
17. Jin Z, Lin DY, Wei LJ, Ying Z. Rank-based inference for the accelerated failure time model. *Biometrika*. 2003;90:341-353.
18. Chiou SH, Kang S, Yan J. Rank-based estimating equations with general weight for accelerated failure time models: an induced smoothing approach. *Stat Med*. 2015;34:1495-1510.
19. Huang Y. Censored regression with the multistate accelerated sojourn times model. *J R Stat Soc Ser B (Methodological)*. 2002;64:17-29.
20. Cox DR. *Renewal Theory*. London: Methuen and Company, Ltd.; 1962.
21. Lin JS, Wei LJ. Linear regression analysis for multivariate failure time observations. *J Am Stat Assoc*. 1992;87:1091-1097.
22. Kalbfleisch J, Prentice R. *The Statistical Analysis of Failure Time Data*. 2nd ed., Wiley series in probability and statistics. Hoboken, N.J.: J. Wiley; 2002.
23. Pollard D. *Convergence of Stochastic Processes*. New York: Springer; 1984.
24. Huang Y. Two-sample multistate accelerated sojourn times model. *J Am Stat Assoc*. 2000;95:619-627.
25. Zeng D, Lin DY. Efficient resampling methods for nonsmooth estimating functions. *Biometrics*. 2008;9:355-363.
26. Tansella M. *Community-based psychiatry. Long-term patterns of care in South Verona*, Psychological Medicine (Monograph Supplement 19). Cambridge, U.K.: Cambridge University Press; 1991.
27. Sturt E, Wykes T, Creer C. *Demographic, Social and Clinical Characteristics of the Sample. in Long-Term Community Care: Experience in a London Borough*. J. k. wing, Psychological Medicine (Monograph Supplement 2). Cambridge, U.K.: Cambridge University Press; 1982.
28. Tansella M, Micciolo R, Biggeri A, Bisoffi G, Balestrieri M. Episode of care for first-ever psychiatric patients. A long-term case-register evaluation in a mainly urban area. *Br J Psychiatry*. 1995;167:220-227.
29. Amaddeo F, Beecham J, Bonizzato P, Fenyo A, Tansella M, Knapp M. The costs of community-based psychiatric care for first-ever patients: a case register study. *Psychological Med*. 1998;28:173-183.
30. Brown BM, Wang YG. Induced smoothing for rank regression with censored survival times. *Stat Med*. 2007;26:828-836.

How to cite this article: Lee CH, Huang C-Y, Xu G, Luo X. Semiparametric regression analysis for alternating recurrent event data. *Statistics in Medicine*. 2018;37:996–1008. https://doi.org/10.1002/sim.7563