Lifetime Data Anal (2017) 23:83–101
DOI 10.1007/s10985-016-9370-3

Regression analysis for bivariate gap time with missing first gap time data

Chia-Hui Huang[1] · **Yi-Hau Chen**[2,3]

Received: 29 June 2015 / Accepted: 06 June 2016 / Published online: 20 June 2016

Abstract We consider ordered bivariate gap time while data on the first gap time are unobservable. This study is motivated by the HIV infection and AIDS study, where the initial HIV contracting time is unavailable, but the diagnosis times for HIV and AIDS are available. We are interested in studying the risk factors for the gap time between initial HIV contraction and HIV diagnosis, and gap time between HIV and AIDS diagnoses. Besides, the association between the two gap times is also of interest. Accordingly, in the data analysis we are faced with two-fold complexity, namely data on the first gap time is completely missing, and the second gap time is subject to induced informative censoring due to dependence between the two gap times. We propose a modeling framework for regression analysis of bivariate gap time under the complexity of the data. The estimating equations for the covariate effects on, as well as the association between, the two gap times are derived through maximum likelihood and suitable counting processes. Large sample properties of the resulting estimators are developed by martingale theory. Simulations are performed to examine the performance of the proposed analysis procedure. An application of data from the HIV and AIDS study mentioned above is reported for illustration.

Keywords Bivariate duration time · Counting process · Dependent censoring · Ordered data

✉ Chia-Hui Huang
chuang2342@mail.ntpu.edu.tw

[1] National Taipei University, Taipei, Taiwan

[2] Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

[3] Institute of Public Health, National Yang-Ming University, Taipei, Taiwan

1 Introduction

In many medical studies, the process of disease evolution can be classified into two systems, "competing-risks" and "serial" systems. In the competing-risks system, only the first occurring event time and the corresponding failure type are observed for each subject, namely each subject may experience one and only one of several dependent failures. On the other hand, the serial system consists of occurrences of a recurrent event, or a certain event that assumes its progression to go through several successive stages, such as "susceptible", "infected", and "recovered" stages. Owing to the ordering structure in data from the serial system, it is much more informative to utilize information regarding the previous event history for studying the risk of the subsequent event occurrence and the underlying disease progression.

The analysis of serial or ordered event data can either be based on time-to-event models where event times are measured from a common time origin (Wang et al. 2001; Wang and Chiang 2002; Zeng and Lin 2006), or on gap time models where durations between successive events or stages are the main interests (Chang and Wang 1999; Huang 2000; Schaubel and Cai 2004; Sun et al. 2006; Cook and Lawless 2007; Huang and Liu 2007). In the latter type of analysis, various conditional proportional hazards models have been developed, which in particular allow for assessing the effects of subject-specific covariates and event history, such as the gap time of the previous event, on the gap time of the subsequent event (Chang and Wang 1999; Huang 2000).

In event time analysis, it is typical to assume that the censoring due to restriction of the follow-up period or absence of subjects is conditionally independent of the failure times given the covariates. However, in the serial system, the residual censoring time depends on the duration times of the prior events. Unless there is no correlation among the multiple event times, the conditional independence assumption between duration times and the residual censoring time would fail. Such a phenomenon, known as "induced informative censorship", poses a significant challenge in ordered events analysis (Visser 1996; Wang and Wells 1998; Huang and Louis 1998; Lin et al. 1999).

In this study, we focus on regression analysis of bivariate duration time where the first duration time is unobserved. The motivation comes from the analysis of ordered duration times in HIV-infected subjects, where the first duration time is defined as the time from the initial contraction of HIV to diagnosis of HIV, and the second duration time is the time interval between HIV and AIDS diagnoses. In practice, the initial HIV contracting time is usually not available and hence the first duration time is missing, although the diagnosis dates of HIV and AIDS are available if the two events are not censored. One of the study focuses is on how the duration time from the initial contraction of HIV to HIV diagnosis would affect the duration between HIV and AIDS diagnoses. Note that in the setting we considered, although the first gap time (i.e., time from the initial contraction of HIV to diagnosis of HIV) itself is unobservable, whether the first event (i.e. HIV diagnosis) had occurred by time t is observable. This is similar to the setting of the current status or type-I interval censored data, where both regression analysis assessing covariate effects on event time as well as association analysis between two event times is feasible and has been studied in literature (see for example Huang 1996 and Wang and Ding 2000).

To address this problem under the two-fold complexity, namely the missingness of data on the first duration time as well as the informative censoring for the second duration time, we propose a novel modeling framework for bivariate gap time analysis. In this framework, a parametric marginal hazard model is assumed for the first duration time, and a conditional hazard model is specified for the second duration time given the first duration time. Additionally, to facilitate analysis under the complexity mentioned above, unlike conventional survival analysis where the censoring time distribution is unspecified, in our proposal a parametric model is assumed for the censoring time distribution. Our simulations suggest that the analysis results from our proposal are in fact not sensitive to moderate mis-specification of the censoring time distribution. We establish maximum likelihood estimation for the proposed modeling framework. The limiting distribution of the resulting estimators can be derived by martingale theory. Simulation results reveal the nice finite sample performances of the proposed method.

The rest of the paper is organized as fellows. In Sect. 2, we introduce the data structure and a general bivariate model for the gap time distribution. The likelihood function based on the counting processes, and the associated score equations are derived in Sect. 3. Simulation study evaluating performances of the estimates under practical sample sizes is reported in Sect. 4, together with an application to the HIV data. Sect. 5 provides some discussions and conclusions.

2 Data and model

Consider the setting where an individual may experience two successive events. Let D_0 be the calendar time of the initiation, and D_1, D_2 the calendar times of events 1 and 2, respectively. We are primarily interested in the durations or gap times $T_1^* := D_1 - D_0$, and $T_2^* := D_2 - D_1$, and how T_k^* may be affected by the covariates Z_k, $k = 1, 2$. Besides, let C be the calendar time of random censoring and $T_C^* = C - D_0$ the censoring time since initiation of the study. As mentioned in the Introduction section, parametric models, i.e., exponential distributions, are imposed for both gap times T_1^* and T_2^* as well as the censoring time T_C^*, to tackle the two-fold complexity involved in the available data: the unobservable D_0 and hence T_1^*, and the induced informative censoring of T_2^*. Such type of data can be found in the infectious disease study where the first gap time T_1^* is usually unknown because its initial contracting time D_0 is unavailable, though the symptom onset time D_1, the subsequent event time D_2, or the censoring time C may usually be available. The maximum likelihood method is then performed for inference of the proposed models.

We first specify the joint distributions of T_1^* and T_2^* through the marginal and conditional hazard functions as in the following:

$$\begin{aligned}\lambda_1(u|Z_1) &= \lim_{\Delta\to 0} P(T_1^* \in [u, u+\Delta)|T_1^* \geq u, Z_1)/\Delta = \exp(\beta_1' Z_1)\lambda_1,\\ \lambda_2(v|Z_2, T_1^*) &= \lim_{\Delta\to 0} P(T_2^* \in [v, v+\Delta)|T_2^* \geq v, T_1^* = u, Z_2)/\Delta\\ &= \left\{\exp(\beta_2' Z_2) + \theta u\right\}\lambda_2, \qquad (1)\end{aligned}$$

where β_1 and β_2 are unknown regression coefficients, λ_1 and λ_2 are unknown positive constant baseline hazards, respectively, and θ is the association parameter measuring the effect of T_1^* on T_2^* and assumed to be an unknown constant.

We further assume that the censoring time T_C^* since initiation of the study is independent of T_1^* and T_2^* conditioning on the covariates and follows a proportional hazard model:

$$\lambda_C(t|Z_C) = \lim_{\Delta\to 0} P(T_C^* \in [t, t+\Delta)|T_C^* \geq t, Z_C)/\Delta = \exp(\beta_C' Z_C)\lambda_C, \tag{2}$$

where λ_C is an unknown positive constant, Z_C is a set of covariates and β_C the corresponding regression coefficients.

We can see from (1) that the distribution of T_1^* is exponential given the covariates, and the hazard function of T_2^* given T_1^* consists of a multiplicative term for the the covariate effects as well as an additive term for the effect from T_1^*. Like the standard additive hazards model (Lin and Ying 1997, pp. 185–198), the overall conditional hazard function $\{\exp(\beta_2' Z_2) + \theta u\}\lambda_2$ is constrained to be positive. This assumption is always met in our numerical studies. For ease of exposition, from now on we consider the simplified case where the three sets of covariates Z_1, Z_2, and Z_C are identical and denoted by Z.

From (1) we can find the marginal survival function of T_2^*:

$$P(T_2^* > v) = \frac{\exp(-\eta_2\lambda_2 v)}{\eta_1\lambda_1 + \theta\lambda_2 v}\,\eta_1\lambda_1,$$

where $\eta_k = \exp(\beta_k' Z)$, $k = 1, 2$. In addition, the joint survival function $S(u, v) = P(T_1^* > u, T_2^* > v)$ and joint density function $f(u, v) = \partial^2 S(u, v)/\partial u \partial v$ are

$$\begin{aligned} S(u, v) &= \frac{\exp(-\eta_2\lambda_2 v)}{\eta_1\lambda_1 + \theta\lambda_2 v}\eta_1\lambda_1 \;\exp(-[\eta_1\lambda_1 + \theta\lambda_2 v]u), \\ f(u, v) &= \eta_1\lambda_1\,(\theta u + \eta_2)\,\lambda_2 \exp\left(-[\eta_1\lambda_1 + \theta\lambda_2 v]u - \eta_2\lambda_2 v\right). \end{aligned} \tag{3}$$

Let $\delta_1 = I(T_1^* \leq T_C^*)$ indicate whether event 1 (D_1) is firstly observed. If D_1 is observed, then either D_2 or C is subsequently observed; otherwise only C is observed.

To denote the time and the occurrence of the second event, let $T_2 = \min(T_2^*, \left(T_C^* - T_1^*\right)\delta_1)$, $\delta_2^* = I(T_2 = T_2^*)$, and $\delta_C^* = I(T_2 = T_C^* - T_1^*)$. Define the two counting processes $N_2^*(v) = \delta_2^* I(T_2 \leq v)$ and $N_C^*(v) = \delta_C^* I(T_2 \leq v)$ for $0 < v \leq \zeta$, where ζ is the maximum follow-up time for event 2 since the occurrence of event 1. The counting processes $N_2^*(v)$ and $N_C^*(v)$ record the type and the gap time of the second event for $0 < v \leq \zeta$. The only observable information for the first-occurring event is δ_1, i.e. the type (infectious or censored) of the first-occurring event, while the duration time for this event is unobserved. The gap time of the second event is available only when $\delta_1 = 1$, namely when the first-occurring event is not a censoring event. Figure 1 depicts three observed event occurrence patterns corresponding to the cases where only the censoring event is observed ($\delta_1 = 0$, $\delta_2^* = 0$, and $\delta_C^* = 0$), event 1 is observed while event 2 is censored ($\delta_1 = 1$, $\delta_2^* = 0$, and $\delta_C^* = 1$), and both event 1 and event 2 are observed ($\delta_1 = 1$, $\delta_2^* = 1$, and $\delta_C^* = 0$), respectively.

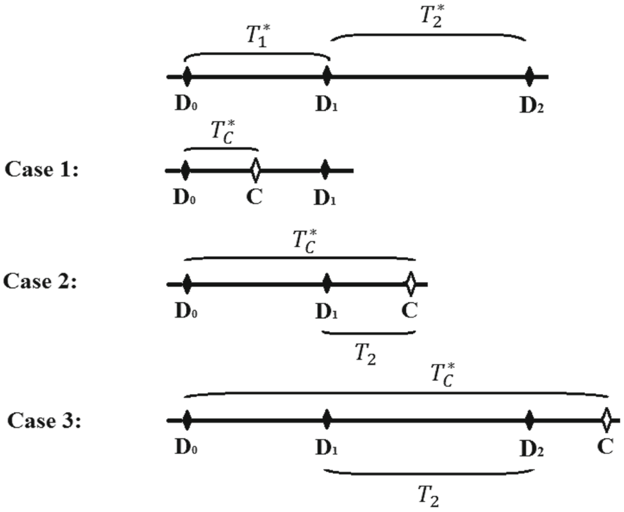

Fig. 1 Three possible patterns of censoring for observed event occurrence; *case 1* $(\delta_1, \delta_2^*, \delta_C^*) = (0, 0, 0)$; *case 2* $(\delta_1, \delta_2^*, \delta_C^*) = (1, 0, 1)$; *case 3* $(\delta_1, \delta_2^*, \delta_C^*) = (1, 1, 0)$

Under the above model and data setups, let $\left(\delta_{1i}, \delta_{2i}^*, \delta_{Ci}^*, T_{2i}, Z_i\right), i = 1, \ldots, n$, be n independent and identically distributed replicates of $\left(\delta_1, \delta_2^*, \delta_C^*, T_2, Z\right)$. For individual i, the data available in the follow-up period are $\{(\delta_{1i}, \delta_{2i}, \delta_{Ci}, T_{2i} \wedge \zeta, Z_i), i = 1, \ldots, n\}$, where $\delta_{2i} = \delta_{2i}^* I(T_{2i} \leq \zeta)$ and $\delta_{Ci} = \delta_{Ci}^* I(T_{2i} \leq \zeta)$, and $a \wedge b \equiv \min(a, b)$. In the next section we develop the maximum likelihood estimation for the model parameters with the observable data. Although the models considered are fully parametric, to develop the maximum likelihood estimation for the model parameters in a neat and systematic way, we will utilize the representation associated with the observable counting processes $N_{2i}(v) = \delta_{2i} I(T_{2i} \leq v)$ and $N_{Ci}(v) = \delta_{Ci} I(T_{2i} \leq v)$, as shown in the next section.

3 Maximum likelihood method

3.1 Intensity and likelihood functions

To establish the likelihood function of the observed bivariate gap time data as described in the previous section, we consider the first-occurring event and then the subsequent gap time conditioning on the previous event type. For the first-occurring event, we can observe which type of event it is and the probability of this event being event 1, i.e. $\delta_{1i} = 1$, is $\eta_{1i}\lambda_1/(\eta_{1i}\lambda_1 + \eta_{Ci}\lambda_C)$ with $\eta_{Ci} = \exp(\beta_C' Z_i)$. On the other hand, the probability that the first-occurring event is a censoring event, i.e. $\delta_{1i} = 0$, is $\eta_{Ci}\lambda_C/(\eta_{1i}\lambda_1 + \eta_{Ci}\lambda_C)$.

Conditioning on $\delta_{1i} = 1$, we are able to observe two counting processes $N_{2i}(\cdot)$ and $N_{Ci}(\cdot)$. We derive the cause-specific intensity function of these two counting processes, so that the unbiased estimating equations can be constructed. According to models (1)–(2), the survival function of T_{2i} and the sub-density functions can be shown, for $0 < v \leq \zeta$, to be

$$\begin{aligned} G(v) = P(T_{2i} > v|\delta_{1i} = 1) &= \frac{\exp\left(-[\eta_{2i}\lambda_2 + \eta_{Ci}\lambda_C]\, v\right)}{\eta_{1i}\lambda_1 + \theta\lambda_2 v + \eta_{Ci}\lambda_C}\ (\eta_{1i}\lambda_1 + \eta_{Ci}\lambda_C)\,, \\ f(T_{2i} = v, \delta_{2i} = 1|\delta_{1i} = 1) &= \frac{\exp\left(-[\eta_{2i}\lambda_2 + \eta_{Ci}\lambda_C]\, v\right)(\eta_{1i}\lambda_1 + \eta_{Ci}\lambda_C)}{\eta_{1i}\lambda_1 + \theta\lambda_2 v + \eta_{Ci}\lambda_C} \\ &\quad \times \left\{\eta_{2i} + \frac{\theta}{\eta_{1i}\lambda_1 + \theta\lambda_2 v + \eta_{Ci}\lambda_C}\right\}\lambda_2, \\ f(T_{2i} = v, \delta_{Ci} = 1|\delta_{1i} = 1) &= \frac{\exp\left(-[\eta_{2i}\lambda_2 + \eta_{Ci}\lambda_C]\, v\right)(\eta_{1i}\lambda_1 + \eta_{Ci}\lambda_C)}{\eta_{1i}\lambda_1 + \theta\lambda_2 v + \eta_{Ci}\lambda_C}\eta_{Ci}\lambda_C. \end{aligned}$$

Therefore, for $0 < v \leq \zeta$, the cause-specific intensity functions of $N_{2i}(v)$ and $N_{Ci}(v)$ are given by $Y_i(v)\tilde{\eta}_{2i}(v)\lambda_2$ and $Y_i(v)\tilde{\eta}_{Ci}(v)\lambda_C$, respectively, where

$$\tilde{\eta}_{2i}(v) = \eta_{2i} + \frac{\theta}{\eta_{1i}\lambda_1 + \theta\lambda_2 v + \eta_{Ci}\lambda_C}, \tag{4}$$

$$\tilde{\eta}_{Ci}(v) = \eta_{Ci}, \tag{5}$$

with $Y_i(v) = I(T_{2i} \geq v)$ the at-risk process for T_{2i}.

Comparing models (2) and (5), it is seen that the cause-specific intensity function for $N_{Ci}(v)$ is the same as its original net intensity. This is expectable because we have assumed independence between (T_1^*, T_2^*) and T_C^* conditioning on the covariates. On the other hand, the cause-specific intensity of $N_{2i}(v)$ is a function depending on the parameters in the models for both T_{1i}^* and T_{Ci}^*. Define

$$\begin{aligned} M_{2i}(t) &= N_{2i}(t) - \int_0^t Y_i(v)\tilde{\eta}_{2i}(v)\lambda_2 dv, \\ M_{Ci}(t) &= N_{Ci}(t) - \int_0^t Y_i(v)\tilde{\eta}_{Ci}(v)\lambda_C dv, \end{aligned} \tag{6}$$

which are martingales with respect to the filtration $\mathscr{F}_t=\sigma\{N_{2i}(v), N_{Ci}(v), Y_i(v), Z_i : v < t, i = 1, \ldots, n\}$.

Set $\lambda_k = \exp(\alpha_k), k = 1, 2$, and $\lambda_C = \exp(\alpha_C)$. Let $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_C)$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_C)$ and $\boldsymbol{\Omega} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \theta)$. We shall assume that two counting processes cannot jump simultaneously. Accordingly, the likelihood function $\mathscr{L}(\boldsymbol{\Omega})$ based on $\{\delta_{1i}, \delta_{2i}, \delta_{Ci}, T_{2i}, Z_i\}$ can be written as

$$\begin{aligned} \mathscr{L}(\boldsymbol{\Omega}) = \prod_{i=1}^{n} & P(\delta_{1i} = 1)^{\delta_{1i}} \{1 - P(\delta_{1i} = 1)\}^{1-\delta_{1i}} \\ & \times \{\tilde{\eta}_{2i}(T_{2i} \wedge \zeta)\exp(\alpha_2)\}^{\delta_{2i}} \{\tilde{\eta}_{Ci}(T_{2i} \wedge \zeta)\exp(\alpha_C)\}^{\delta_{Ci}}\, G(T_{2i} \wedge \zeta)^{\delta_{1i}}. \end{aligned}$$

The first two terms in the likelihood correspond to the first-occurring event, and the other terms correspond to the second-occurring event conditioning on the first event.

Expressed explicitly in this way, the loglikelihood function $\ell(\boldsymbol{\Omega}) = \ell_1(\boldsymbol{\Omega}) + \ell_2(\boldsymbol{\Omega})$, where

$$\begin{aligned}
\ell_1(\boldsymbol{\Omega}) &= \sum_{i=1}^{n} \delta_{1i}\left(\alpha_1 + \beta_1' Z_i\right) + (1-\delta_{1i})\left(\alpha_C + \beta_C' Z_i\right) \\
&\quad - \log\left\{\exp\left(\alpha_1 + \beta_1' Z_i\right) + \exp(\alpha_C + \beta_C' Z_i)\right\}, \\
\ell_2(\boldsymbol{\Omega}) &= \sum_{i=1}^{n} \int_0^{\zeta} \{\alpha_2 + \log \tilde{\eta}_{2i}(v)\}\, dN_{2i}(v) - Y_i(v)\tilde{\eta}_{2i}(v)\exp(\alpha_2) dv \\
&\quad + \sum_{i=1}^{n} \int_0^{\zeta} \left\{\alpha_C + \beta_C' Z_i\right\} dN_{Ci}(v) - Y_i(v)\exp\left(\beta_C' Z_i\right)\exp(\alpha_C) dv. \quad (7)
\end{aligned}$$

3.2 Estimation and asymptotic properties

To find the estimators, we apply the maximum likelihood method. The score functions, given below, are obtained by taking the derivative of $\ell(\boldsymbol{\Omega})$ with respect to $\boldsymbol{\Omega} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \theta)$:

$$\begin{aligned}
U_{\alpha_1} &= \sum_{i=1}^{n} \left\{\delta_{1i} - \frac{\exp(\alpha_1 + \beta_1' Z_i)}{\exp(\alpha_1 + \beta_1' Z_i) + \exp(\alpha_C + \beta_C' Z_i)}\right\} + \sum_{i=1}^{n} \int_0^{\zeta} \widetilde{X}_{1i}(v) dM_{2i}(v), \\
U_{\alpha_2} &= \sum_{i=1}^{n} \int_0^{\zeta} \left\{\widetilde{X}_{2i}(v) + 1\right\} dM_{2i}(v), \\
U_{\alpha_C} &= \sum_{i=1}^{n} \left\{(1-\delta_{1i}) - \frac{\exp(\alpha_C + \beta_C' Z_i)}{\exp(\alpha_1 + \beta_1' Z_i) + \exp(\alpha_C + \beta_C' Z_i)}\right\} \\
&\quad + \sum_{i=1}^{n} \int_0^{\zeta} \left\{\widetilde{X}_{Ci}(v) dM_{2i}(v) + dM_{Ci}(v)\right\}, \\
U_{\beta_1} &= \sum_{i=1}^{n} Z_i \left\{\delta_{1i} - \frac{\exp(\alpha_1 + \beta_1' Z_i)}{\exp(\alpha_1 + \beta_1' Z_i) + \exp(\alpha_C + \beta_C' Z_i)}\right\} \\
&\quad + \sum_{i=1}^{n} \int_0^{\zeta} \widetilde{Z}_{1i}(v) dM_{2i}(v), \\
U_{\beta_2} &= \sum_{i=1}^{n} \int_0^{\zeta} \widetilde{Z}_{2i}(v) dM_{2i}(v), \\
U_{\beta_C} &= \sum_{i=1}^{n} Z_i \left\{(1-\delta_{1i}) - \frac{\exp(\alpha_C + \beta_C' Z_i)}{\exp(\alpha_1 + \beta_1' Z_i) + \exp(\alpha_C + \beta_C' Z_i)}\right\} \\
&\quad + \sum_{i=1}^{n} \int_0^{\zeta} \left\{\widetilde{Z}_{Ci}(v) dM_{2i}(v) + Z_i\ dM_{Ci}(v)\right\},
\end{aligned}$$

$$U_\theta = \sum_{i=1}^{n} \int_0^\zeta \widetilde{W}_i(v) dM_{2i}(v),$$

where $\widetilde{X}_{ki}(v) = (\partial/\partial\alpha_k) \log \tilde{\eta}_{2i}(v)$, $\widetilde{Z}_{ki}(v) = (\partial/\partial\beta_k) \log \tilde{\eta}_{2i}(v)$, $k = 1, 2$, and similarly for $\widetilde{X}_{Ci}(v)$ and $\widetilde{Z}_{Ci}(v)$, and $\widetilde{W}_i(v) = (\partial/\partial\theta) \log \tilde{\eta}_{2i}(v)$. We use the MATLAB function "fminunc" to optimize the objective function given by the negative loglikelihood, with the score equations and the hessian matrix, which is the second derivate of the objective function, explicitly provided. Let $\widehat{\boldsymbol{\Omega}} = \left(\widehat{\boldsymbol{\alpha}}, \ \widehat{\boldsymbol{\beta}}, \hat{\theta}\right)$ be the solution of the system of the score equations. It can be seen that the score functions are martingales. Hence we can apply the martingale central limit theorem to establish the large sample properties of $\widehat{\boldsymbol{\Omega}}$, as shown in Theorem 1 below.

Let

$$\pi_i = \frac{\exp(\alpha_1 + \beta_1' Z_i)}{\exp(\alpha_1 + \beta_1' Z_i) + \exp(\alpha_C + \beta_C' Z_i)},$$
$$\bar{\pi}_i = \frac{\exp(\alpha_C + \beta_C' Z_i)}{\exp(\alpha_1 + \beta_1' Z_i) + \exp(\alpha_C + \beta_C' Z_i)},$$

and

$$\begin{aligned} \boldsymbol{S}_{1i}(\boldsymbol{\Omega}) &= \left(\delta_{1i} - \pi_i, \ 0, 1 - \delta_{1i} - \bar{\pi}_i, \ Z_i'(\delta_{1i} - \pi_i), \ 0, \ Z_i'(1 - \delta_{1i} - \bar{\pi}_i), \ 0\right)', \\ \boldsymbol{S}_{2i}(v, \boldsymbol{\Omega}) &= \left(\widetilde{X}_{1i}(v), \ \widetilde{X}_{2i}(v) + 1, \ \widetilde{X}_{Ci}(v), \ \widetilde{Z}_{1i}'(v), \ \widetilde{Z}_{2i}'(v), \ \widetilde{Z}_{Ci}'(v), \ \widetilde{W}_i(v)\right)', \\ \boldsymbol{S}_{Ci}(\boldsymbol{\Omega}) &= \left(0, \ 0, \ 1 \ , 0, \ 0, \ Z_i', \ 0\right)'. \end{aligned} \tag{8}$$

Hence $\widehat{\boldsymbol{\Omega}}$ is the root of the following estimating functions

$$\sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\Omega}) = \sum_{i=1}^{n} \left\{ \boldsymbol{S}_{1i}(\boldsymbol{\Omega}) + \int_0^\zeta \boldsymbol{S}_{2i}(v, \boldsymbol{\Omega}) \, dM_{2i}(v) + \int_0^\zeta \boldsymbol{S}_{Ci}(\boldsymbol{\Omega}) \, dM_{Ci}(v) \right\},$$

where $\boldsymbol{S}_{1i}$, $\boldsymbol{S}_{2i}$, and $\boldsymbol{S}_{Ci}$ are components of the estimating functions associated with the first-occurring event, the gap time between the first and second events, and the censoring time, respectively.

Let $\boldsymbol{\Omega}_0 = (\boldsymbol{\alpha}_0, \ \boldsymbol{\beta}_0, \ \theta_0)$, where $\boldsymbol{\alpha}_0, \ \boldsymbol{\beta}_0, \ \theta_0$ are the true values of $\boldsymbol{\alpha}, \ \boldsymbol{\beta}, \ \theta$, respectively. The consistency and weak convergence properties are established in the following theorem, with the proof provided in the Appendix. The following assumptions are required.

(A1) Z is bounded and left-continuous. Also the true values of parameters are in the interior of a known compact set.

(A2) $P(\delta_1 = 1|Z) > 0$, $P(Y(\zeta) = 1|Z) > 0$, $P(\delta_2 = 0, T_2 \geq \zeta|Z) > 0$.

(A3) There exists a deterministic $\boldsymbol{s}_1$ such that

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \boldsymbol{S}_{1i}(\boldsymbol{\Omega}_0)}{\partial \boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{s}_1.$$

(A4) There exists a deterministic, integrable functions $\boldsymbol{s}_2$ and $\boldsymbol{s}_C$, such that

$$\frac{1}{n}\sum_{i=1}^{n} Y_i(v)\tilde{\eta}_{2i}(v)\lambda_2 \boldsymbol{S}_{2i}(v, \boldsymbol{\Omega_0})^{\otimes 2} \xrightarrow{p} \boldsymbol{s}_2(v),$$
$$\frac{1}{n}\sum_{i=1}^{n} Y_i(v)\tilde{\eta}_{Ci}(v)\lambda_C \boldsymbol{S}_{Ci}(\boldsymbol{\Omega_0})^{\otimes 2} \xrightarrow{p} \boldsymbol{s}_C(v),$$

where $a^{\otimes 2} = aa'$ for a column vector a.

(A5) There exists a positive-definite matrix $\mathscr{I}_0$ such that

$$\mathscr{I}_0 = \boldsymbol{s}_1 + \int_0^{\zeta} \boldsymbol{s}_2(v)dv + \int_0^{\zeta} \boldsymbol{s}_C(v)dv.$$

Theorem 1 *Assume that conditions* (A1)–(A5) *hold. Then there exists a neighborhood of $\boldsymbol{\Omega}_0$, within which $\widehat{\boldsymbol{\Omega}}$ is a unique solution to $\sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\Omega}) = 0$, and $\widehat{\boldsymbol{\Omega}} \to \boldsymbol{\Omega}_0$ almost surely. Furthermore,*

$$\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0,\ \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0,\ \hat{\theta} - \theta_0)$$

is asymptotically normal with mean zero and variance-covariance matrix $\mathscr{I}_0^{-1}$. The information matrix $\mathscr{I}_0$ can be consistently estimated by $\widehat{\mathscr{I}}/n$, where $\widehat{\mathscr{I}} = -\sum_{i=1}^{n} \partial \boldsymbol{U}_i(\boldsymbol{\Omega})/\partial \boldsymbol{\Omega}$ evaluated at $\widehat{\boldsymbol{\Omega}}$ is the observed information matrix.

4 Numerical examples

We conduct simulations under several scenarios to evaluate the performance of the proposed estimator and the associated large sample theory, including standard error estimation, the coverage probability of the Wald-type and the bootstrap-based confidence intervals. We also apply the proposed method to the HIV data mentioned in the Introduction.

4.1 Simulation

We consider the covariate Z, which is a two-dimensional vector with the first component generated from a standard normal distribution truncated at ±2, and the second component a bernoulli trial with probability 0.5. The first gap time T_1^* for event 1 and the censoring time T_C^* are respectively simulated from (1) and (2) with the given values of $(\alpha_1, \alpha_C, \beta_1, \beta_C)$. If $\delta_1 = 1$, we continue to simulate the second gap time T_2^*

for event 2 from the conditional hazard function (1), with θ ranging from 0.2 to 0.8. The censoring rate for T_2^* ranges from 36 to 46 %, where a quarter of the subjects may be censored before event 1 occurs, i.e., $\delta_1 = 0$.

Table 1 shows the summary statistics based on 5000 simulation replications. In the scenario where $\theta = 0.4$, about 71 % of the subjects experience event 1 and then 56 % of the subjects experience the subsequent event 2. The estimates for the regression coefficients are nearly unbiased and the empirical standard errors are quite close to the estimated standard errors based on the observed information matrix. The coverage probabilities of the Wald-type confidence interval based on Theorem 1 are close to the desired level. Compared to the regression parameter estimates, the association parameter estimate $\hat{\theta}$ has larger bias, its asymptotic standard error underestimates the empirical standard error, and the coverage probability of the confidence interval based on asymptotic theory is slightly lower than the nominal 95 % level under the sample size 1000. We also apply the bootstrap method to estimate the standard error of the parameters and obtain the bootstrap-based confidence interval. We see that the bootstrap standard error for $\hat{\theta}$ matches better with the simulation standard error, and the coverage probability of the bootstrap-based confidence interval for θ is better than that based on asymptotic theory when $n = 1000$. When the sample size increases to 2000, the bias decreases and the coverage probability of the confidence interval based on asymptotic theory becomes closer to the desired value. We further consider the scenario with $\theta = 0.6$; the results in this scenario are consistent with those in the former one.

We also consider a setting where the covariate effects do not exist and the value of θ is 0.2 or 0.8. The results are given in Table 2, showing that the large sample properties work well. From Tables 1 and 2, we see that it requires a larger sample size to ensure the theoretical properties, which may be due to that the first duration time is unobservable. This, however, will not cause a major limitation on the proposed analysis, since data on the first duration time are already allowed to be missing, more subjects are eligible to be included in the analysis.

Since the proposed analysis further assumes the parametric model (2) for the censoring time T_C^*, there is a potential concern with robustness against mis-specification for such a model. To examine this problem, we perform simulations where T_C^* does not follow an exponential distribution as in model (2), but follows a gamma or Weibull distribution with the shape and scale parameters chosen to yield a percentage around 30 % of observing event 2. Data on T_1^* and T_2^* are still generated from the model (1), The sample sizes considered are $n = 1000$, 2000, and 3000. Table 3 gives the analysis results from the proposed models (1) and (2). It is seen that results for the covariate effects β_1 and β_2 remain satisfactory when T_C^* is generated from either the gamma or Weibull distribution, while in analysis it is wrongly assumed to follow the exponential distribution (2). The estimate for θ has a larger bias when $n = 1000$; however, the bias decreases when sample size increases. The estimate for the baseline hazard λ_1 is severely biased and the coverage probability of the confidence interval is much lower than the desired level of 95 %, which seems to be expectable given that data on T_1^* is fully unobservable and hence the associated analysis is highly model-dependent. The sensitivity analysis is also performed in the setting when the covariate effects do not exist and $\theta = 0.5$, and the values of the shape and scale parameters in gamma or

Table 1 Results are based on $n = 1000, 2000$ and $\theta = 0.4, 0.6$

Parameter	True value	$n = 1000$						$n = 2000$					
		Bias	SE	$\widehat{\text{SE}}$	CP(%)	BSE	BCP(%)	Bias	SE	$\widehat{\text{SE}}$	CP(%)	BSE	BCP(%)
Scenario 1: 71 % of event 1, 56 % of event 2													
β_1	−0.5	−0.003	0.081	0.080	94.5	0.080	94.9	−0.002	0.056	0.057	96.0	0.057	95.5
	1	0.007	0.149	0.151	96.2	0.153	95.7	0.001	0.109	0.109	95.0	0.109	94.9
β_2	−0.5	0.000	0.067	0.068	94.6	0.068	94.2	−0.001	0.044	0.045	95.6	0.045	95.0
	1	0.003	0.133	0.133	94.7	0.134	94.1	0.002	0.091	0.087	94.9	0.088	94.5
β_C	−0.5	−0.003	0.062	0.060	94.9	0.060	94.1	0.000	0.041	0.042	95.2	0.041	95.1
	1	0.007	0.114	0.114	95.4	0.115	94.9	−0.001	0.078	0.079	94.8	0.079	95.2
λ_1	0.3	0.002	0.037	0.039	94.6	0.039	94.3	0.003	0.028	0.027	94.0	0.027	93.4
λ_2	0.3	0.004	0.049	0.048	93.9	0.049	93.3	0.001	0.033	0.031	94.6	0.031	92.9
λ_C	0.1	0.000	0.011	0.011	94.7	0.011	94.6	0.000	0.007	0.007	95.3	0.007	94.9
θ	0.4	0.031	0.234	0.217	91.5	0.238	92.0	0.013	0.110	0.102	92.5	0.107	92.7
Scenario 2: 71 % of event 1, 57 % of event 2													
β_1	−0.5	−0.003	0.079	0.079	95.0	0.080	95.0	−0.003	0.055	0.055	95.2	0.055	94.5
	1	0.007	0.149	0.149	95.3	0.151	95.2	0.005	0.103	0.103	95.1	0.104	94.9
β_2	−0.5	0.000	0.071	0.071	94.8	0.072	94.8	0.001	0.053	0.052	95.0	0.052	94.6
	1	0.000	0.139	0.140	95.0	0.142	94.9	−0.007	0.102	0.103	94.9	0.103	94.8
β_C	−0.5	−0.003	0.061	0.061	95.2	0.061	95.0	−0.001	0.043	0.043	95.0	0.043	94.9
	1	0.006	0.115	0.116	95.2	0.116	95.3	0.002	0.080	0.083	96.1	0.083	95.4
λ_1	0.3	0.002	0.039	0.039	94.9	0.040	95.0	0.000	0.027	0.028	95.2	0.028	94.8
λ_2	0.3	0.006	0.051	0.051	95.3	0.053	94.5	0.005	0.037	0.037	95.1	0.038	94.5
λ_C	0.1	0.000	0.011	0.011	95.1	0.011	94.8	0.000	0.008	0.008	94.1	0.008	93.6
θ	0.6	0.022	0.281	0.278	91.7	0.304	92.5	0.008	0.237	0.235	93.3	0.244	92.9

The statistics are based on the 5000 replications, where SE is the simulation standard error, $\widehat{\text{SE}}$ is the average of estimated standard error, CP is the coverage probability of 95 % confidence interval based on asymptotic standard error, BSE is the bootstrap standard error, and BCP is the coverage probability of the bootstrap-based confidence interval

Table 2 Results are based on $n = 2000$, 3000 and $\theta = 0.2, 0.8$

Parameter	True value	$n = 2000$						$n = 3000$					
		Bias	SE	$\widehat{\text{SE}}$	CP(%)	BSE	BCP(%)	Bias	SE	$\widehat{\text{SE}}$	CP(%)	BSE	BCP(%)
Scenario 1: 75 % of event 1, 60 % of event 2													
β_1	0	−0.002	0.053	0.052	94.8	0.053	94.5	−0.001	0.042	0.042	94.7	0.043	94.3
	0	0.000	0.097	0.099	95.5	0.102	95.6	−0.002	0.079	0.081	95.7	0.082	96.1
β_2	0	0.001	0.042	0.042	95.3	0.042	95.1	0.001	0.035	0.034	95.4	0.034	94.8
	0	0.001	0.079	0.080	95.8	0.080	95.3	0.000	0.066	0.065	96.0	0.065	95.3
β_C	0	0.000	0.040	0.041	95.1	0.040	94.4	−0.001	0.034	0.033	94.6	0.033	93.9
	0	0.001	0.076	0.078	95.2	0.078	95.0	−0.001	0.064	0.063	94.9	0.063	94.4
λ_1	0.3	0.001	0.028	0.028	95.5	0.029	95.6	0.002	0.023	0.023	94.9	0.023	95.3
λ_2	0.3	0.010	0.047	0.044	91.5	0.046	92.9	0.008	0.040	0.036	93.3	0.038	93.4
λ_C	0.1	0.000	0.007	0.007	94.5	0.007	94.6	0.000	0.006	0.006	94.8	0.006	95.3
θ	0.2	−0.005	0.104	0.100	93.0	0.106	93.2	−0.006	0.090	0.083	94.9	0.087	93.9
Scenario 2: 75 % of event 1, 65 % of event 2													
β_1	0	0.002	0.047	0.046	95.1	0.047	95.1	0.000	0.037	0.038	95.3	0.038	95.3
	0	−0.001	0.093	0.089	94.0	0.089	93.6	0.000	0.074	0.073	94.9	0.073	94.6
β_2	0	0.002	0.065	0.064	95.3	0.064	95.1	0.000	0.051	0.052	95.8	0.052	95.4
	0	0.004	0.121	0.122	95.9	0.122	95.0	−0.006	0.100	0.099	95.1	0.099	94.5
β_C	0	0.000	0.042	0.043	95.3	0.043	94.4	0.000	0.035	0.035	94.7	0.035	94.2
	0	0.000	0.083	0.083	94.4	0.083	94.2	0.000	0.067	0.068	95.6	0.068	95.1
λ_1	0.3	0.001	0.030	0.030	94.9	0.030	94.8	0.001	0.024	0.024	96.0	0.024	95.4
λ_2	0.3	0.004	0.038	0.037	95.2	0.038	95.3	0.005	0.030	0.030	95.4	0.031	95.3
λ_C	0.1	0.000	0.008	0.008	94.7	0.008	94.2	0.000	0.007	0.007	96.0	0.007	95.2
θ	0.8	0.005	0.195	0.190	93.3	0.193	93.2	−0.006	0.152	0.152	93.7	0.154	93.6

The statistics are based on the 5000 replications, where SE is the simulation standard error, $\widehat{\text{SE}}$ is the average of estimated standard error, CP is the coverage probability of 95 % confidence interval based on asymptotic standard error, BSE is the bootstrap standard error, and BCP is the coverage probability of the bootstrap-based confidence interval

Table 3 Simulation results when the distribution of T_C^* is misspecified

n	Parameter	Scenario 1: Gamma (2, 3)				Scenario 2: Weibull (3, 0.5)			
		Bias	SE	$\widehat{\text{SE}}$	CP(%)	Bias	SE	$\widehat{\text{SE}}$	CP(%)
1000	$\beta_1 : z_1 = -0.5$	−0.034	0.106	0.108	94.48	0.028	0.131	0.122	92.66
	$\beta_1 : z_2 = 1$	0.072	0.194	0.200	94.56	−0.068	0.244	0.232	92.72
	$\beta_2 : z_1 = -0.5$	−0.008	0.130	0.125	91.76	0.008	0.174	0.164	91.88
	$\beta_2 : z_2 = 1$	0.040	0.326	0.312	93.02	−0.020	0.398	0.389	91.38
	$\lambda_1 : 0.1$	0.031	0.018	0.019	66.00	−0.046	0.010	0.010	2.56
	$\lambda_2 : 0.1$	0.008	0.046	0.042	88.64	0.011	0.055	0.050	91.56
	$\theta : 0.5$	0.178	0.712	0.577	89.08	0.136	0.712	0.615	86.06
2000	$\beta_1 : z_1 = -0.5$	−0.034	0.074	0.076	93.60	0.029	0.092	0.086	92.10
	$\beta_1 : z_2 = 1$	0.073	0.136	0.141	92.88	−0.076	0.169	0.163	91.32
	$\beta_2 : z_1 = -0.5$	−0.006	0.090	0.089	95.44	0.005	0.119	0.114	93.84
	$\beta_2 : z_2 = 1$	0.031	0.218	0.216	95.66	−0.035	0.271	0.261	92.30
	$\lambda_1 : 0.1$	0.030	0.013	0.013	34.46	−0.046	0.007	0.007	0.02
	$\lambda_2 : 0.1$	0.004	0.032	0.030	93.54	0.005	0.036	0.034	94.14
	$\theta : 0.5$	0.085	0.366	0.342	92.92	0.043	0.378	0.333	88.96
3000	$\beta_1 : z_1 = -0.5$	−0.031	0.061	0.062	92.82	0.034	0.073	0.070	90.84
	$\beta_1 : z_2 = 1$	0.068	0.112	0.115	91.46	−0.073	0.142	0.133	88.80
	$\beta_2 : z_1 = -0.5$	−0.007	0.074	0.072	95.30	0.008	0.094	0.092	93.74
	$\beta_2 : z_2 = 1$	0.034	0.181	0.175	95.96	−0.039	0.214	0.210	92.74
	$\lambda_1 : 0.1$	0.030	0.010	0.011	15.80	−0.046	0.006	0.006	0
	$\lambda_2 : 0.1$	0.002	0.025	0.024	94.60	0.003	0.028	0.028	94.92
	$\theta : 0.5$	0.067	0.290	0.271	94.08	0.014	0.264	0.251	91.14

In scenario 1, there are 55 % of study subjects experiencing event 1 and 36 % experiencing event 2; in scenario 2, there are 35 % of study subjects experiencing event 1 and 23 % experiencing event 2

Weibull distribution are chosen so that the percentage of observing event 2 is about 30 %. The results shown in Table 4 lead to conclusions quite similar to those based on Table 3.

4.2 Data analysis

A dataset of 3255 subjects is analyzed to illustrate the proposed method. In this analysis, event 1 is defined as the HIV diagnosis, event 2 is the AIDS diagnosis, and censoring is caused by death. The data were collected from hospitals, centers for disease control, and prevention and public health centers from 1984 to 2001 in Taiwan. Although the diagnosis dates of HIV and AIDS were recorded if they had occurred, the initial date of HIV contraction was unavailable for all subjects. Some of the subjects had the record of death time during the study period. About 1.2 % of subjects died before HIV diagnosis. In the end of 2001, 16.4 % of subjects were diagnosed with AIDS and 12.7

Table 4 Simulation results when the distribution of T_C^* is misspecified

n	Parameter	Scenario 1: Gamma (0.8, 12.5)				Scenario 2: Weibull (6, 2/3)			
		Bias	SE	$\widehat{\text{SE}}$	CP(%)	Bias	SE	$\widehat{\text{SE}}$	CP(%)
1000	$\beta_1 : z_1 = 0$	0.001	0.112	0.108	94.14	−0.002	0.121	0.115	93.85
	$\beta_1 : z_2 = 0$	−0.001	0.213	0.208	94.44	−0.002	0.229	0.220	94.17
	$\beta_2 : z_1 = 0$	−0.001	0.215	0.196	95.48	−0.002	0.270	0.224	95.09
	$\beta_2 : z_2 = 0$	0.007	0.407	0.387	97.08	0.003	0.472	0.450	98.50
	$\lambda_1 : 0.1$	−0.009	0.014	0.014	86.10	−0.025	0.012	0.012	43.44
	$\lambda_2 : 0.1$	0.006	0.036	0.035	95.08	0.011	0.046	0.043	95.33
	$\theta : 0.5$	0.057	0.511	0.374	88.76	0.094	0.950	0.605	85.99
2000	$\beta_1 : z_1 = 0$	0.000	0.079	0.076	94.00	−0.001	0.085	0.080	93.84
	$\beta_1 : z_2 = 0$	0.001	0.146	0.146	94.78	0.003	0.158	0.154	94.80
	$\beta_2 : z_1 = 0$	0.000	0.140	0.132	94.68	−0.002	0.159	0.148	95.20
	$\beta_2 : z_2 = 0$	−0.004	0.258	0.254	96.40	−0.002	0.296	0.286	96.78
	$\lambda_1 : 0.1$	−0.009	0.010	0.010	80.94	−0.025	0.009	0.008	19.42
	$\lambda_2 : 0.1$	0.004	0.025	0.024	94.72	0.007	0.030	0.029	95.36
	$\theta : 0.5$	0.002	0.210	0.198	89.74	0.001	0.271	0.244	88.14
3000	$\beta_1 : z_1 = 0$	−0.001	0.062	0.062	94.70	0.001	0.069	0.065	93.86
	$\beta_1 : z_2 = 0$	0.002	0.119	0.118	94.90	0.001	0.129	0.126	94.10
	$\beta_2 : z_1 = 0$	−0.001	0.109	0.107	95.02	0.002	0.126	0.118	94.80
	$\beta_2 : z_2 = 0$	0.000	0.209	0.206	95.78	0.001	0.234	0.227	95.88
	$\lambda_1 : 0.1$	−0.009	0.008	0.008	76.16	−0.025	0.007	0.007	7.48
	$\lambda_2 : 0.1$	0.003	0.020	0.020	94.70	0.006	0.024	0.023	95.54
	$\theta : 0.5$	0.000	0.166	0.157	91.88	−0.015	0.263	0.188	88.98

In scenario 1, there are 49 % of study subjects experiencing event 1 and 33 % experiencing event 2; in scenario 2, there are 65 % of study subjects experiencing event 1 and 36 % experiencing event 2

% died between HIV and AIDS diagnoses. The average age of being diagnosed with HIV was 34.4 and the standard deviation was 12.2.

We apply the proposed analysis to the data by considering potential risk factors for HIV, AIDS, and death, including: homosexuality (Z_1), heterosexuality (Z_2), bisexuality (Z_3). According to the documentation, the proportions of the above patient groups are 32.23, 47.50, and 13.43 %, respectively, and the reference group consists of the subjects who did not respond to this question about sexual orientation and the proportion is 6.85 %. Table 5 shows the estimates of regression coefficients and the corresponding estimated standard errors. We can see that there is no statistically significant effect found in analyzing the gap time between time of contraction and diagnosis time of HIV-positive. However, subjects with bisexuality seem to have a higher risk of becoming HIV-positive, compared to individuals in other categories. After HIV diagnosis, heterosexual and bisexual subjects have a significantly higher risk of being AIDS-positive than subjects with unknown sexual orientation, while the increased risk for subjects with homosexuality is not significant at 5 % significance level.

Table 5 Analysis of HIV and AIDS positive data

Variable	Time to HIV positive		Time to AIDS positive after HIV infected		Time to death	
	$\hat{\beta}_1$	SE($\hat{\beta}_1$)	$\hat{\beta}_2$	SE($\hat{\beta}_2$)	$\hat{\beta}_C$	SE($\hat{\beta}_C$)
Z_1	0.3046	0.6899	0.4474	0.2286	−0.4428	0.2014*
Z_2	0.0455	0.5997	0.6573	0.2216*	0.2375	0.1796
Z_C	1.2297	0.8174	0.9041	0.2461*	0.5181	0.2046*
Baseline (λ)	2.2715	1.2845	0.0242	0.0053*	0.0321	0.0053*
The first gap time	–	–	0.2327	0.1295	–	–

Z_1: homosexuality, Z_2: heterosexuality, Z_3: bisexuality. $*$: p-value less than 5 %

The duration between initial contraction of HIV and HIV diagnosis does not have a statistically significant effect on the duration between HIV diagnosis and AIDS diagnosis. For the death time, the analysis shows that subjects with bisexuality are associated with an significantly increased risk for death, while homosexual subjects have a lower risk for death compared with the reference group of subjects (i.e., subjects with unknown sexual orientation).

To examine the validity of the model assumptions employed, we use the martingale residuals $\hat{M}_{2i}(\zeta)$ and $\hat{M}_{Ci}(\zeta)$ $(i = 1, \ldots, n)$, which are obtained directly from (6) with the involved parameters substituted with their estimates, and plot the residuals against the linear predictors $\hat{\beta}_2' Z_i$ and $\hat{\beta}_C' Z_i$ $(i = 1, \ldots, n)$. There are no systematic patterns revealed in these residual plots (see Figs. 2, 3), suggesting the adequacy of the model assumptions.

5 Conclusion

This paper is concerned with the setting where the process of disease progression goes through two successive stages, but data on the first gap time are unavailable. The dependence of the second gap time on the first gap time is one of the study interests. We propose a simple bivariate model for the two gap times, allowing the second gap time to depend on the previous one. To overcome the two-fold difficulty caused by missing data on the first gap time, and the induced informative censoring arising in the system of serial events, we impose an additional model for the distribution of the censoring time. However, our simulation results show that, the proposed analysis for the covariate effects on the event times, as well as the effect of the first gap time on the second one, are insensitive to moderate deviation from the assumed model for the censoring time distribution, provided that the sample size is sufficiently large. We derive the cause-specific intensity functions for the available event times based on the proposed modeling framework, by which the observed-data likelihood is obtained. Using the counting process approach and martingale theory, the large sample properties of the maximum likelihood estimator can be readily established.

Due to the two-fold difficulty (missing first-gap time data plus dependent censoring) mentioned above, the current work focuses specifically on parametric models based on

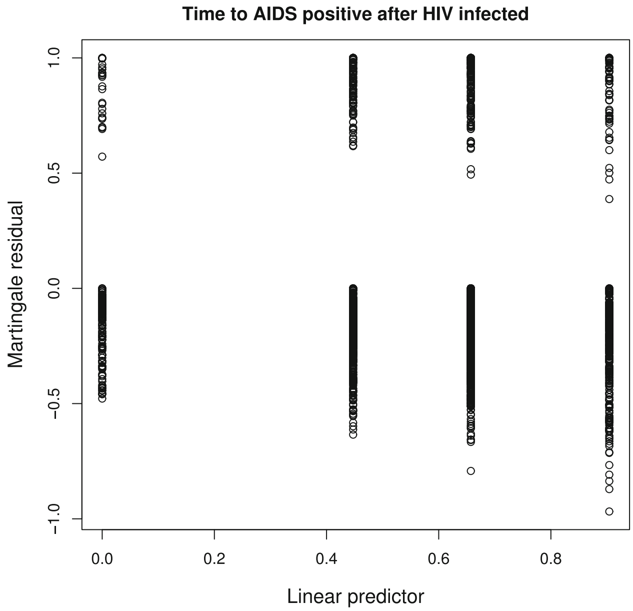

Fig. 2 The estimated martingale residual $\hat{M}_{2i}(\zeta)$ against the linear predictor $\hat{\beta}_2' Z_i$

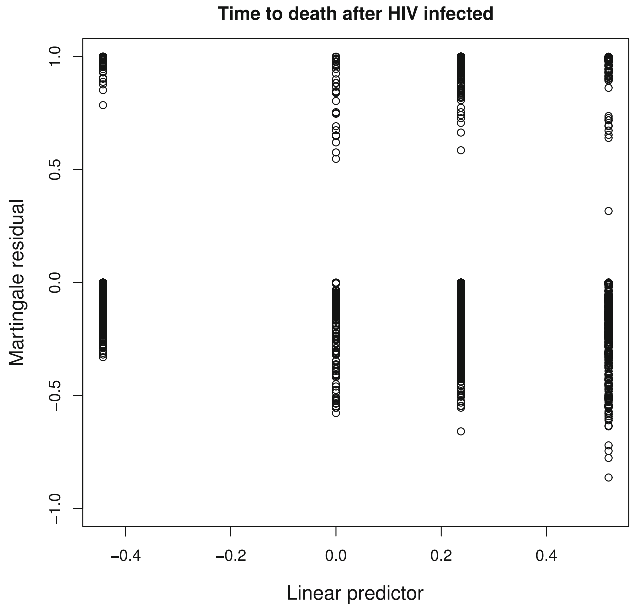

Fig. 3 The estimated martingale residual $\hat{M}_{Ci}(\zeta)$ against the linear predictor $\hat{\beta}_C' Z_i$

exponential distributions. The idea underlying the proposed approach can be similarly extended to other parametric models. However, we are unaware of if there is any other model that will also lead to an analytically tractable model formulation as in the exponential models we consider. Note that the martingale representation of the score functions can facilitate convenient model diagnosis for the proposed method.

Acknowledgements This research was supported by Taiwan Ministry of Science and Techonology Grant 103-2118-M-305-003.

Appendix 1: Proof of Theorem 1

First we shall show that $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{U}_i(\boldsymbol{\Omega})$ at $\boldsymbol{\Omega}_0$ converges to a multivariate normal distribution, where

$$\boldsymbol{U}_i(\boldsymbol{\Omega})=\boldsymbol{S}_{1i}(\boldsymbol{\Omega})+\int_0^{\zeta}\boldsymbol{S}_{2i}(v,\boldsymbol{\Omega})\,dM_{2i}(v)+\int_0^{\zeta}\boldsymbol{S}_{Ci}(\boldsymbol{\Omega})\,dM_{Ci}(v).$$

Note that $\boldsymbol{S}_{ki}$, $k=1,2,3$, have been defined in (8). By the multivariate central limit theorem, we have $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{S}_{1i}(\boldsymbol{\Omega}_0)\to N\left(0, E\left\{-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\boldsymbol{S}_{1i}(\boldsymbol{\Omega}_0)}{\partial\boldsymbol{\Omega}}\right\}\right)$. Since we assume that dN_{2i}, dN_{Ci} cannot jump simultaneously and two counting processes are conditionally independent to δ_{1i} given Z_i,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{S}_{1i}(\boldsymbol{\Omega}_0),\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\zeta}\boldsymbol{S}_{2i}(v,\boldsymbol{\Omega}_0)\,dM_{2i}(v),\text{ and}$$
$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\zeta}\boldsymbol{S}_{Ci}(\boldsymbol{\Omega}_0)\,dM_{Ci}(v),$$

are asymptotically independent. Based on martingale central limit theorem and regular assumptations (A1) − (A5), we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\boldsymbol{S}_{1i}(\boldsymbol{\Omega}_0)+\int_0^{\zeta}\boldsymbol{S}_{2i}(v,\boldsymbol{\Omega}_0)\,dM_{2i}(v)+\boldsymbol{S}_{Ci}(\boldsymbol{\Omega}_0)\,dM_{Ci}(v)\right\}\to N\left(0,\ \mathscr{I}_0\right).$$

Furthermore, the differential of $-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{U}_i(\boldsymbol{\Omega})$ with respect to $\boldsymbol{\Omega}$ is

$$\begin{aligned}&-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\boldsymbol{S}_{1i}(\boldsymbol{\Omega})}{\partial\boldsymbol{\Omega}}\\&+\frac{1}{n}\sum_{i=1}^{n}\int_0^{\zeta}\left\{Y_i(v)\tilde{\eta}_{2i}(v)\lambda_2\boldsymbol{S}_{2i}(v,\boldsymbol{\Omega})^{\otimes 2}+Y_i(v)\tilde{\eta}_{Ci}(v)\lambda_C\boldsymbol{S}_{Ci}(\boldsymbol{\Omega})^{\otimes 2}\right\}dv\\&-\frac{1}{n}\sum_{i=1}^{n}\int_0^{\zeta}\left\{\frac{\partial\boldsymbol{S}_{2i}(v,\boldsymbol{\Omega})}{\partial\boldsymbol{\Omega}}dM_{2i}(v)+\frac{\partial\boldsymbol{S}_{Ci}(\boldsymbol{\Omega})}{\partial\boldsymbol{\Omega}}dM_{Ci}(v)\right\}.\end{aligned}\tag{9}$$

By (A3)–(A4) we have the first two terms on the right-hand side of (9) converge to $\mathscr{I}_0$, and the third term is $o_p(1)$ because the sum is scaled by n^{-1}. So with the Taylor expansion of $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{U}_i(\widehat{\boldsymbol{\Omega}})$ around $\boldsymbol{\Omega}_0$, we have

$$\sqrt{n}\Big(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0, \widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0, \hat{\theta}-\theta_0\Big) = \Big\{-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \boldsymbol{U}_i(\boldsymbol{\Omega}^*)}{\partial \boldsymbol{\Omega}}\Big\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{U}_i(\boldsymbol{\Omega}_0),$$

where $\boldsymbol{\Omega}^*$ is on the line segment between $\widehat{\boldsymbol{\Omega}}$ and $\boldsymbol{\Omega}_0$. Therefore, by Slutsky's theorem and $\widehat{\boldsymbol{\Omega}} \to \boldsymbol{\Omega}_0$ almost surely, Theorem 1 is established.

Appendix 2: Information matrix

For $s, l \in \{1, 2, C\}$, the differential of $-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{U}_i(\boldsymbol{\Omega})$ for each parameter are given

$$\begin{aligned}
I_{\alpha_s\alpha_l} &= (-1)^{I(s\neq l)}I(s\neq 2, l\neq 2)\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\exp(\alpha_1+\beta_1'Z_i)\exp(\alpha_C+\beta_C'Z_i)}{\left\{\exp(\alpha_1+\beta_1'Z_i)+\exp(\alpha_C+\beta_C'Z_i)\right\}^2}\right]\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\int_0^{\zeta}Y_i(v)\tilde{\eta}_{2i}(v)\exp(\alpha_2)\left\{\widetilde{X}_{si}(v)+I(s=2)\right\}\left\{\widetilde{X}_{li}(v)+I(l=2)\right\}dv\\
&\quad+ I(s=l=C)\frac{1}{n}\sum_{i=1}^{n}\int_0^{\zeta}Y_i(v)\tilde{\eta}_{Ci}(v)\exp(\alpha_C)dv\\
&\quad-\frac{1}{n}\sum_{i=1}^{n}\int_0^{\zeta}\left\{\frac{1}{\tilde{\eta}_{2i}(v)}\frac{\partial^2\tilde{\eta}_{2i}(v)}{\partial\alpha_s\partial\alpha_l}-\widetilde{X}_{si}(v)\widetilde{X}_{li}(v)\right\}dM_{2i}(v),\\
I_{\alpha_s\beta_l} &= (-1)^{I(s\neq l)}I(s\neq 2, l\neq 2)\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\exp(\alpha_1+\beta_1'Z_i)\exp(\alpha_C+\beta_C'Z_i)}{\left\{\exp(\alpha_1+\beta_1'Z_i)+\exp(\alpha_C+\beta_C'Z_i)\right\}^2}\right]Z_i\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\int_0^{\zeta}Y_i(v)\tilde{\eta}_{2i}(v)\exp(\alpha_2)\left\{\widetilde{X}_{si}(v)+I(s=2)\right\}\widetilde{Z}_{li}(v)\,dv\\
&\quad+ I(s=l=C)\frac{1}{n}\sum_{i=1}^{n}\int_0^{\zeta}Y_i(v)\tilde{\eta}_{Ci}(v)\exp(\alpha_C)Z_i\,dv\\
&\quad-\frac{1}{n}\sum_{i=1}^{n}\int_0^{\zeta}\left\{\frac{1}{\tilde{\eta}_{2i}(v)}\frac{\partial^2\tilde{\eta}_{2i}(v)}{\partial\alpha_s\partial\beta_l}-\widetilde{X}_{si}(v)\widetilde{Z}_{li}(v)\right\}dM_{2i}(v),\\
I_{\alpha_s\theta} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\zeta}Y_i(v)\tilde{\eta}_{2i}(v)\exp(\alpha_2)\left\{\widetilde{X}_{si}(v)+I(s=2)\right\}\widetilde{W}_i'(v)\,dv\\
&\quad-\frac{1}{n}\sum_{i=1}^{n}\int_0^{\zeta}\left\{\frac{1}{\tilde{\eta}_{2i}(v)}\frac{\partial^2\tilde{\eta}_{2i}(v)}{\partial\alpha_s\partial\theta}-\widetilde{X}_{si}(v)\widetilde{W}_i'(v)\right\}dM_{2i}(v),\\
I_{\beta_s\beta_l} &= (-1)^{I(s\neq l)}I(s\neq 2, l\neq 2)\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\exp(\alpha_1+\beta_1'Z_i)\exp(\alpha_C+\beta_C'Z_i)}{\left\{\exp(\alpha_1+\beta_1'Z_i)+\exp(\alpha_C+\beta_C'Z_i)\right\}^2}\right]Z_i^{\otimes 2}\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\int_0^{\zeta}Y_i(v)\tilde{\eta}_{2i}(v)\exp(\alpha_2)\widetilde{Z}_{si}(v)\widetilde{Z}_{li}'(v)\,dv
\end{aligned}$$

$$+ I(s = l = C)\frac{1}{n}\sum_{i=1}^{n}\int_0^{\zeta} Y_i(v)\tilde{\eta}_{Ci}(v)\exp(\alpha_C)Z_i^{\otimes 2}dv$$

$$-\frac{1}{n}\sum_{i=1}^{n}\int_0^{\zeta}\left\{\frac{1}{\tilde{\eta}_{2i}(v)}\frac{\partial^2\tilde{\eta}_{2i}(v)}{\partial\beta_s\partial\beta_l} - \widetilde{Z}_{si}(v)\widetilde{Z}'_{li}(v)\right\}dM_{2i}(v),$$

$$\begin{aligned}
I_{\beta_s\theta} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\zeta} Y_i(v)\tilde{\eta}_{2i}(v)\exp(\alpha_2)\widetilde{Z}_{si}(v)\widetilde{W}'_i(v)\,dv \\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\zeta}\left\{\frac{1}{\tilde{\eta}_{2i}(v)}\frac{\partial^2\tilde{\eta}_{2i}(v)}{\partial\beta_s\partial\theta} - \widetilde{Z}_{si}(v)\widetilde{W}'_i(v)\right\}dM_{2i}(v), \\
I_{\theta\theta} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\zeta} Y_i(v)\tilde{\eta}_{2i}(v)\exp(\alpha_2)\widetilde{W}_i^2(v)dv \\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\zeta}\left\{\frac{1}{\tilde{\eta}_{2i}(v)}\frac{\partial^2\tilde{\eta}_{2i}(v)}{\partial^2\theta} - \widetilde{W}_i^2(v)\right\}dM_{2i}(v),
\end{aligned}$$

References

Chang SH, Wang MC (1999) Conditional regression analysis for recurrence time data. J Am Stat Assoc 94:1221–1230

Cook RJ, Lawless JF (2007) The statistical analysis of recurrent events. Springer, New York

Huang J (1996) Efficient estimation for the proportional hazards model with interval censoring. Ann Stat 24:540–568

Huang Y, Louis TA (1998) Nonparametri estimation of the joint distribution of survival time and mark variables. Biometrika 85:785–798

Huang Y (2000) Multistate accelerated sojourn time model. J Am Stat Assoc 95:619–627

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

Lin DY, Ying Z (1997) Additive regression models for survival data Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis. Springer, New York

Lin DY, Sun W, Ying Z (1999) Nonparametric estimation of gap time distributions for serial events with censored data. Biometrika 86:59–70

Schaubel DE, Cai J (2004) Regression methods for gap time hazard functions of sequentially ordered multivariate failure time data. Biometrika 91:291–303

Sun LQ, Park DH, Sun JG (2006) The additive hazards model for recurrent gap times. Stat Sin 16:919–932

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

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

Wang W, Ding AA (2000) On assessing the association for bivariate current status data. Biometrika 87:879–893

Wang MC, Qin J, Chiang CT (2001) Analyzing recurrent event data with informative censoring. J Am Stat Assoc 96:1057–1065

Wang MC, Chiang CT (2002) Nonparametric methods for recurrent event data with informative and non-informative censorings. Stat Med 21:445–456

Zeng D, Lin DY (2006) Efficient estimation of semiparametric transformation models for counting processes. Biometrika 93:627–640