Statistica Sinica Preprint No: SS-2018-0224	
Title	Generalized scale-change models for recurrent event processes under informative censoring
Manuscript ID	SS-2018-0224
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0224
Complete List of Authors	Gongjun Xu Sy Han Chiou Jun Yan Kieren Marr and Chiung-Yu Huang
Corresponding Author	Gongjun Xu
E-mail	gongjun@umich.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

Statistica Sinica
doi:http://dx.doi.org/10.5705/ss.20XX.XXX

Generalized scale-change models for recurrent event processes under informative censoring

Gongjun Xu[1], Sy Han Chiou[2], Jun Yan[3], Kieren Marr[4], and Chiung-Yu Huang[5].

[1] *University of Michigan,* [2] *University of Texas at Dallas,* [3] *University of Connecticut,*

[4] *Johns Hopkins University, and* [5] *University of California, San Francisco*

Abstract: Two major challenges arise in regression analyses of recurrent event data. First, popular existing models, such as the Cox proportional rates model, may not fully capture the covariate effects on the underlying recurrent event process. Second, the censoring time remains informative about the risk of experiencing recurrent events after accounting for covariates. We address both challenges using a general class of semiparametric scale-change models that allow both scale-change and multiplicative covariate effects. The proposed model is flexible, and includes several existing models as special cases, including the popular proportional rates model, accelerated mean model, and accelerated rate model. Moreover, it accommodates informative censoring through a subject-level latent frailty, the distribution of which is left unspecified. A robust estimation procedure is proposed to estimate the model parameters that does not require a parametric assumption on the distribution of the frailty, or a Poisson assumption on the recurrent event process. The asymptotic properties of the resulting estimator are established, with the asymptotic variance estimated using a novel resampling approach.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

As a byproduct, the structure of the model provides a model selection approach for the submodels that employs hypothesis testing of the model parameters. Numerical studies show that the proposed estimator and the model selection procedure perform well under both noninformative and informative censoring scenarios. Lastly, the methods are applied to data from two transplant cohorts to study the risk of infection after transplantation.

Key words and phrases: Accelerated failure time model; Accelerated rate model; Cox model; Frailty model; Hypothesis testing; Model selection; Resampling.

1. Introduction

The importance of analyzing recurrent event data is widely recognized in many fields, including medicine, public health, cybersecurity, engineering, and the social sciences (Wei and Glidden, 1997; Cook and Lawless, 2007). Examples of recurrent events include opportunistic infections experienced by patients who undergo a hematopoietic stem cell transplantation (Marr, 2012), repeated cardiovascular events in survivors of myocardial infarction (Rogers et al., 2012), episodes of schizophrenia in chronic schizophrenic patients (Eaton et al., 1992), cyber attacks on network systems (Benjamin et al., 2016), and breakdowns of repairable systems (Nelson, 2003). Various regression models have been proposed to evaluate covariate effects on the risk of recurrent events. Pepe and Cai (1993), Lawless and Nadeau (1995), and Lin et al. (2000) considered Cox-type proportional rates/intensities models that postulate multiplica-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

tive covariate effects on the baseline rate/intensity function of the recurrent event process. Lin et al. (1998) studied the accelerated mean model, where covariates modify the timescale of the cumulative mean function. Chen and Wang (2000) and Ghosh (2004) presented accelerated rate/intensity models that formulate covariate effects to change the time-scale directly on the baseline rate/intensity function. These three types of models are covered as special cases in a general class of regression models proposed by Sun and Su (2008). Other covariate effect formulations include the additive rate models (Schaubel et al., 2006) and the additive-multiplicative rate models (Liu et al., 2010). The aforementioned methods all require a noninformative censoring assumption; that is, the censoring time is conditionally independent of the recurrent event process, given the observed covariates.

In many applications, an informative censoring event or a terminal event, such as graft failure or death, respectively, can terminate an observation. Failing to account for informative censoring can lead to substantial bias in inferences and, thus, invalid conclusions (e.g., Cook and Lawless, 1997; Ghosh and Lin, 2002; Luo et al., 2010). A popular approach to accommodate informative censoring is joint modeling, where the association between the failure event and the recurrent event process is modeled via a shared frailty (Lancaster and Intrator, 1998; Liu et al., 2004; Ye et al., 2007; Zeng and Lin, 2009; Kalbfleisch et al., 2013). An advantage of the frailty is that it accounts for heterogeneity that cannot be explained by the observed covariates. Nonetheless,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

inferences on the shared frailty model often require a parametric assumption on the frailty distribution and the correct modeling of the terminal event, which are a nuisance when the primary interest is the covariate effects on the risk of recurrent events. Despite the importance for inferences of formal checks for the frailty distribution and the model specification of the terminal event, research on this topic remains scarce. An alternative approach is to relax the parametric assumption on the shared frailty model. For instance, Wang et al. (2001), Huang and Wang (2004), and Huang et al. (2010) considered Cox-type models, and Xu et al. (2017) studied a joint scale-change model of the recurrent event process and the terminal event.

We propose an approach that allows a flexible form of informative censoring in a generalized scale-change model for recurrent event processes. Our model encompasses two types of covariate effects: a scale-change effect, which alters the timescale, and a multiplicative effect, which modifies the risk. A similar modeling approach has been studied for univariate survival data (Chen and Jewell, 2001) and recurrent event data (Sun and Su, 2008) under conditionally independent censoring, given the observed covariates. Similarly to the method of Sun and Su (2008), the proposed formulation includes Cox-type models, the accelerated mean model, and the accelerated rate model as special cases. However, in contrast to Sun and Su (2008), the recurrent event process in our study is associated with the censoring time through an unobserved, subject-specific frailty, and no parametric assumption about the frailty distribution is required.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

Using the mean structure shared by the conditional distribution of recurrent events and the order statistics of a set of right-truncated failure times, we embed the problem into an estimation using right-truncated data, and develop a novel semiparametric estimation procedure that does not require information about the frailty variable. The asymptotic normality of the resulting estimator is established without the strong Poisson-type assumption for the recurrent event process, or parametric assumptions for the frailty variable. The asymptotic variance is estimated from an efficient resampling-based sandwich estimator. The structure of the model facilitates model selection from among the submodels via hypothesis testing of the model parameters. Our numerical studies confirm the validity of the proposed methods and the model selection procedure.

2. Model Setup

Suppose $[0, \tau]$ is the period of a study during which recurrent events can potentially be observed up to time τ. For a subject, let $N(t)$ be the number of events in interval $[0, t]$, and X be a $p \times 1$ covariate vector. Let C be a noninformative censoring time, such as the end of the study, which is independent of $N(\cdot)$, given X. Let D be an informative censoring time, such as death, which is associated with $N(\cdot)$, even after conditioning on X. Define the follow-up time as $Y = \min(C, D, \tau)$. The observed data are independent and identically distributed (i.i.d.) copies $\{N_i(t), Y_i, X_i : t \leq Y_i,\ i = 1, \ldots, n\}$. Let m_i

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

be the number of events of subject i before time Y_i. If $m_i > 0$, the jump times of $N_i(t)$ are the observed event times t_{ij}, for $j = 1, \ldots, m_i$.

Our model formulates the rate function for the counting process $N_i(t)$, $\lambda_i(t)\,\mathrm{d}t = E\{\mathrm{d}N_i(t) \mid Z_i, X_i\}$, given the covariate vector X_i and an unobserved subject-specific nonnegative frailty Z_i. Specifically, we postulate that

$$\lambda_i(t) = Z_i \lambda_0(t e^{X_i^\top \alpha}) e^{X_i^\top \beta}, \qquad t \in [0, \tau], \tag{2.1}$$

where Z_i has an unspecified distribution, with $E(Z_i^2) < \infty$, α and β are both $p \times 1$ vectors of parameters, and $\lambda_0(t)$ is an unspecified, nonWeibull baseline rate function. As in Sun and Su (2008), the Weibull baseline $\lambda_0(t) \propto t^q$, for some q, is excluded for identifiability between α and β. Define the corresponding cumulative baseline rate function as $\Lambda_0(t) = \int_0^t \lambda_0(u)\,\mathrm{d}u$. For identifiability between Z_i and $\lambda(t)$, we assume $\Lambda_0(\tau) = 1$ and $E(Z_i \mid X_i) = \mu_Z$; that is, the conditional mean of Z, given X, does not depend on X. Because Z_i has a multiplicative effect on the rate function, Model (2.1) allows the event occurrence rate to be inflated (or deflated) by the frailty variable Z_i, with an arbitrary distribution. We assume that Y_i is independent of $N_i(\cdot)$, given (Z_i, X_i). The dependence between Y_i and $N_i(\cdot)$, unconditional on (Z_i, X_i), can be either positive or negative, depending on whether the association between Y_i and Z_i is positive or negative, given X_i.

Model (2.1) offers great flexibility and includes several popular semiparametric models for recurrent event processes as special cases. When $\beta = 0$, it reduces to the

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

frailty accelerated rate model $\lambda_i(t) = Z_i\lambda_0(te^{X_i^\top\alpha})$, the special case of which, with a degenerate frailty distribution, was considered in Ghosh (2004). Covariate effects under the accelerated rate model modify the timescale of the rate function, and allow for identical risk at time zero, a desirable property for modeling recurrent events in randomized clinical trials. When $\alpha = \beta$, Model (2.1) reduces to the frailty accelerated mean model $\lambda_i(t) = Z_i\lambda_0(te^{X_i^\top\alpha})e^{X_i^\top\alpha}$ proposed by Xu et al. (2017), where the covariate effects modify the timescale of the cumulative mean function of the recurrent event process by a factor of $e^{X_i^\top\alpha}$. When $\alpha = 0$, Model (2.1) reduces to the popular frailty Cox-type regression model $\lambda_i(t) = Z_i\lambda_0(t)e^{X_i^\top\beta}$; see Lancaster and Intrator (1998), Wang et al. (2001), and Huang and Wang (2004). In this case, the covariate effects modify the magnitude of the rate function by a factor of $e^{X_i^\top\beta}$. Similarly to Sun and Su (2008), the three submodels coincide if and only if $\lambda_0(t)$ is of the Weibull form. The flexible formulation of Model (2.1) offers a new framework to test, diagnose, and compare these submodels using hypothesis tests on the parameters α and β; see Section 3.

Interpretations of the covariate effects under Model (2.1) involve two types of modifications on the rate function: a scale-change effect, which alters the timescale by a factor of $e^{X_i^\top\alpha}$, and a multiplicative effect, which modifies the magnitude of the rate function by a factor of $e^{X_i^\top\beta}$. The effects are easily seen when X contains a single treatment indicator. In this case, e^β characterizes the risk ratio between

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

treated subjects ($X = 1$) at time t and untreated subjects ($X = 0$) at time te^{α}. When $\alpha = \beta$, the combined changes in timescale and magnitude are such that the resulting cumulative mean function has a timescale modification. This motivates a useful alternative presentation of Model (2.1). Let $\Lambda_i(t) = E\{N_i(t) \mid Z_i, X_i\}$ be the conditional cumulative mean function. Then,

$$\Lambda_i(t) = Z_i \Lambda_0(te^{X_i^\top \alpha}) e^{X_i^\top \gamma}, \qquad t \in [0, \tau],$$

where $\gamma = \beta - \alpha$. The parameters α and γ can then be interpreted as the scale-change effect parameter and the multiplicative effect parameter, respectively, on the conditional cumulative mean function. In the case of a single treatment indicator covariate, the expected number of events by time t among the treated subjects ($X = 1$) is equal to e^{γ} multiplied by the expected number of events by time te^{α} in the control group ($X = 0$).

3. Estimation and Inference

3.1 Parameter Estimation

For a p-dimensional vector a, consider the transformations $t_{ij}^*(a) = t_{ij} e^{X_i^\top a}$ and $Y_i^*(a) = Y_i e^{X_i^\top a}$. Let $R_i^*(t, a) = \sum_{j=1}^{m_i} I\{t_{ij}^*(a) \le t \le Y_i^*(a)\}$; hereafter, when $m_i = 0$, we define the summation operator $\sum_{j=1}^{m_i}$ to be zero. Define the counting process for the transformed event times as $N_i^*(t, a) = \sum_{j=1}^{m_i} I\{t_{ij}^*(a) \le t \wedge Y_i^*(a)\}$, such that $N_i^*(t, a) = N_i(te^{-X_i^\top a} \wedge Y_i)$, where $\wedge$ is the minimum operator. Under Model (2.1), for

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

$t \leq Y_i$,

$$E\{N_i(t) \mid X_i, Z_i, Y_i\} = \int_0^t Z_i \lambda_0(ue^{X_i^\top \alpha}) e^{X_i^\top \beta} \, \mathrm{d}u = Z_i \Lambda_0(te^{X_i^\top \alpha}) e^{X_i^\top (\beta - \alpha)},$$

and, thus, for $t \leq Y_i^*$, $E\{N_i^*(t, a) \mid X_i, Z_i, Y_i^*\} = Z_i \Lambda_0\{te^{X_i^\top (\alpha - a)}\} e^{X_i^\top (\beta - \alpha)}$.

We develop a novel semiparametric estimation procedure that does not require a distributional assumption about the frailty variable. To motivate the procedure, we first consider the case where the true parameter is known, or $a = \alpha$. For this special case, we suppress α in the notation whenever there is no ambiguity, and write $t_{ij}^* = t_{ij} e^{X_i^\top \alpha}$, $Y_i^* = Y_i e^{X_i^\top \alpha}$, $N_i^*(t) = N_i^*(t, \alpha)$, and $R_i^*(t) = R_i^*(t, \alpha)$. When $a = \alpha$ we have $E\{N_i^*(t) \mid X_i, Z_i, Y_i\} = Z_i \Lambda_0(t) e^{X_i^\top (\beta - \alpha)}$, for $t \leq Y_i$, which implies that the rate function of the underlying transformed process follows the Cox-type proportional rates model with a multiplicative frailty. Because the informative censoring time Y_i depends on both X_i and Z_i, conventional methods that require independent censoring may lead to a biased estimation. Moreover, the estimation procedures in Xu et al. (2017) cannot be applied, because the transformed process remains dependent on X_i.

We now embed the problem into a seimparametric estimation with clustered right-truncated data by using a mean structure shared by the conditional distribution of the recurrent events and the order statistics of a set of (possibly correlated) right-truncated failure times. Given (m_i, Y_i^*), consider a set of m_i random variables, denoted

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

by $\widetilde{t}^*_{i1}, \ldots \widetilde{t}^*_{im_i}$, such that each has a *marginal* density function free of (Z_i, X_i):

$$\frac{\lambda_0(t)}{\Lambda_0(Y_i^*)}, \qquad 0 \le t \le Y_i^*. \tag{3.2}$$

It follows from $\Lambda_0(\tau) = 1$ that $\Lambda_0(t)$ defines a proper distribution function and, thus, (3.2) can be viewed as a truncation density function (Wang et al., 2001; Xu et al., 2017). In other words, $\widetilde{t}^*_{ij}$ can be viewed as a right-truncated failure time with truncation time Y_i^*. We show in Proposition 1 that the transformed recurrent event times $\{t^*_{i1}, \ldots, t^*_{im_i}\}$ of the ith individual share the same mean structure with the right-truncated failure times $\{\widetilde{t}^*_{i1}, \ldots, \widetilde{t}^*_{im_i}\}$.

Proposition 1. *Consider the counting process induced by the right-truncated random variables $\widetilde{t}^*_{ij}$, for $j = 1, \ldots, m_i$; that is, $\widetilde{N}^*_i(t) = \sum_{i=1}^{m_i} I(\widetilde{t}^*_{ij} \le t)$, for $t \le Y_i^*$. Then, we have $E\{\widetilde{N}^*_i(t) \mid Z_i, X_i, Y_i^*\} = E\{N^*_i(t) \mid Z_i, X_i, Y_i^*\}$.*

Because $E(m_i \mid Z_i, X_i, Y_i^*) = Z_i \Lambda_0(Y^*) e^{X_i^\top(\beta-\alpha)}$, Proposition 1 follows from $E\{\widetilde{N}^*_i(t) \mid Z_i, X_i, Y_i^*\} = E[E\{\widetilde{N}^*_i(t) \mid m_i, Z_i, X_i, Y_i^*\} \mid Z_i, X_i, Y_i^*] = E\{m_i \Lambda_0(t)/\Lambda_0(Y_i^*) \mid Z_i, X_i, Y_i^*\} = Z_i \Lambda_0(t) e^{X_i^\top(\beta-\alpha)} = E\{N^*_i(t) \mid Z_i, X_i, Y_i^*\}$. This mathematical equivalence motivates us to extend the methods for independent right-truncated survival data (Kalbfleisch and Lawless, 1991; Wang, 1989) to the context of clustered right-truncated data in order to construct the estimating equations. Specifically, define $\widetilde{N}^*_{ij}(t) = I(\widetilde{t}^*_{ij} \le t \wedge Y_i^*)$, $\widetilde{R}^*_{ij}(t) = I(\widetilde{t}^*_{ij} \le t \le \widetilde{Y}^*_i)$, and $\widetilde{R}^*_i(t) = \sum_{j=1}^{m_i} \widetilde{R}^*_{ij}(t)$. It is known that, with right-truncated data, $\widetilde{N}^*_{ij}(\tau - t) - \int_{\tau-t}^{M} \widetilde{R}^*_{ij}(u)\, \mathrm{d}H(u)$, where $H(u) = \log \Lambda_0(u)$,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

defines a martingale on $[0,\tau]$. Simple algebra, with a change of variable, yields $E\{\mathrm{d}\widetilde{N}^*_{ij}(t) - \widetilde{R}^*_{ij}(t)\,\mathrm{d}H(t)\} = 0$, for $t \in [0,\tau]$, which defines a zero-mean process $\widetilde{M}^*_{ij}(t) = \widetilde{N}^*_{ij}(t) - \int_t^\tau \widetilde{R}^*_{ij}(u)\,\mathrm{d}H(u)$. Furthermore, the stochastic process based on the clustered right-truncated data $\{\tilde{t}^*_{i1},\ldots,\tilde{t}^*_{im_i}\}$,

$$\widetilde{M}^*_i(t) = \sum_{j=1}^{m_i} \widetilde{M}^*_{ij}(t) = \sum_{j=1}^{m_i} \widetilde{N}^*_{ij}(t) - \sum_{j=1}^{m_i} \int_t^\tau \widetilde{R}^*_{ij}(u)\,\mathrm{d}H(u) = \widetilde{N}^*_i(t) - \int_0^t \widetilde{R}^*_i(u)\,\mathrm{d}H(u),$$

has a zero mean. Following the above discussion, we can treat the observations $\{t^*_{i1},\ldots,t^*_{im_i}\}$ as the order statistics of $\{\tilde{t}^*_{i1},\ldots,\tilde{t}^*_{im_i}\}$; in this case, $N^*_i(t) = \widetilde{N}^*_i(t)$ and $R^*_i(t) = \widetilde{R}^*_i(t)$. Therefore, the stochastic process $M^*_i(t) = N^*_i(t) - \int_0^t R^*_i(u)\,\mathrm{d}H(u)$, for $t \in [0,\tau]$, though not a martingale, also has a zero mean. The proof is given in Appendix S1 of the Supplementary Material. Moreover, for all $t \in [0,\tau]$, we have the following equations:

$$E\left\{\sum_{i=1}^n \int_0^t \mathrm{d}M^*_i(u)\right\} = 0 \text{ and } E\left\{\sum_{i=1}^n \int_0^\tau X_i\,\mathrm{d}M^*_i(u)\right\} = 0. \tag{3.3}$$

Note that t^*_{ij} may be correlated, but the estimating equations in (3.3) remain unbiased. The first term in (3.3) introduces a consistent estimator for H, via

$$\mathrm{d}\hat{H}(t) = \frac{\sum_{i=1}^n \mathrm{d}N^*_i(t)}{\sum_{i=1}^n R^*_i(t)}.$$

The proof of the consistency of $\hat{H}(\cdot)$ is given in Appendix S1 of the Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

Replacing H with $\hat{H}$ in (3.3), we estimate α by solving the following equation:

$$S_n(a) := n^{-1} \sum_{i=1}^{n} \int_0^\tau \left\{ X_i - \frac{\mathcal{R}_n^{(1)}(u,a)}{\mathcal{R}_n^{(0)}(u,a)} \right\} \mathrm{d}N_i^*(t,a) = 0, \tag{3.4}$$

where $\mathcal{R}_n^{(k)}(t,a) = \sum_{i=1}^n X_i^k R_i^*(t,a)$, for $k \in \{0,1\}$. The estimating function $S_n(a)$ is similar to the accelerated failure time model with truncated data (Lai and Ying, 1991), and can be solved using, for example, the derivative-free algorithm of Barzilai and Borwein (1988), implemented in Varadhan and Gilbert (2009). Let $\hat{\alpha}_n$ be the solution to (3.4). It is easy to see that H can be estimated by

$$\hat{H}_n(t;\hat{\alpha}_n) = -\int_t^\tau \frac{\sum_{i=1}^n \mathrm{d}N_i^*(u;\hat{\alpha}_n)}{\sum_{i=1}^n R_i^*(u;\hat{\alpha}_n)},$$

and, thus, Λ_0 can be estimated by $\hat{\Lambda}_n(t) = \exp\{\hat{H}_n(t;\hat{\alpha}_n)\}$.

With α estimated, we now estimate γ, defined earlier as $\gamma = \beta - \alpha$. It follows from (2.1) that

$$\begin{aligned} E\left[m_i \Lambda_0^{-1}(Y_i^*) \mid X_i\right] &= E\left[E\{m_i \mid X_i, Y_i^*, Z_i\}\Lambda_0^{-1}(Y_i^*) \mid X_i\right] \\ &= E\left[Z_i \exp(X_i^\top \gamma) \mid X_i\right] = \exp(\bar{X}_i^\top \theta), \end{aligned}$$

where $\bar{X}_i^\top = (1, X_i^\top)$ and $\theta^\top = (\log \mu_Z, \gamma^\top)$. This expectation suggests the following estimating equation if α and Λ_0 are known: $n^{-1}\sum_{i=1}^n \bar{X}_i^\top \left\{m_i \Lambda_0^{-1}(Y_i^*) - \exp(\bar{X}_i^\top \theta)\right\} = 0$. The estimator for θ, denoted by $\hat{\theta}_n$, can be obtained by solving the following estimating equation, with α and Λ_0 replaced by their estimators from the first step:

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

$$U_n(\theta;\hat{\alpha}_n) := n^{-1}\sum_{i=1}^{n}\bar{X}_i^{\top}\left[m_i\hat{\Lambda}_n^{-1}\{Y_i^*(\hat{\alpha}_n)\} - \exp(\bar{X}_i^{\top}\theta)\right] = 0. \tag{3.5}$$

Then, β can be estimated by $\hat{\beta}_n = \hat{\alpha}_n + \hat{\gamma}_n$. Given $\hat{\alpha}_n$ and $\hat{\Lambda}_n$, the estimating equation in (3.5) is monotone and continuously differentiable with respect to θ; hence, its root can be obtained easily using standard software.

3.2 Asymptotic Theory and Variance Estimation

To study the large-sample properties of the proposed estimators, we impose the following regularity conditions.

Condition 1 $\Pr(Y^* \geq \tau) > 0$, where $Y^* = Ye^{X^{\top}\alpha}$.

Condition 2 The covariate X is bounded; the latent variable Z is positive, with $E(Z^2) < \infty$.

Condition 3 The conditional probability density function of Y, given (Z, X), is continuous and uniformly bounded.

Condition 4 The rate function $\lambda_0(t)$, for $t \in [0, \tau]$, is strictly bounded below by zero, and has a bounded second derivative function.

Condition 5 The matrices J and J_2, defined in Appendix S1 of the Supplementary Material, are nonsingular.

Conditions 1–5 are common assumptions in survival models. Condition 4 imposes

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

the bounded second derivative function of $\lambda_0(t)$, which is usually required by the accelerated failure time model to evaluate the asymptotic covariance matrix. Based on these regularity conditions, we have the following asymptotic results; the proofs are provided in Appendix S1 of the Supplementary Material.

Theorem 1. *Under Conditions 1–5, $n^{1/2}(\hat{\alpha}_n - \alpha, \hat{\beta}_n - \beta)$ converges weakly to a multivariate normal distribution with mean zero and covariance matrix $\Sigma(\alpha, \beta)$, specified in Appendix S1 of the Supplementary Material. Furthermore, for the estimated baseline rate function, we have that $n^{1/2}\{\hat{\Lambda}_n(t, \hat{\alpha}_n) - \Lambda_0(t)\}$, for $t \in [0, \tau]$, converges weakly to a zero-mean Gaussian process.*

Theorem 1 allows us to use the asymptotic joint Gaussian distribution of $n^{1/2}(\hat{\alpha}_n - \alpha,\ \hat{\beta}_n - \beta)$ to make inferences on the model parameters. Because the limiting covariance matrix $\Sigma(\alpha, \beta)$ depends on the unknown density functions of the censoring time, it may be computationally difficult and inefficient to estimate it directly from the data. Therefore, we propose an efficient resampling approach to estimate the covariance matrix $\Sigma(\alpha, \beta)$.

We first describe an approach to estimate the covariance of $n^{1/2}(\hat{\alpha}_n - \alpha,\ \hat{\theta}_n - \theta)$, denoted by $\Sigma(\alpha, \theta)$, and then use it to retrieve the estimation of $\Sigma(\alpha, \beta)$. From the proof of Theorem 1,

$$n^{1/2} \begin{pmatrix} \hat{\alpha}_n - \alpha \\ \hat{\theta}_n - \theta \end{pmatrix} = n^{1/2} J_{\alpha,\theta}^{-1} \begin{pmatrix} S_n(\alpha) \\ U_n(\theta; \alpha) \end{pmatrix} + o_p(1),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

where $J_{\alpha,\theta}$ is the slope matrix

$$J_{\alpha,\theta} = \begin{pmatrix} J & 0 \\ J_1 & J_2 \end{pmatrix},$$

with J, J_1, and J_2 defined in Appendix S1 of the Supplementary Material. This implies that $\Sigma(\alpha,\theta)$ has the following sandwich form: $J_{\alpha,\theta}^{-1} V_{\alpha,\theta} (J_{\alpha,\theta}^{-1})^\top$, where $V_{\alpha,\theta}$ is the limiting covariance matrix of $n^{1/2}\{S_n^\top(\alpha), U_n^\top(\theta;\alpha)\}$. The proposed resampling approach estimates the two components $V_{\alpha,\theta}$ and $J_{\alpha,\theta}$ separately, and requires neither density estimations nor intensive computation.

Step 1: Estimation of $V_{\alpha,\theta}$ Let $(\xi_1, \ldots, \xi_n)$ be a set of i.i.d. positive random variables with unit mean and unit variance (e.g., standard exponential). Then, we define the perturbed estimating functions as follows:

$$S_n^\dagger(\alpha) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \int_0^\tau \left\{ \xi_i X_i - \frac{\sum_{k=1}^{n} \sum_{l=1}^{m_k} \xi_k X_k R_{k,l}^*(t,\alpha)}{\sum_{k=1}^{n} \sum_{l=1}^{m_k} \xi_k R_{k,l}^*(t,\alpha)} \right\} \mathrm{d}N_{ij}^*(t,\alpha),$$

and

$$U_n^\dagger(\theta;\alpha) = n^{-1} \sum_{i=1}^{n} \xi_i \bar{X}_i^\top \left[\frac{m_i}{\hat{\Lambda}_n^\dagger\{Y_i^*(\alpha)\}} - \exp(\bar{X}_i^\top \theta) \right],$$

where

$$\hat{\Lambda}_n^\dagger(t) = \exp\{\hat{H}_n^\dagger(t;\hat{\alpha}_n)\} \text{ and } \hat{H}_n^\dagger(t;\hat{\alpha}_n) = \int_0^t \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} \xi_i \, \mathrm{d}N_{ij}^*(u;\hat{\alpha}_n)}{\sum_{i=1}^{n} \sum_{j=1}^{m_i} \xi_i R_{ij}^*(u;\hat{\alpha}_n)}.$$

Following the arguments in Zeng and Lin (2008), $n^{1/2}\{S_n^\dagger(\hat{\alpha}_n), U_n^\dagger(\hat{\theta}_n;\hat{\alpha}_n)\}$, conditional on the observed data, has the same asymptotic distribution as $n^{1/2}\{S_n(\alpha), U_n(\theta;\alpha)\}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

evaluated at the true parameters. Thus, a consistent estimator of $V_{\alpha,\theta}$, denoted by $\hat{V}_{\hat{\alpha}_n,\hat{\theta}_n}$, is given by the sample variance of the perturbed replicates of the derivative-free Barzilai–Borwein spectral algorithm, $n^{-1/2}\{S_n^{\dagger}(\hat{\alpha}_n), U_n^{\dagger}(\hat{\theta}_n;\hat{\alpha}_n)\}$.

Step 2: Estimation of $J_{\alpha,\theta}$ Estimating the slope matrix, $J_{\alpha,\theta}$, is challenging, owing to the nonsmoothness of the estimating functions. For a $(2p+1)$-dimensional vector, $s=(s_1,s_2)\in R^{2p+1}$, and (a,r) in a small neighborhood of (α,θ), such that $\|(a,r)-(\alpha,\theta)\|\to 0$, the proof of Theorem 1 implies that the estimating functions can be uniformly decomposed into

$$n^{1/2}\begin{pmatrix} S_n(a+n^{-1/2}s_1)-S_n(a), \\ U_n(r+n^{-1/2}s_2;a+n^{-1/2}s_1)-U_n(r;a)\end{pmatrix} = J_{\alpha,\theta}\begin{pmatrix} s_1 \\ s_2\end{pmatrix}+o_p(1).$$

Because $S_n(\hat{\alpha}_n)=0$ and $U_n(\hat{\gamma}_n,\hat{\alpha}_n)=0$, we have

$$n^{1/2}\begin{pmatrix} S_n(\hat{\alpha}_n+n^{-1/2}s_1) \\ U_n(\hat{\theta}_n+n^{-1/2}s_2;\hat{\alpha}_n+n^{-1/2}s_1)\end{pmatrix} = J_{\alpha,\theta}\begin{pmatrix} s_1 \\ s_2\end{pmatrix}+o_p(1).$$

The above equation presents an asymptotic linear relationship between the estimating equations. Motivated by these results, the jth row of $J_{\alpha,\theta}$ can be approximated by regressing the jth component of $n^{1/2}\{S_n(\hat{\alpha}_n+n^{-1/2}s_1), U_n(\hat{\theta}_n+n^{-1/2}s_2;\hat{\alpha}_n+n^{-1/2}s_1)\}$ on s, which is generated from a $(2p+1)$-dimensional standard normal distribution. Putting the estimated regression coefficients into a matrix gives an estimator $\hat{J}_{\hat{\alpha}_n,\hat{\theta}_n}$ of $J_{\alpha,\theta}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

The target sandwich variance matrix $\Sigma(\alpha, \theta)$ is then estimated by $\hat{\Sigma}(\hat{\alpha}_n, \hat{\theta}_n) = \hat{J}^{-1}_{\hat{\alpha}_n,\hat{\theta}_n} \hat{V}_{\hat{\alpha}_n,\hat{\theta}_n} (\hat{J}^{-1}_{\hat{\alpha}_n,\hat{\theta}_n})^\top$. Compared with the conventional bootstrap methods, which require that we solve estimating equations repeatedly, the proposed resampling approach is computationally much more efficient, because it requires only evaluations of (rather than a solution to) the perturbed estimating functions and that we perform least squares regressions. Using $\hat{\Sigma}(\hat{\alpha}_n, \hat{\theta}_n)$, the estimated covariance matrix of $\Sigma(\alpha, \beta)$ can be obtained as $\hat{\Sigma}(\hat{\alpha}_n, \hat{\beta}_n) = A\hat{\Sigma}(\hat{\alpha}_n, \hat{\theta}_n)A^\top$, with

$$A = \begin{pmatrix} I_p & 0_{p\times 1} & 0_{p\times p} \\ I_p & 0_{p\times 1} & I_p \end{pmatrix},$$

where I_p is the $p \times p$ identity matrix. Following the above discussion and the argument in Zeng and Lin (2008), the estimator $\hat{\Sigma}(\hat{\alpha}_n, \hat{\beta}_n)$ is consistent under Conditions 1–5.

3.3 Hypothesis Testing of Submodels

The asymptotic results enable model selection for the nested submodels. For example, the Cox-type proportional rates assumption can be tested using $H_0 : \alpha = 0$ vs. $H_1 : \alpha \neq 0$ under the proposed model. In this case, a test statistic can be constructed as $T_{\text{cox}} = \hat{\alpha}_n^\top \hat{\Sigma}(\hat{\alpha}_n)^{-1} \hat{\alpha}_n$, where $\hat{\Sigma}(\hat{\alpha}_n)$ is the estimated covariance matrix of $n^{1/2}(\hat{\alpha}_n - \alpha)$. Under the null hypothesis, T_{cox} converges weakly to a Chi-square distribution χ^2_p, with p degrees of freedom. To evaluate the power of the test statistics, consider the true local alternative $\alpha = n^{-1/2}h$, where $h \in R^p$. Then, by the central limit theorem, T_{cox} converges weakly to a noncentral Chi-square distribution, with p

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

degrees of freedom and noncentrality parameter $h^\top \Sigma(\alpha)^{-1} h$. Therefore, the power of the test goes to one if $h^\top \Sigma(\alpha)^{-1} h \to \infty$ or $\|n^{1/2}\alpha\| \to \infty$.

The other two submodels can be tested and diagnosed in a similar way. For the accelerated mean model, we consider $H_0 : \gamma = 0$ vs. $H_1 : \gamma \neq 0$, with the test statistic $T_{\text{am}} = \hat{\gamma}_n^\top \hat{\Sigma}(\hat{\gamma}_n)^{-1} \hat{\gamma}_n$, where $\hat{\Sigma}(\hat{\gamma}_n)$ is the estimated covariance matrix of $n^{1/2}(\hat{\gamma}_n - \gamma)$. For the accelerated rate model, we consider $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$ and $T_{\text{ar}} = \hat{\beta}_n^\top \hat{\Sigma}(\hat{\beta}_n)^{-1} \hat{\beta}_n$, where $\hat{\Sigma}(\hat{\beta}_n)$ is the estimated covariance matrix of $n^{1/2}(\hat{\beta}_n - \beta)$. Following similar arguments, T_{am} and T_{ar} both converge weakly to the Chi-square distribution χ^2_p. In addition, the power of each test goes to one when the true parameters satisfy $\|n^{1/2}\gamma\| \to \infty$ and $\|n^{1/2}\beta\| \to \infty$, respectively.

4. Numerical Studies

Simulations were conducted to evaluate the performance of the proposed method. The recurrent event process is generated from a nonstationary Poisson process, with intensity function $\lambda(t) = Z\lambda_0(te^{\alpha_1 X_1 + \alpha_2 X_2})e^{\beta_1 X_1 + \beta_2 X_2}$, where $\lambda_0(t) = [2(1+t)]^{-1}$, and X_1 and X_2 are generated from independent standard normal distributions. The subject-specific latent variable Z is either set as $Z = 1$, or is generated from a gamma distribution with mean 1 and variance 0.25. The latter yields a scenario of informative censoring, whereas the former yields a scenario of noninformative censoring. In these settings, we use $E(Z) = 1$ rather than $\Lambda_0(\tau) = 1$, which we assumed in Section 2 for

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

ease of discussion. Because we only require one of these identifiability conditions, the proposed estimation procedure remains valid. We altered the regression coefficients α and β to generate data from either the proposed generalized scale-change model or the submodels discussed in Section 2. The censoring time is generated from an exponential distribution, with mean $60e^{-X_1}/Z$. We set $\tau = 60$.

For $n \in \{200, 400\}$ and 1000 replications, Tables 1 and 2 summarize the results for a generalized scale-change model and a Cox-type proportional rates model, respectively. The results for the accelerated rate model and accelerated mean model are presented in Appendix S2 of the Supplementary Material. Under our settings, the average number of observed recurrent events ranges from 1.5 to 5.7. The standard errors of the proposed method are obtained using the proposed resampling approach with 200 bootstrap samples. For comparison purposes, we also report the results of the estimator proposed by Sun and Su (2008), which requires a noninformative censoring assumption. Both the proposed estimator and the estimator proposed by Sun and Su (2008) were obtained by solving the corresponding estimating equations using the derivative-free Barzilai–Borwein spectral algorithm, implemented in Varadhan and Gilbert (2009). Zero vectors were used as the initial values for the equation solver. We also present results based on using the true values as the initial values, in order to investigate the stability of the estimating equations.

[Table 1 about here.]

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

[Table 2 about here.]

[Figure 1 about here.]

The proposed estimator is virtually unbiased for all scenarios considered, regardless of the choice of initial values. The average standard errors for the proposed estimators are reasonably close to their empirical counterparts, indicating that the proposed variance estimator performs well, even with a moderate bootstrap sample of size 200. Furthermore, the proposed estimator yields empirical coverage probabilities that are close to the nominal level of 95%, suggesting that the normal approximation for the distribution of the proposed estimators is appropriate. When $n = 200$, the empirical coverage probabilities for $\hat{\beta}_n$ are closer to the anticipated level of 95% than those are for $\hat{\alpha}_n$, suggesting that the normality approximation may require a larger sample size for $\hat{\alpha}_n$ than it does for $\hat{\beta}_n$. Similar trends are observed in the scenarios presented in Appendix S2 of the Supplementary Material. The estimates of the baseline cumulative rate functions for all scenarios are also presented in Appendix S2 of the Supplementary Material. The averages of $\hat{\Lambda}_0(t)$ are indistinguishable from the truth, for all cases considered.

The results of our simulation studies show that the estimator of Sun and Su (2008) is sensitive to the choice of initial values. As seen in Tables 1 and 2, the estimator of Sun and Su (2008) yields a large bias when the initial values are set to be zero. The average standard errors for the estimator of Sun and Su (2008) were obtained

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

using the classical bootstrap approach with 200 bootstrap samples. For most cases, the bootstrap standard errors are not close to their empirical counterparts. The inconsistency of the bootstrap standard errors reflects the instability of the estimator of Sun and Su (2008). As a result, the subsequent coverage probabilities are far from the nominal level. When the initial value was specified at the true value, their estimator yields small biases when $Z = 1$, but moderate biases when the noninformative censoring assumption is not met. The estimator of Sun and Su (2008) yields smaller empirical standard errors for all cases when the initial value was specified at the true value, which is not realistic in practice.

In Appendix S2 of the Supplementary Material, we report the results of additional simulation studies. Here, we considered a gamma frailty of variance 0.5 and 1, implying a larger degree of heterogeneity among subjects. For all settings considered, our estimator remains virtually unbiased, with estimated standard errors reasonably close to the empirical standard errors. The magnitude of the estimated standard errors seems to increase with the variance of the frailty variable, but the empirical coverage rates remain close to the nominal level in all scenarios. We also considered additional scenarios with a lower average number of events per subject than those in the earlier settings, as well as scenarios in which the recurrent events were generated from a nonPoisson process, given the frailty. In the later case, we include a second latent variable in the rate model and/or change the distribution of the inter-arrival time to

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

something other than the exponential distribution. The performance of our estimator remains satisfactory in all scenarios.

Now that we have shown that the proposed estimator is robust in most practical settings, we evaluate its performance when testing the nested submodels. We consider the same simulation settings, but with $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$. In addition, we focus only on the informative censoring scenarios. Based on 1000 replications, Figure 1 displays the rejection rate at a 0.05 significance level for the tests discussed in Section 3, fixing either α or β at $(0, 0)^\top$, and the other at $(k, k)^\top$, for some constant k. We set $k = 0$ initially, and then move k away from zero in both directions to denote a gradual deviation from the accelerated mean model, and a migration to the Cox-type model or accelerated rate model. When $\alpha = \beta = 0$, all rejection proportions are close to the nominal level of 0.05. In Figure 1a, the rejection proportions for the T_{cox} test are close to the nominal level of 0.05, reflected in the true value of $\alpha = 0$. As β deviates from zero in Figure 1a, both the T_{am} and T_{ar} tests increase in power, with slightly higher power for the T_{am} test. Similarly, in Figure 1b, the T_{cox} test and T_{am} test increase in power as α deviates from zero, whereas the T_{ar} test remains at the nominal level of 0.05 throughout. Among the tests, the T_{am} test appears to have the highest rejection proportion, indicating that our method is more likely to reject the accelerated mean model. For a given k, the rejection proportion for the T_{ar} test is higher than that for the T_{cox} test, which is because $\hat{\beta}_n$ is usually associated with smaller standard errors.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

5. Application

Serious infection is a major source of complications after a transplant, and is known to be associated with increased risk of allograft failure and death. A prospective cohort study was conducted at the Johns Hopkins Hospital to evaluate morbidity and mortality after transplant. In this study, patients who consented to an IRB-approved protocol were contacted every three months to obtain information on serious infection episodes. This preliminary cohort contained 161 kidney transplant recipients and 164 patients who had undergone hematopoietic stem cell transplant (HSCT) at the Johns Hopkins Hospital in 2012. Patients were followed until death, graft failure, or the end of the study. The median follow-up time was 20.2 months for the kidney transplant cohort, and 12.2 months for the HSCT cohort. During the study, the kidney transplant recipients experienced a total of 206 infection episodes (1.3 per recipient), and the HSCT recipients experienced a total of 290 infection episodes (1.8 per recipient). There were 42 deaths observed during the study period, among which 36 were in the HSCT cohort.

[Figure 2 about here.]

[Table 3 about here.]

We first analyze the infection process of the kidney transplant cohort. Of the 161 kidney transplant recipients, 91 (56.5%) were white, 47 (29.2%) had hyperten-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

sion, and 11 (6.8%) had diabetes at the time of the transplant. The age at transplant among the kidney transplant recipients ranged from 19.7 to 81.8 years, with a median of 53.5 years. Other potential risk factors used in the analysis include human leukocyte antigen (HLA) incompatibility and the high-risk cytomegalovirus (CMV) serostatus (CMV-negative recipients and CMV-positive donors vs. others). There were 31 (19.3%) HLA-incompatible patients and 21 (13.0%) patients with high-risk CMV serostatus. The age variable was centered and scaled to have unit variance. Figure 2a depicts the longitudinal patterns of recurrent infection episodes by HLA compatibility in the kidney transplant cohort. The plot suggests that HLA-incompatible transplant recipients tend to have a higher frequency of serious infections than HLA-compatible recipients do. The upper panel of Table 3 summarizes the estimated covariate effects for the kidney transplant cohort, with standard errors estimated using the proposed resampling approach, with 500 bootstraps. Using Wald's Chi-square test, the p-values used to test $H_0 : \alpha = 0$, $H_0 : \beta = 0$, and $H_0 : \gamma = 0$ are all less than 0.001. The hypothesis testing results suggest that none of the submodels are appropriate for the data, and that the covariates modify both the timescale of the infection process and the magnitude of the rate of infections. The estimated coefficients for age and HLA incompatibility are both significant and positive, implying that patients who were older or who were HLA incompatible were more likely to experience infections sooner, and more frequently throughout the follow-up period. In particular, for a one standard

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

deviation increase in age (12.8 years), their time-to-infection episodes accelerate by a factor of 0.36, and their risk increases to 1.75. Similarly, patients who underwent an HLA-incompatible kidney transplantation experienced time-to-infection episodes that were accelerated by a factor of 0.17, and an elevated risk of 3.95. Patients with a high-risk of the CMV disease or with hypertension tend to have time-to-infection episodes that were decelerated by a factor of 5.00 and 6.46, respectively.

A similar analysis was performed for the HSCT cohort, where, instead of HLA incompatibility and CMV serostatus, the type of stem cell transplant (allogeneic vs. autologous) was included as a covariate. Among the 164 HSCT patients, 126 (76.8%) were white, 93 (56.7%) were male, 128 (78.1%) had an allogeneic transplant, and 42 (25.6%) had lymphomas disease at the time of the transplant. The age at transplant ranged from 19.2 to 75.5 years, with a median of 52.2 years. We used the standardized age in this analysis. Figure 2b depicts the longitudinal patterns of the recurrent infection episodes and death, by type of HSCT transplantation. The figure shows that HSCT patients who underwent an allogeneic transplant tended to experience serious infections at a higher frequency. The lower panel of Table 3 summarizes the parameter estimates and their standard errors. The p-values used to test $H_0 : \beta = 0$ is 0.37, and the p-values used to test testing $H_0 : \alpha = 0$ and $H_0 : \gamma = 0$ are both less than 0.001. The hypothesis testing results suggest that the covariates are not significantly associated with the rate of infection, and the proposed model reduces to

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

the accelerated rate model. Because the proposed procedure estimates the timescale effect parameter without requiring an estimation of the multiplicative effect parameter, our inferences for the regression coefficients in the accelerated rate model are still valid. The only significant risk factor is the allogeneic transplant, which decelerated the time-to-infection episodes by a factor of 0.13.

Finally, we conduct a graphical assessment of whether the baseline rate function, $\lambda_0(t)$, is in the Weibull class. The assessment is motivated by the fact that, under the Weibull model, the proposed model reduces to the Cox-type model of Wang et al. (2001) and $\log\{\hat{\Lambda}_0(t)\}$ is linear in $\log(t)$. The plots of $\log\{\hat{\Lambda}_0(t)\}$ versus $\log(t)$ in Appendix S3 of the Supplementary Material suggest that $\lambda_0(t)$ is not Weibull.

6. Discussion

The proposed model addresses the need to characterize covariate effects in a flexible modeling framework and to account for informative censoring in recurrent event data analyses using a generalized scale-change model with an unspecified frailty. The estimation procedure is novel, and avoids requiring information about the frailties by exploiting the model structure. The asymptotic properties of the proposed estimator are established, and inferences are based on a computationally efficient resampling method. Because the model encompasses several popular models as special cases, the proposed approach includes model specification tests for the submodels that employ

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

various restrictions on the model parameters.

The proposed estimation procedure is based on a quasi-conditional likelihood, conditioning on (X_i, Z_i) and Y_i. Thus, our model is simple, in the sense that model specifications for the censoring event time and the frailty are not needed, because they can be treated as nuisances. For the same reason, our model is robust against the misspecification of the censoring time distribution, making it an appealing alternative to most joint modeling approaches that model the risk of recurrent events and the informative time jointly. The proposed method can be extended easily to a joint modeling framework when a joint analysis of the covariate effects on the recurrent events and the terminal event is of interest. For instance, in addition to assuming Model (2.1) for the underlying recurrent event process, we may consider the accelerated failure time model of Xu et al. (2017). Then we can specify the hazard function of the terminal event D as

$$h(t) = Zh_0(te^{X^\top\zeta})e^{X^\top\zeta}, \quad t \in [0, \tau], \tag{6.6}$$

where ζ is a $p\times 1$ vector of model parameters, and $h_0(t)$ is the baseline hazard function. Under the joint models, the recurrent event model can still be estimated by applying the estimation procedure described in Section 3. In addition, Model (6.6) can be estimated using the "borrow-strength" technique, originally proposed in Huang and Wang (2004), and later adopted in Xu et al. (2017). This is an interesting extension to pursue in future research.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

There are also several other possible research directions. The robustness of the proposed method comes at the cost of an efficiency loss. Thus, it would be of interest to evaluate this efficiency loss by comparing the performance of the proposed methods with that of the likelihood-based joint analyses of recurrent and terminal events, because the latter is expected to yield the most efficient parameter estimation under a correct model specification. In particular, the current estimation of α does not depend on the estimation of β, because in our carefully devised estimation procedure, β is not involved in the embedded seimparametric estimation in clustered right-truncated data. A more efficient estimator may be constructed by incorporating the knowledge about β in the estimation of α, thus, developing an iterative algorithm for estimating both β and α. Nonetheless, such a procedure is difficult to derive without additional assumptions, because $e^{\beta^\top X_i}$ and the unobserved frailty variable Z_i are coupled together in the rate function. It is expected that an additional assumption on the distribution of Z_i will be needed in order to exploit the information about β. For instance, one may use a likelihood-based estimation approach by assuming a distribution for the frailty variable. However, likelihood-based inferences using the proposed model have not been investigated in the literature, either for univariate survival data, or for recurrent event data, and thus warrants further research. Given that the current method can deal only with time-independent baseline covariates, it would also be of interest to extend the proposed method to allow both time-independent and time-dependent covariates

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

(Huang et al., 2010). From a model identifiability perspective, we recommend that the Weibull model should be fitted and diagnosed first. Then, if it is rejected, the proposed model can be fitted. Lastly, because graphical diagnoses are often subjective, a formal goodness-of-fit test for the Weibull model with frailty would be useful before applying the proposed model.

Supplementary Material

The online Supplementary Material contains the proof of Theorem 1, additional simulation results, and a graphical diagnosis of the Weibull model.

Acknowledgments

Xu and Chiou contributed equally to this work. The authors acknowledge the computing time granted on the Beowulf cluster of the Department of Statistics, University of Connecticut, partially supported by NSF grant DMS 0723557. The authors also thank Dr. Shao-Hsuan Wang for his computing support in an earlier exploration of this project. This research was partially supported by NSF SES1659328 and DMS 1712717 for Xu, and NIH R01CA193888 for Huang. The transplantation cohort study was supported, in part, by NIH K24AI085118.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

References

Barzilai, J. and Borwein, J. M. (1988). Two-point step size gradient methods. *IMA Journal of Numerical Analysis* **8**, 141–148.

Benjamin, V., Zhang, B., Nunamaker Jr, J. F., and Chen, H. (2016). Examining hacker participation length in cybercriminal internet-relay-chat communities. *Journal of Management Information Systems* **33**, 482–510.

Chen, Y. Q. and Jewell, N. P. (2001). On a general class of semiparametric hazards regression models. *Biometrika* **88**, 687–702.

Chen, Y. Q. and Wang, M.-C. (2000). Estimating a treatment effect with the accelerated hazards models. *Controlled Clinical Trials* **21**, 369–380.

Cook, R. J. and Lawless, J. (2007). *The Statistical Analysis of Recurrent Events*. New York: Wiley.

Cook, R. J. and Lawless, J. F. (1997). Marginal analysis of recurrent events and a terminating event. *Statistics in Medicine* **16**, 911–924.

Eaton, W. W., Mortensen, P. B., Herrman, H., Freeman, H., Bilker, W., Burgess, P., and Wooff, K. (1992). Long-term course of hospitalization for schizophrenia: Part I. risk for rehospitalization. *Schizophrenia Bulletin* **18**, 217–228.

Ghosh, D. (2004). Accelerated rates regression models for recurrent failure time data. *Lifetime Data Analysis* **10**, 247–261.

Ghosh, D. and Lin, D. (2002). Marginal regression models for recurrent and terminal events. *Statistica Sinica*

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

12, 663–688.

Huang, C.-Y., Qin, J., and Wang, M.-C. (2010). Semiparametric analysis for recurrent event data with time-dependent covariates and informative censoring. *Biometrics* **66**, 39–49.

Huang, C.-Y. and Wang, M.-C. (2004). Joint modeling and estimation for recurrent event processes and failure time data. *Journal of the American Statistical Association* **99**, 1153–1165.

Kalbfleisch, J. and Lawless, J. (1991). Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Statistica Sinica* **1**, 19–32.

Kalbfleisch, J. D., Schaubel, D. E., Ye, Y., and Gong, Q. (2013). An estimating function approach to the analysis of recurrent and terminal events. *Biometrics* **69**, 366–374.

Lai, T. L. and Ying, Z. (1991). Rank regression methods for left-truncated and right-censored data. *The Annals of Statistics* **19**, 531–556.

Lancaster, T. and Intrator, O. (1998). Panel data with survival: Hospitalization of HIV-positive patients. *Journal of the American Statistical Association* **93**, 46–53.

Lawless, J. F. and Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics* **37**, 158–168.

Lin, D., Wei, L., Yang, I., and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B* **62**, 711–730.

Lin, D. Y., Wei, L. J., and Ying, Z. (1998). Accelerated failure time models for counting processes. *Biometrika* **85**, 605–618.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

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

Liu, Y., Wu, Y., Cai, J., and Zhou, H. (2010). Additive–multiplicative rates model for recurrent events. *Lifetime Data Analysis* **16**, 353–373.

Luo, X., Wang, M.-C., and Huang, C.-Y. (2010). A comparison of various rate functions of a recurrent event process in the presence of a terminal event. *Statistical Methods in Medical Research* **19**, 167–182.

Marr, K. A. (2012). Delayed opportunistic infections in hematopoietic stem cell transplantation patients: A surmountable challenge. *ASH Education Program Book* **2012**, 265–270.

Nelson, W. B. (2003). *Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications*. SIAM.

Pepe, M. S. and Cai, J. (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association* **88**, 811–820.

Rogers, J. K., McMurray, J. J., Pocock, S. J., Zannad, F., Krum, H., van Veldhuisen, D. J., Swedberg, K., Shi, H., Vincent, J., and Pitt, B. (2012). Eplerenone in patients with systolic heart failure and mild symptoms analysis of repeat hospitalizations. *Circulation* **126**, 2317–2323.

Schaubel, D. E., Zeng, D., and Cai, J. (2006). A semiparametric additive rates model for recurrent event data. *Lifetime Data Analysis* **12**, 389–406.

Sun, L. and Su, B. (2008). A class of accelerated means regression models for recurrent event data. *Lifetime Data Analysis* **14**, 357–375.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

Varadhan, R. and Gilbert, P. (2009). BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *Journal of Statistical Software* **32**, 1–26.

Wang, M.-C. (1989). A semiparametric model for randomly truncated data. *Journal of the American Statistical Association* **84**, 742–748.

Wang, M.-C., Qin, J., and Chiang, C.-T. (2001). Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association* **96**, 1057–1065.

Wei, L. and Glidden, D. V. (1997). An overview of statistical methods for multiple failure time data in clinical trials. *Statistics in Medicine* **16**, 833–839.

Xu, G., Chiou, S. H., Huang, C.-Y., Wang, M.-C., and Yan, J. (2017). Joint scale-change models for recurrent events and failure time. *Journal of the American Statistical Association* **112**, 794–805.

Ye, Y., Kalbfleisch, J. D., and Schaubel, D. E. (2007). Semiparametric analysis of correlated recurrent and terminal events. *Biometrics* **63**, 78–87.

Zeng, D. and Lin, D. (2009). Semiparametric transformation models with random effects for joint analysis of recurrent and terminal events. *Biometrics* **65**, 746–752.

Zeng, D. and Lin, D. Y. (2008). Efficient resampling methods for nonsmooth estimating functions. *Biostatistics* **9**, 355–363.

Department of Statistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

E-mail: gongjun@umich.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75080, U.S.A.

E-mail: schiou@utdallas.edu

Department of Statistics, University of Connecticut, Storrs, CT 06269, U.S.A.

E-mail: jun.yan@uconn.edu

Johns Hopkins University School of Medicine, Baltimore, MD 21205, U.S.A.

E-mail: kmarr4@jhmi.edu

Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94158, U.S.A.

E-mail: ChiungYu.Huang@ucsf.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

Table 1: Simulation results with $\alpha = (-1, -1)^\top$ and $\beta = (1, 1)^\top$. Columns without an asterisk $(*)$ present results using the zero vector as the initial value; columns with an asterisk present results using the true value as the initial value; Bias is the empirical bias; ESE is the empirical standard error; ASE is the average of the standard error obtained from resampling; CP is the empirical coverage probability (%) of the 95% confidence intervals.

		Proposed						Sun and Su (2008)					
n		Bias	ESE	ASE	CP	Bias*	ESE*	Bias	ESE	ASE	CP	Bias*	ESE*
						$Z = 1$							
200	α_1	0.005	0.307	0.295	93.0	−0.011	0.308	0.678	0.533	0.218	19.2	0.003	0.095
	α_2	0.007	0.277	0.264	93.1	−0.013	0.273	0.646	0.499	0.214	19.9	0.003	0.091
	β_1	0.005	0.181	0.184	93.7	0.004	0.179	0.373	0.332	0.142	33.5	−0.010	0.091
	β_2	0.005	0.175	0.174	94.2	−0.002	0.174	0.375	0.323	0.140	34.7	0.007	0.094
400	α_1	0.019	0.218	0.210	94.6	−0.004	0.217	0.218	0.251	0.161	39.5	0.001	0.086
	α_2	0.011	0.194	0.186	94.2	0.003	0.189	0.194	0.239	0.155	41.4	−0.008	0.087
	β_1	0.006	0.135	0.131	95.1	0.005	0.131	0.123	0.172	0.109	57.1	−0.001	0.077
	β_2	0.008	0.129	0.128	95.0	0.000	0.130	0.113	0.165	0.105	57.4	−0.002	0.077
						$Z \sim \text{Gamma}(4, 4)$							
200	α_1	0.001	0.314	0.301	93.7	0.001	0.320	0.683	0.410	0.224	18.8	−0.113	0.159
	α_2	−0.010	0.281	0.267	93.3	0.023	0.281	0.536	0.386	0.230	25.1	0.095	0.146
	β_1	−0.008	0.221	0.219	93.7	−0.006	0.223	0.311	0.305	0.173	44.1	0.103	0.169
	β_2	0.003	0.213	0.209	93.5	0.013	0.214	0.311	0.301	0.180	49.6	0.096	0.168
400	α_1	0.005	0.232	0.215	94.5	−0.008	0.224	0.432	0.257	0.162	23.9	−0.147	0.156
	α_2	0.002	0.203	0.191	94.4	−0.009	0.205	0.264	0.218	0.158	41.7	−0.091	0.144
	β_1	0.005	0.163	0.158	94.4	−0.006	0.168	0.169	0.190	0.136	62.6	−0.090	0.136
	β_2	−0.001	0.151	0.151	94.7	−0.000	0.147	0.134	0.188	0.138	70.4	0.097	0.141

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

Table 2: Simulation results with $\alpha = (0, 0)^\top$ and $\beta = (-1, -1)^\top$. Columns without an asterisk $(*)$ present results using the zero vector as the initial value; columns with an asterisk present results using the true value as the initial value; Bias is the empirical bias; ESE is the empirical standard error; ASE is the average of the standard error obtained from resampling; CP is the empirical coverage probability (%) of the 95% confidence intervals.

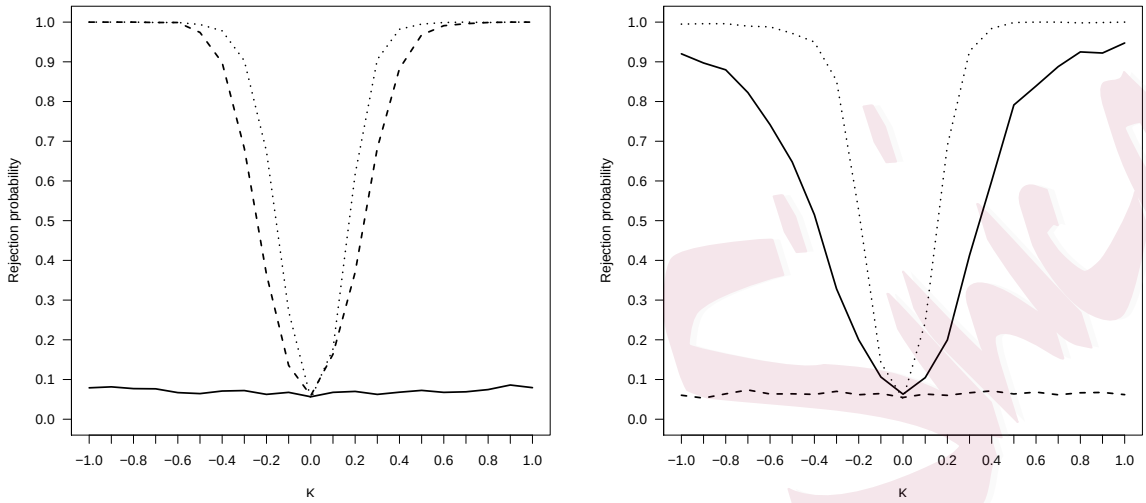
		Proposed						Sun and Su (2008)					
n		Bias	ESE	ASE	CP	Bias*	ESE*	Bias	ESE	ASE	CP	Bias*	ESE*
						$Z = 1$							
200	α_1	−0.004	0.161	0.155	92.5	−0.006	0.148	0.358	0.240	0.092	15.2	−0.019	0.049
	α_2	−0.001	0.153	0.151	92.5	0.005	0.150	0.394	0.237	0.097	11.6	−0.018	0.043
	β_1	−0.001	0.118	0.114	92.9	−0.003	0.109	0.241	0.154	0.081	20.3	−0.013	0.069
	β_2	−0.003	0.112	0.113	92.5	0.003	0.112	0.262	0.158	0.082	18.2	−0.012	0.072
400	α_1	0.001	0.104	0.100	94.9	0.001	0.110	0.187	0.215	0.078	33.3	−0.016	0.039
	α_2	−0.005	0.102	0.096	93.6	−0.001	0.102	0.199	0.214	0.077	30.4	0.008	0.042
	β_1	−0.001	0.078	0.073	94.8	0.001	0.081	0.127	0.145	0.064	39.8	−0.010	0.062
	β_2	−0.003	0.076	0.071	94.6	−0.001	0.076	0.135	0.148	0.063	39.4	0.005	0.058
						$Z \sim \text{Gamma}(4, 4)$							
200	α_1	0.007	0.167	0.146	92.4	−0.011	0.169	0.414	0.219	0.118	15.0	−0.042	0.062
	α_2	0.002	0.155	0.141	94.0	−0.001	0.155	0.388	0.253	0.124	18.2	−0.045	0.077
	β_1	0.007	0.149	0.124	91.7	−0.000	0.142	0.273	0.168	0.127	35.9	−0.044	0.129
	β_2	0.003	0.146	0.123	90.7	0.005	0.143	0.255	0.202	0.131	40.5	−0.043	0.121
400	α_1	0.005	0.113	0.101	92.2	0.001	0.112	0.302	0.208	0.093	20.3	−0.051	0.071
	α_2	0.001	0.108	0.099	93.6	0.001	0.105	0.249	0.231	0.100	32.0	−0.049	0.074
	β_1	0.000	0.106	0.095	91.9	0.005	0.102	0.189	0.168	0.100	42.7	−0.044	0.096
	β_2	−0.005	0.102	0.099	91.8	0.006	0.100	0.174	0.186	0.102	49.4	−0.055	0.102

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

Table 3: Summary of the infection data; $\hat{\alpha}$ and $\hat{\beta}$ are point estimators; SE($\hat{\alpha}$) and SE($\hat{\beta}$) are the corresponding standard errors; the age variable is standardized to have mean zero and standard deviation one.

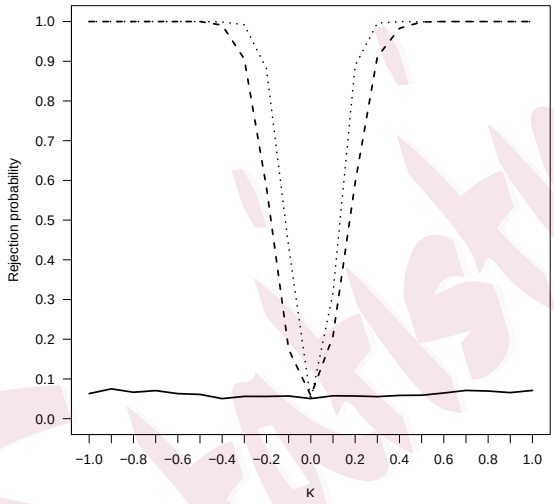
	Proposed Model			
	$\hat{\alpha}$	SE($\hat{\alpha}$)	$\hat{\beta}$	SE($\hat{\beta}$)
Kidney transplant cohort				
Age	1.025	0.354	0.557	0.263
White	−1.729	0.952	−0.631	0.600
HLA incompatible	1.757	0.651	1.374	0.478
CMV	−1.609	0.654	−0.087	0.487
Diabetes	1.076	1.383	0.019	0.821
Hypertension	−1.864	0.918	−0.917	0.729
HSCT cohort				
Age	−0.320	0.567	0.075	0.149
White	−0.966	0.871	−0.281	0.504
Male	−2.237	2.781	−0.864	0.773
Allogeneic	−2.038	0.804	0.517	0.907
Lymphomas disease	−1.048	1.347	−0.478	0.565

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224

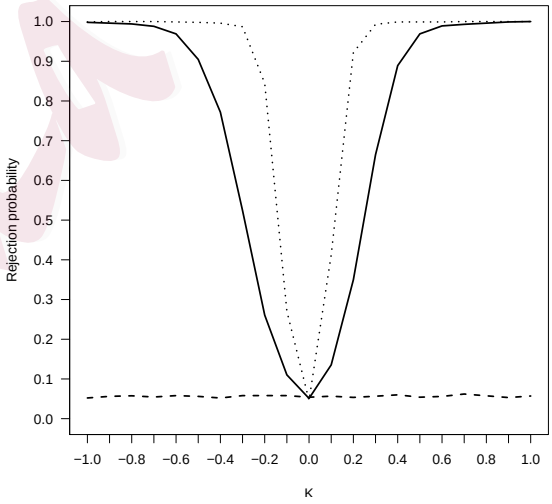


(a) $n = 200, \alpha = (0,0)^\top, \beta = (k,k)^\top$.

(b) $n = 200, \alpha = (k,k)^\top, \beta = (0,0)^\top$.



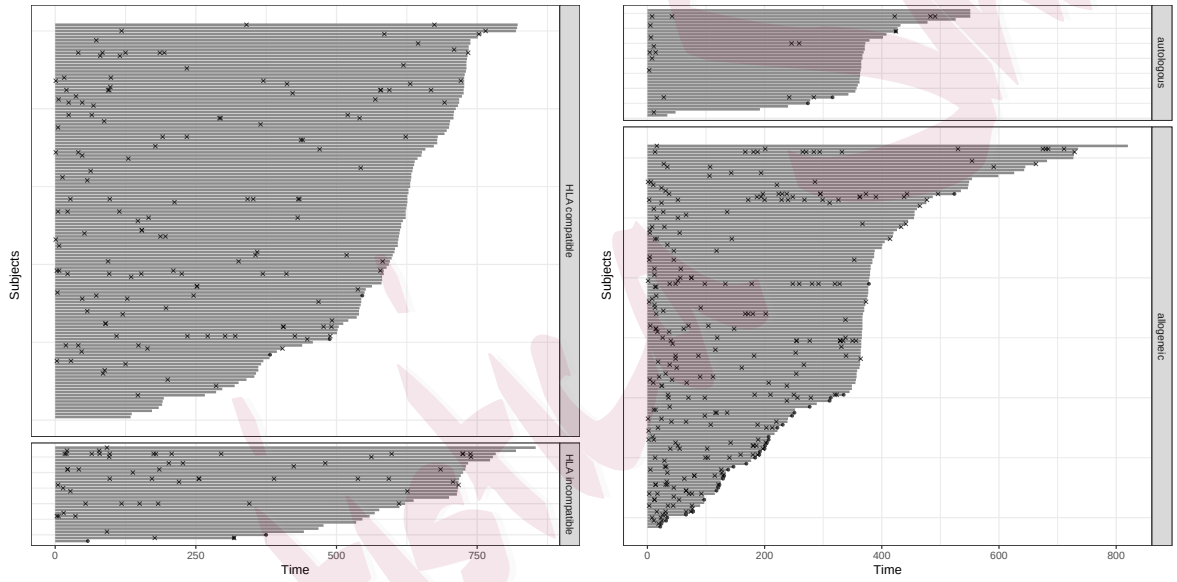
(c) $n = 400, \alpha = (0,0)^\top, \beta = (k,k)^\top$.



(d) $n = 400, \alpha = (k,k)^\top, \beta = (0,0)^\top$.

Figure 1: Rejection rates based on 1000 replications at the 0.05 significance level using the hypothesis testing procedures described in Section 3. Solid lines (—) present the rejection rates for $H_0 : \alpha = 0$, which is used to test the Cox-type proportional rates assumption; Dashed lines (- - -) present the rejection rates for $H_0 : \beta = 0$, which is used to test the accelerated rate assumption; Dotted lines (·····) present the rejection rates for $H_0 : \alpha = \beta$ or $\gamma = 0$, which is used to test the accelerated mean assumption.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0224



(a) Kidney transplant c ohort by HLA compatibil- (b) HSCT cohort by type of stem cell transplant.
ity.

Figure 2: Longitudinal plots of the infection data; horizontal gray lines indicate the time elapsed from transplant to end of follow-up; × represents an infection episode; • represents death.