Computational Statistics and Data Analysis 89 (2015) 39–50

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A semiparametric additive rates model for multivariate recurrent events with missing event categories

Peng Ye [a], Xingqiu Zhao [b,c,*], Liuquan Sun [a], Wei Xu [d]

[a] Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China
[b] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong
[c] Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China
[d] Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

ARTICLE INFO

Article history:
Received 14 June 2014
Received in revised form 22 November 2014
Accepted 3 March 2015
Available online 15 March 2015

Keywords:
Additive rates model
Marginal models
Missing at random
Multivariate recurrent events
Weighted estimating equation

ABSTRACT

Multivariate recurrent event data arise in many clinical and observational studies, in which subjects may experience multiple types of recurrent events. In some applications, event times can be always observed, but types for some events may be missing. In this article, a semiparametric additive rates model is proposed for analyzing multivariate recurrent event data when event categories are missing at random. A weighted estimating equation approach is developed to estimate parameters of interest, and the resulting estimators are shown to be consistent and asymptotically normal. In addition, a lack-of-fit test is presented to assess the adequacy of the model. Simulation studies demonstrate that the proposed method performs well for practical settings. An application to a platelet transfusion reaction study is provided.

1. Introduction

Multivariate recurrent event data are often encountered in biomedical studies when subjects may experience several different types of recurrent events (Cai and Schaubel, 2004; Chen et al., 2005; Schaubel and Cai, 2006a). For example, infections in bone marrow transplantation can be subtyped by underlying causes (e.g. bacterial, fungal and viral infections). Childhood asthma outcomes may be differentiated by severity (e.g. hospital admissions and physician office visits). The occurrence of a technical failure in continuous ambulatory peritoneal dialysis study may be classified by causes (e.g. peritonitis, abdominal complications, inadequate dialysis and other). For analyzing this kind of data, it is usually more informative to study the category-specific recurrent event processes separately, rather than aggregating across event categories, because the type-specific covariate effects may be not equal in many situations.

Several methods have been proposed in the literature to analyze multivariate recurrent event data (Abu-Libdeh et al., 1990; Cai and Schaubel, 2004; Sun et al., 2009; Zhu et al., 2010; Chen et al., 2012; Zhao et al., 2012). For example, Abu-Libdeh et al. (1990) suggested a nonhomogeneous mixed Poisson process to model the dependence among different types of recurrent events. Cai and Schaubel (2004) proposed a class of proportional marginal means and rates models for assessing the effect of covariates on the event processes. Sun et al. (2009) presented a semiparametric multiplicative rates model with time-varying covariate effects. Zhu et al. (2010) considered a joint modeling approach for regression analysis of multivariate recurrent event data in the presence of a dependent terminal event. Chen et al. (2012) proposed a general additive marginal

* Corresponding author at: Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong.
E-mail address: xingqiu.zhao@polyu.edu.hk (X. Zhao).

http://dx.doi.org/10.1016/j.csda.2015.03.002
0167-9473/

rate model for the multiple type recurrent events. Zhao et al. (2012) studied a joint semiparametric frailty-based proportional intensity model with both time-dependent and time-independent covariates.

The aforementioned semiparametric regression models assumed that the event category was always known. In reality, however, the occurrence of an event is always observed but the specific type is unknown. For example, in the Dialysis Outcomes and Practice Patterns Study (Schaubel and Cai, 2006b), the dates of hospital admission of all patients are available, but for some patients, the reasons for hospital admission may not be recorded or may be difficult to determine. Another example is a clinical trial involving different treatment strategies for asthma patients (Chen and Cook, 2009), where asthma exacerbations are subtyped by the analysis of sputum cell counts as either non-eosinophilic or eosinophilic. If patients took rescue medication for exacerbations before providing a sputum sample, the cellular analysis of sputum was not valid, and thus the exacerbation type was not determined. When the event category may be missing, a naive way of handling such a situation is the complete-case analysis, which treats events of unknown category as censored. The complete-case analysis leads to biased estimators unless event categories are missing completely at random (Little and Rubin, 2002), which assume that category missingness occurs randomly among events.

When the category missingness mechanism depends on the observed data but not on the missing categories, it is termed missing at random (MAR) (Little and Rubin, 2002). In recent years, some methods have been developed to analyze multivariate recurrent event data with missing event categories under the MAR assumption (Schaubel and Cai, 2006a,b); (Chen and Cook, 2009; Lin et al., 2013). For example, Schaubel and Cai (2006a,b) studied the multiple-event-category proportional means and rates model using weighted estimating equations and multiple imputation methods, respectively. Chen and Cook (2009) considered the multivariate random effects model based on a likelihood approach. Lin et al. (2013) proposed a nonparametric estimation of the mean function in which the missingness mechanism is completely unspecified. Note that a useful and important alternative to the proportional rates model is the additive rates model. In this article, we propose a semiparametric additive rates model to analyze multivariate recurrent event data with missing event categories under the MAR assumption. A weighted estimating equation approach is developed to estimate parameters of interest. The resulting estimators have closed forms and are easy to implement.

The rest of the article is organized as follows. In Section 2, we specify the model and propose an estimating equation approach for estimation of model parameters. Section 3 presents the asymptotic properties of the resulting estimators with proofs outlined in the Appendix. A model checking technique is given in Section 4, and some simulation results to evaluate the proposed method are reported in Section 5. An application to a platelet transfusion reaction study is provided in Section 6, and some concluding remarks are made in Section 7.

2. Model and estimation procedures

Suppose that there are n independent subjects with K recurrent event categories. Let $N^*_{ik}(t)$ denote the number of category k events over the interval $[0, t]$ for subject i, and C_{ik} be the right censoring time for event type k for subject i. Usually, $C_{ik} = C_i$ for $k = 1, \ldots, K$. Let $Y_{ik}(t) = I(C_{ik} \geq t)$ be the at-risk process, where $I(\cdot)$ is the indicator function. The observed event processes are given by $N_{ik}(t) = \int_0^t Y_{ik}(s)\, dN^*_{ik}(s)$. Assume that $dN^*_{ik}(t) \in \{0, 1\}$ and that $dN^*_{ik}(t)dN^*_{il}(t) = 0$ for $k \neq l$. Let $Z_{ik}(t)$ be the $p \times 1$ vector of external time-dependent covariates (Kalbfleisch and Prentice, 2002). The proposed semiparametric additive rates model takes the form

$$E[dN^*_{ik}(t)|Z_{ik}(t)] = d\mu_{0k}(t) + \beta_0' Z_{ik}(t)dt, \tag{1}$$

for $k = 1, \ldots, K$, where β_0 is a vector of unknown regression parameters, and $\mu_{0k}(t)$ is an unspecified baseline mean function. In the case where all data are observed, model (1) has been studied by Chen et al. (2012). In addition, when $K \equiv 1$, model (1) reduces to that considered by Schaubel et al. (2006).

Remark 1. Although model (1) is each written in terms of a regression parameter vector which is common across event categories, category-specific parameter vector can be incorporated upon appropriate expansion of the covariate vector. In addition, for the case of common baseline mean function across event categories, we have

$$E[dN^*_{ik}(t)|Z_{ik}(t)] = d\mu_0(t) + \beta_0' Z_{ik}(t)dt. \tag{2}$$

The proposed estimation procedure can be extended in a straightforward manner to deal with model (2).

We consider the setting where event times are always observed, but event categories may be missing under the MAR assumption. Let $\delta_i(t)$ denote the type of the event which occurred to subject i at time t, and set $\delta_{ik}(t) = I(\delta_i(t) = k)$. Define $\xi_i(t) = 1$ when an event occurs at time t and $\delta_i(t)$ is known, and 0 otherwise. For a random sample of n subjects, the observed data consist of $\{N_{ik}(t), C_{ik}, Z_{ik}(t), \xi_i(t), \xi_i(t)\delta_i(t); t \leq C_{ik}, i = 1, \ldots, n, k = 1, \ldots, K\}$. When some of the event categories are missing, a complete case analysis may not only lose efficiency due to discarding all events with missing categories, but may also yield biased estimators when the event categories are MAR.

Define $dN_{i\cdot}(t) = \sum_{k=1}^K dN_{ik}(t)$. Since $dN_{ik}(t)dN_{il}(t) = 0$ for $k \neq l$, it follows that $dN_{ik}(t) = \delta_{ik}(t)dN_{i\cdot}(t)$, and

$$dN_{ik}(t) = \xi_i(t)dN_{ik}(t) + \delta_{ik}(t)dN^c_{i\cdot}(t),$$

where $dN_{i.}^c(t) = (1 - \xi_i(t))dN_{i.}(t)$. Thus, under model (1), we have

$$E\Big[\xi_i(t)dN_{ik}(t) + \delta_{ik}(t)dN_{i.}^c(t) - Y_{ik}(t)\{d\mu_{0k}(t) + \beta_0' Z_{ik}(t)dt\}\Big] = 0.$$

However, it does not lead to a feasible estimating equation since we are unable to observe $\delta_{ik}(t)$ when $dN_{i.}^c(t) = 1$. Let $W_i(t)$ be a vector denoting the pertinent information in the event history at time t for subject i. Then under the MAR assumption,

$$\begin{aligned} E[\delta_{ik}(t)|dN_{i.}^c(t) = 1, W_i(t)] &= E[\delta_{ik}(t)|dN_{i.}(t) = 1, \xi_i(t) = 0, W_i(t)] \\ &= E[\delta_{ik}(t)|dN_{i.}(t) = 1, W_i(t)] \\ &= E[\delta_{ik}(t)|dN_{i.}(t) = 1, \xi_i(t) = 1, W_i(t)], \end{aligned}$$

which implies that $E[\delta_{ik}(t)|dN_{i.}^c(t) = 1, W_i(t)]$ can be estimated based on events where categories are not missing. Let $\pi_{ik}(t) = E[\delta_{ik}(t)|dN_{i.}(t) = 1, W_i(t)]$. Here we assume that $\pi_{ik}(t)$ can be parametrically modeled as $\pi_{ik}(t; \gamma_0)$, where γ_0 is the true parameter value. As discussed in Schaubel and Cai (2006a,b), we propose to model $\pi_{ik}(t; \gamma_0)$ through the following generalized logits model:

$$\log\left\{\frac{\pi_{ik}(t; \gamma_0)}{\pi_{i1}(t; \gamma_0)}\right\} = \gamma_0' W_{ik}(t), \quad k = 2, \ldots, K, \tag{3}$$

where the vector $W_{ik}(t)$ contains the elements of $W_i(t)$ which pertain to category k, and $k = 1$ is arbitrarily selected as the reference category. Note that the generalized logits model is a very flexible approach, and other parametric models can also be easily accommodated. Let $\hat{\gamma}$ be the solution to the following estimating equation:

$$\sum_{i=1}^{n}\sum_{k=2}^{K}\int_0^\tau W_{ik}(t)[\delta_{ik}(t) - \pi_{ik}(t; \gamma)]\xi_i(t)dN_{i.}(t) = 0. \tag{4}$$

Then $\hat{\gamma}$ is a consistent estimator of γ_0. Also the event category probabilities can be estimated by

$$\pi_{ik}(t; \hat{\gamma}) = \frac{\exp\{\hat{\gamma}' W_{ik}(t)\}}{\sum_{l=1}^{K}\exp\{\hat{\gamma}' W_{il}(t)\}}, \quad k = 1, \ldots, K,$$

where $W_{i1}(t) = 0$. Define

$$dM_{ik}(t; \beta, \gamma) = \xi_i(t)dN_{ik}(t) + \pi_{ik}(t; \gamma)dN_{i.}^c(t) - Y_{ik}(t)\{d\mu_{0k}(t) + \beta' Z_{ik}(t)dt\}.$$

Under models (1) and (3), we have $E[M_{ik}(t; \beta_0, \gamma_0)] = 0$. By applying the generalized estimating equation approach (Liang and Zeger, 1986) and the consistency of $\hat{\gamma}$ for γ_0, we specify the following estimating equations for $\mu_{0k}(t)$ and β_0:

$$\sum_{i=1}^{n}\int_0^t dM_{ik}(s; \beta, \hat{\gamma}) = 0, \quad 0 \le t \le \tau, \tag{5}$$

and

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau Z_{ik}(t)dM_{ik}(t; \beta, \hat{\gamma}) = 0, \tag{6}$$

where τ is a prespecified constant such that $P(Y_{ik}(\tau) = 1) > 0$ for $k = 1, \ldots, K$ and $i = 1, \ldots, n$. For given β, it follows from (5) that

$$\hat{\mu}_{0k}(t; \beta, \hat{\gamma}) = \sum_{i=1}^{n}\int_0^t \frac{\xi_i(s)dN_{ik}(s) + \pi_{ik}(s; \hat{\gamma})dN_{i.}^c(s) - Y_{ik}(s)\beta' Z_{ik}(s)ds}{\sum_{j=1}^{n} Y_{jk}(s)}. \tag{7}$$

Substituting (7) into (6), we obtain the following weighted estimating function for β_0:

$$U(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau \{Z_{ik}(t) - \bar{Z}_k(t)\}\{\xi_i(t)dN_{ik}(t) + \pi_{ik}(t; \hat{\gamma})dN_{i.}^c(t) - Y_{ik}(t)\beta' Z_{ik}(t)dt\}, \tag{8}$$

where $\bar{Z}_k(t) = \sum_{i=1}^n Y_{ik}(t)Z_{ik}(t)/\sum_{i=1}^n Y_{ik}(t)$. Let $\hat{\beta}$ be the solution to $U(\beta) = 0$, which has a closed form

$$\hat{\beta} = \left[\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau Y_{ik}(t)\{Z_{ik}(t) - \bar{Z}_k(t)\}^{\otimes 2}dt\right]^{-1}\left[\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau \{Z_{ik}(t) - \bar{Z}_k(t)\}\Big\{\xi_i(t)dN_{ik}(t) + \pi_{ik}(t; \hat{\gamma})dN_{i.}^c(t)\Big\}\right],$$

where $a^{\otimes 2} = aa'$ for any vector a. The corresponding estimator of $\mu_{0k}(t)$ is then given by $\hat{\mu}_{0k}(t) = \hat{\mu}_{0k}(t; \hat{\beta}, \hat{\gamma})$.

3. Asymptotic properties

In this section, we establish the asymptotic properties of the proposed estimators, and assume that the following regularity conditions hold for $i = 1, \ldots, n$ and $k = 1, \ldots, K$.

(C1) $\{N_{ik}(\cdot), C_{ik}, Z_{ik}(\cdot)\}_{k=1}^{K}$ are independent and identically distributed for $i = 1, \ldots, n$.
(C2) $P(C_{ik} \geq \tau) > 0$, and $N_{ik}(\tau) < \eta < \infty$ almost surely.
(C3) $Z_{ik}(t)$ and $W_{ik}(t)$ are almost surely of bounded variation on $[0, \tau]$.
(C4) A and $\Omega(\gamma_0)$ are nonsingular, where

$$A = E\left[\sum_{k=1}^{K}\int_0^{\tau} Y_{ik}(t)\{Z_{ik}(t) - \bar{z}_k(t)\}^{\otimes 2}dt\right],$$

$$\Omega(\gamma) = E\left[\sum_{k=1}^{K}\int_0^{\tau} W_{ik}(t)\pi_{ik}(t;\gamma)\left\{W_{ik}(t) - \sum_{l=1}^{K} W_{il}(t)\pi_{il}(t;\gamma)\right\}'\xi_i(t)dN_{i\cdot}(t)\right],$$

and $\bar{z}_k(t)$ is the limit of $\bar{Z}_k(t)$.

The asymptotic properties of $\hat{\beta}$ are summarized in the following theorem with the proof outlined in the Appendix.

Theorem 1. *Under the regularity conditions* (C1)–(C4), *$\hat{\beta}$ is strongly consistent to β_0, and $n^{1/2}(\hat{\beta} - \beta_0)$ is asymptotically normal with mean zero and covariance matrix $A^{-1}\Sigma A^{-1}$, where $\Sigma = E[(\sum_{k=1}^{K}\Phi_{ik}(\beta_0, \gamma_0))^{\otimes 2}]$,*

$$\Phi_{ik}(\beta,\gamma) = \int_0^{\tau}\{Z_{ik}(t) - \bar{z}_k(t)\}dM_{ik}(t;\beta,\gamma) + \Psi_k(\gamma)\Omega(\gamma)^{-1}\sum_{l=1}^{K}\Gamma_{il}(\gamma),$$

$$\Gamma_{ik}(\gamma) = \int_0^{\tau} W_{ik}(t)\{\delta_{ik}(t) - \pi_{ik}(t;\gamma)\}\xi_i(t)dN_{i\cdot}(t),$$

$$\hat{\Psi}_k(\gamma) = n^{-1}\sum_{i=1}^{n}\int_0^{\tau}\{Z_{ik}(t) - \bar{Z}_k(t)\}\left\{W_{ik}(t) - \sum_{l=1}^{K} W_{il}(t)\pi_{il}(t;\gamma)\right\}'\pi_{ik}(t;\gamma)dN_{i\cdot}^{c}(t),$$

and $\Psi_k(\gamma)$ is the limit of $\hat{\Psi}_k(\gamma)$.

The asymptotic covariance matrix $A^{-1}\Sigma A^{-1}$ can be consistently estimated by $\hat{A}^{-1}\hat{\Sigma}\hat{A}^{-1}$, where $\hat{\Sigma} = n^{-1}\sum_{i=1}^{n}[\sum_{k=1}^{K}\hat{\Phi}_{ik}(\hat{\beta},\hat{\gamma})]^{\otimes 2}$,

$$\hat{A} = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^{\tau} Y_{ik}(t)\{Z_{ik}(t) - \bar{Z}_k(t)\}^{\otimes 2}dt,$$

$$\hat{\Phi}_{ik}(\beta,\gamma) = \int_0^{\tau}\{Z_{ik}(t) - \bar{Z}_k(t)\}d\hat{M}_{ik}(t;\beta,\gamma) + \hat{\Psi}_k(\gamma)\hat{\Omega}(\gamma)^{-1}\sum_{l=1}^{K}\Gamma_{il}(\gamma),$$

$$\hat{\Omega}(\gamma) = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^{\tau} W_{ik}(t)\pi_{ik}(t;\gamma)\left\{W_{ik}(t) - \sum_{l=1}^{K} W_{il}(t)\pi_{il}(t;\gamma)\right\}'\xi_i(t)dN_{i\cdot}(t),$$

and

$$d\hat{M}_{ik}(t;\beta,\gamma) = \xi_i(t)dN_{ik}(t) + \pi_{ik}(t;\gamma)dN_{i\cdot}^{c}(t) - Y_{ik}(t)\{d\hat{\mu}_{0k}(t;\beta,\gamma) + \beta' Z_{ik}(t)dt\}.$$

Define $\hat{H}_k(t) = -\int_0^t \bar{Z}_k(s)ds$ and $\bar{Y}_k(t) = n^{-1}\sum_{i=1}^{n} Y_{ik}(t)$. Let $H_k(t)$ and $\bar{y}_k(t)$ be the limits of $\hat{H}_k(t)$ and $\bar{Y}_k(t)$, respectively. The asymptotic properties of $\hat{\mu}_{0k}(t)$ are given in the next theorem.

Theorem 2. *$\hat{\mu}_{0k}(t)$ converges almost surely to $\mu_{0k}(t)$ uniformly in $t \in [0, \tau]$, and $n^{1/2}(\hat{\mu}_{0k}(t) - \mu_{0k}(t))$ converges weakly on $[0, \tau]$ to a zero-mean Gaussian process with covariance function at (s, t) equal to $\omega_k(s,t) = E[\phi_{ik}(s;\beta_0,\gamma_0)\phi_{ik}(t;\beta_0,\gamma_0)]$, where*

$$\phi_{ik}(t;\beta,\gamma) = H_k(t)'A^{-1}\sum_{k=1}^{K}\Phi_{ik}(\beta,\gamma) + Q_k(t;\gamma)\Omega(\gamma)^{-1}\sum_{l=1}^{K}\Gamma_{il}(\gamma) + \int_0^t \bar{y}_k(s)^{-1}dM_{ik}(s;\beta,\gamma),$$

$$\hat{Q}_k(t;\gamma) = n^{-1}\sum_{i=1}^{n}\int_0^t \bar{Y}_k(s)^{-1}\pi_{ik}(s;\gamma)\left\{W_{ik}(s) - \sum_{l=1}^{K} W_{il}(s)\pi_{il}(s;\gamma)\right\}'dN_{i\cdot}^{c}(s),$$

and $Q_k(t;\gamma)$ is the limit of $\hat{Q}_k(t;\gamma)$.

The covariance function $\omega_k(s,t)$ can be consistently estimated by

$$\hat{\omega}_k(s,t) = n^{-1}\sum_{i=1}^{n}\hat{\phi}_{ik}(s)\hat{\phi}_{ik}(t),$$

where

$$\hat{\phi}_{ik}(t) = \hat{H}_k(t)'\hat{A}^{-1}\sum_{k=1}^{K}\hat{\Phi}_{ik}(\hat{\beta},\hat{\gamma}) + \hat{Q}_k(t;\hat{\gamma})\hat{\Omega}(\hat{\gamma})^{-1}\sum_{l=1}^{K}\Gamma_{il}(\hat{\gamma}) + \int_0^t \bar{Y}_k(s)^{-1}d\hat{M}_{ik}(s;\hat{\beta},\hat{\gamma}).$$

Remark 2. Note that $\hat{\mu}_{0k}(t)$ could have negative increments. However, as discussed in Lin and Ying (1994), Yin and Cai (2004) and Schaubel et al. (2006), simple modifications can be made to ensure monotonicity while preserving the established asymptotic properties, that is,

$$\tilde{\mu}_{0k}(t) = \max_{0\le u\le t}\hat{\mu}_{0k}(u;\hat{\beta},\hat{\gamma}).$$

Following the similar arguments to those in Lin and Ying (1994), it can be shown that $\{\tilde{\mu}_{0k}(t)-\hat{\mu}_{0k}(t)\} = o_p(n^{-1/2})$ uniformly in $t\in[0,\tau]$.

4. Model checking

Note that model inadequacy could result from the event category model. To check the generalized logits model (3), we can use some model checking procedures such as the Hosmer–Lemeshow test, the classification table and the ROC curve (Hosmer and Lemeshow, 2000). Here, we propose a formal lack-of-fit test for assessing the adequacy of model (1). Following Lin et al. (1993), we consider the following cumulative sums of residuals:

$$\mathcal{L}_k(t,z) = n^{-1/2}\sum_{i=1}^{n}\int_0^t I(Z_{ik}(s)\le z)d\hat{M}_{ik}(s;\hat{\beta},\hat{\gamma}),$$

where the event $I(Z_{ik}(s)\le z)$ means that each component of $Z_{ik}(s)$ is no larger than the corresponding component of z (Lin et al., 2000). Define the null hypothesis as the correct specification of model (1) under the assumption that model (3) are correctly specified. For the null distribution of $\mathcal{L}_k(t,z)$, we have the following theorem with the proof given in the Appendix.

Theorem 3. *Under the assumptions of Theorem 1, the null distribution of $\mathcal{L}_k(t,z)$ converges weakly to a zero-mean Gaussian process with covariance function at (t_1,z_1) and (t_2,z_2) equal to $E\{\sigma_{ik}(t_1,z_1)\sigma_{ik}(t_2,z_2)\}$, where*

$$\begin{aligned}\sigma_{ik}(t,z) &= \int_0^t\left\{I(Z_{ik}(s)\le z) - \frac{D_k(s,z)}{\bar{y}_k(s)}\right\}dM_{ik}(s;\beta_0,\gamma_0)\\ &\quad+\left\{E_k(t,z)-F_k(t,z)\right\}\Omega(\gamma_0)^{-1}\sum_{l=1}^{K}\Gamma_{il}(\gamma_0) - L_k(t,z)'A^{-1}\sum_{l=1}^{K}\Phi_{il}(\beta_0,\gamma_0),\\ E_k(t,z) &= E\left\{\int_0^t I(Z_{ik}(s)\le z)\Big[W_{ik}(s)-\sum_{l=1}^{K}\pi_{il}(s;\gamma_0)W_{il}(s)\Big]'\pi_{ik}(s;\gamma_0)dN_{i\cdot}^c(s)\right\},\\ F_k(t,z) &= E\left\{\int_0^t \frac{D_k(s,z)}{\bar{y}_k(s)}\Big[W_{ik}(s)-\sum_{l=1}^{K}\pi_{il}(s;\gamma_0)W_{il}(s)\Big]'\pi_{ik}(s;\gamma_0)dN_{i\cdot}^c(s)\right\},\\ L_k(t,z) &= E\left\{\int_0^t Y_{ik}(s)I(Z_{ik}(s)\le z)\{Z_{ik}(s)-\bar{z}_k(s)\}ds\right\},\end{aligned}$$

and $D_k(t,z) = E\{Y_{ik}(t)I(Z_{ik}(t)\le z)\}$.

It follows from Theorem 3 that the null distribution of $\mathcal{L}_k(t,z)$ can be approximated by the zero-mean Gaussian process

$$\tilde{\mathcal{L}}_k(t,z) = n^{-1/2}\sum_{i=1}^{n}\hat{\Upsilon}_{ik}(t,z),$$

where

$$\begin{aligned}\hat{\Upsilon}_{ik}(t,z) &= \int_0^t\left\{I(Z_{ik}(s)\le z) - \frac{\hat{D}_k(s,z)}{\bar{Y}_k(s)}\right\}d\hat{M}_{ik}(s;\hat{\beta},\hat{\gamma})\\ &\quad+\left\{\hat{E}_k(t,z)-\hat{F}_k(t,z)\right\}\hat{\Omega}(\hat{\gamma})^{-1}\sum_{l=1}^{K}\Gamma_{il}(\hat{\gamma}) - \hat{L}_k(t,z)'\hat{A}^{-1}\sum_{l=1}^{K}\hat{\Phi}_{il}(\hat{\beta},\hat{\gamma}),\end{aligned}$$

$$\hat{E}_k(t,z) = n^{-1}\sum_{i=1}^{n}\int_0^t I(Z_{ik}(s)\le z)\Big[W_{ik}(s) - \sum_{l=1}^{K} W_{il}(s)\pi_{il}(s;\hat{\gamma})\Big]'\pi_{ik}(s;\hat{\gamma})dN_{i\cdot}^c(s),$$

$$\hat{F}_k(t,z) = n^{-1}\sum_{i=1}^{n}\int_0^t \frac{\hat{D}_k(s,z)}{\bar{Y}_k(s)}\Big[W_{ik}(s) - \sum_{l=1}^{K} \pi_{il}(s;\hat{\gamma})W_{il}(s)\Big]'\pi_{ik}(s;\hat{\gamma})dN_{i\cdot}^c(s),$$

$$\hat{L}_k(t,z) = n^{-1}\sum_{i=1}^{n}\int_0^t Y_{ik}(s)I(Z_{ik}(s)\le z)\{Z_{ik}(s) - \bar{Z}_k(s)\}ds,$$

and $\hat{D}_k(t,z) = n^{-1}\sum_{i=1}^{n} Y_{ik}(t)I(Z_{ik}(t)\le z)$.

It is difficult to estimate the asymptotic covariance function of $\mathcal{L}_k(t,z)$ analytically because the limiting process of $\mathcal{L}_k(t,z)$ does not have an independent increment structure. To this end, we can utilize the resampling approach (Lin et al., 1993, 2000). Let $\{G_1,\ldots,G_n\}$ be independent standard normal variables which are independent of the observed data. Then it can be shown that the null distribution of $\mathcal{L}_k(t,z)$ can be approximated by the conditional distribution of $\hat{\mathcal{L}}_k(t,z)$, where

$$\hat{\mathcal{L}}_k(t,z) = n^{-1/2}\sum_{i=1}^{n}\hat{\Upsilon}_{ik}(t,z)G_i. \tag{9}$$

Thus, we can obtain a large number of realizations from $\hat{\mathcal{L}}_k(t,z)$ by repeatedly generating the standard normal random sample $\{G_1,\ldots,G_n\}$ while fixing the observed data. To evaluate the lack-of-fit of model (1), we could plot $\mathcal{L}_k(t,z)$ along with a few number of realizations of $\hat{\mathcal{L}}_k(t,z)$ to see if there are some unusual patterns. Since $\mathcal{L}_k(t,z)$ is expected to fluctuate randomly around 0 under the assumed model, a formal lack-of-fit test can be constructed based on the supremum statistic $\sup_{0\le t\le\tau,z}|\mathcal{L}_k(t,z)|$, with which the p-value can be obtained by comparing the observed value of $\sup_{0\le t\le\tau,z}|\mathcal{L}_k(t,z)|$ to a large number of realizations from $\sup_{0\le t\le\tau,z}|\hat{\mathcal{L}}_k(t,z)|$.

5. Simulation studies

Simulation studies were conducted to examine the finite-sample properties of the proposed estimators. In the study, we considered the situation where there exist $K = 2$ event categories. The covariates were taken as $Z_{i1} = (Z_i, 0)'$ and $Z_{i2} = (0, Z_i)'$, where Z_i follows a Bernoulli distribution with success probability 0.5. Let R_i be a gamma random variable with mean 1 and variance σ_R^2, which was introduced to induce positive correlation among the within-subject events. To avoid yielding too many recurrent events for one subject, we set $R_i^* = \min(R_i, 1.5)$ with $\sigma_R^2 = 0.5$ and 1. The kth type recurrent events were generated from a Poisson process with the intensity function

$$\lambda_{ik}(t) = R_i^* + \lambda_{0k} + \beta_0' Z_{ik}, \quad k = 1, 2,$$

where $\beta_0 = (\beta_1, \beta_2)' = (0.5, 0.3)'$, $\lambda_{01} = 0.25$ or 0.5, and $\lambda_{02} = 0.5$. It can be verified that $N_{ik}^*(t)$ satisfies the additive rates model (1) with $d\mu_{0k}(t) = \{E(R_i^*) + \lambda_{0k}\}dt$. The censoring time C_{ik} was generated from a uniform distribution $U(0,\tau)$ with $\tau = 5$. Under the preceding settings, the average number of observed events per subject ranged from 3.2 to 5.4 for $k = 1$ and from 3.6 to 5.2 for $k = 2$. The event categories were set to missing with probability

$$P\{\xi_i(t) = 0 | dN_{i\cdot}(t) = 1, Z_i\} = \frac{\exp\{\rho_0 + Z_i\rho_z\}}{1 + \exp\{\rho_0 + Z_i\rho_z\}},$$

where $\rho_0 = -1$, and $\rho_z = 0, \log(1.5)$ or $\log(2)$. Under the above settings, the proportion of events with missing types varied from 27% to 36%.

For comparison, three methods were used to estimate β_0: (i) the full-data (FF) analysis, which is based on data with all event categories being always observed; (ii) the complete-case (CC) analysis, which treats events of unknown category as censored; (iii) the proposed weighted estimating equation (WEE) method. For the WEE method, the event category probability was fitted by the following logistic model:

$$\log\left\{\frac{\pi_{i2}(t;\gamma_0)}{\pi_{i1}(t;\gamma_0)}\right\} = \gamma_0' W_{i2}(t),$$

where $W_{i2}(t) = (1, Z_i)'$. The results presented below are based on 500 replications with sample sizes $n = 100$ and 200.

All the simulation results are summarized in Tables 1 and 2. In these tables, Bias is the sample mean of the estimate minus the true value; ESD is the empirical standard deviation of the estimate; ASE is the average estimated standard error; RE is the relative efficiency (computed as the ratio of empirical variances); and CP is the 95% empirical coverage probability based on the normal approximation.

Table 1 presents the comparison results on estimation of β_1 only, as those for the estimate of β_2 are very similar. It can be seen from Table 1 that the CC estimator is nearly unbiased only when $\rho_z = 0$ (i.e., the event categories are missing completely at random). However, the CC estimator is highly biased when $\rho_z > 0$, and the bias increases as the correlation

Table 1
Comparison results on estimation of $\beta_1 = 0.5$.

n	σ_R^2	λ_{01}	ρ_z	Bias			ESD			RE:WEE	
				FF	WEE	CC	FF	WEE	CC	vs.FF	vs.CC
100	0.5	0.25	0	−0.002	−0.001	0.005	0.186	0.200	0.212	0.86	1.12
			log(1.5)	−0.002	−0.003	−0.242	0.186	0.205	0.216	0.82	1.11
			log(2)	−0.002	−0.003	−0.454	0.186	0.208	0.218	0.80	1.10
		0.5	0	−0.007	−0.011	−0.014	0.197	0.213	0.223	0.86	1.10
			log(1.5)	−0.007	−0.012	−0.286	0.197	0.215	0.229	0.84	1.13
			log(2)	−0.007	−0.011	−0.523	0.197	0.220	0.234	0.80	1.13
	1	0.25	0	0.002	−0.001	−0.003	0.172	0.183	0.192	0.88	1.10
			log(1.5)	0.002	−0.002	−0.163	0.172	0.185	0.197	0.86	1.13
			log(2)	0.002	−0.001	−0.305	0.172	0.189	0.201	0.83	1.13
		0.5	0	0.018	0.022	0.019	0.197	0.204	0.216	0.93	1.12
			log(1.5)	0.018	0.022	−0.170	0.197	0.205	0.218	0.92	1.13
			log(2)	0.018	0.022	−0.343	0.197	0.207	0.224	0.91	1.17
200	0.5	0.25	0	0.008	0.006	0.005	0.128	0.139	0.147	0.85	1.12
			log(1.5)	0.008	0.006	−0.235	0.128	0.141	0.152	0.82	1.16
			log(2)	0.008	0.005	−0.448	0.128	0.144	0.156	0.79	1.17
		0.5	0	0.004	0.002	0.004	0.137	0.147	0.157	0.87	1.14
			log(1.5)	0.004	0.002	−0.265	0.137	0.151	0.162	0.82	1.15
			log(2)	0.004	0.002	−0.504	0.137	0.153	0.170	0.80	1.23
	1	0.25	0	−0.005	−0.008	−0.006	0.133	0.141	0.148	0.89	1.10
			log(1.5)	−0.002	0.005	−0.158	0.132	0.146	0.156	0.82	1.14
			log(2)	−0.002	0.006	−0.299	0.132	0.147	0.157	0.81	1.14
		0.5	0	−0.007	−0.003	−0.004	0.135	0.143	0.149	0.89	1.09
			log(1.5)	−0.003	−0.007	−0.200	0.144	0.154	0.165	0.87	1.15
			log(2)	−0.003	−0.005	−0.368	0.144	0.157	0.169	0.84	1.16

Table 2
Simulation results for the accuracy of the asymptotic approximation to the distributions of the WEE estimator.

n	σ_R^2	λ_{01}	ρ_z	β_1			β_2		
				ASE	ESD	CP	ASE	ESD	CP
100	0.5	0.25	0	0.197	0.200	0.940	0.202	0.210	0.928
			log(1.5)	0.201	0.205	0.936	0.205	0.212	0.928
			log(2)	0.205	0.208	0.934	0.209	0.215	0.938
		0.5	0	0.207	0.213	0.936	0.206	0.209	0.946
			log(1.5)	0.211	0.215	0.944	0.211	0.213	0.940
			log(2)	0.215	0.220	0.942	0.214	0.220	0.938
	1	0.25	0	0.195	0.183	0.962	0.201	0.190	0.960
			log(1.5)	0.198	0.185	0.960	0.204	0.191	0.954
			log(2)	0.201	0.189	0.952	0.206	0.190	0.964
		0.5	0	0.206	0.204	0.940	0.203	0.208	0.940
			log(1.5)	0.209	0.205	0.942	0.206	0.212	0.924
			log(2)	0.212	0.207	0.954	0.208	0.215	0.928
200	0.5	0.25	0	0.141	0.139	0.946	0.145	0.147	0.932
			log(1.5)	0.144	0.141	0.948	0.147	0.149	0.938
			log(2)	0.146	0.144	0.938	0.150	0.151	0.944
		0.5	0	0.148	0.147	0.958	0.145	0.144	0.950
			log(1.5)	0.151	0.151	0.952	0.148	0.150	0.954
			log(2)	0.154	0.153	0.954	0.151	0.153	0.954
	1	0.25	0	0.140	0.141	0.950	0.143	0.143	0.942
			log(1.5)	0.142	0.146	0.928	0.146	0.147	0.948
			log(2)	0.144	0.147	0.934	0.148	0.150	0.938
		0.5	0	0.147	0.143	0.954	0.144	0.150	0.934
			log(1.5)	0.149	0.154	0.930	0.146	0.149	0.928
			log(2)	0.152	0.157	0.932	0.148	0.151	0.932

between the missingness probability and Z_i increases. Both the FF and WEE estimators are essentially unbiased in all settings. Furthermore, the WEE estimator is more efficient than the CC estimator, and is only slightly less efficient than the FF estimator.

Table 2 gives the simulation results on the accuracy of the asymptotic approximation to the distributions of the WEE estimator. The results show that the WEE method performs well for the situations considered here. Specifically, the average estimated standard errors are very close to the empirical standard deviations, and the 95% empirical coverage probabilities are reasonable. The performance of the WEE estimator becomes better when the sample size increases from 100 to 200, We also considered other setups and the results were similar to those given above.

Table 3
Analysis results for the FNHTRs data.

Type	Covariate	CC method			WEE method		
		Est $\times 10^2$	SE $\times 10^2$	p-value	Est $\times 10^2$	SE $\times 10^2$	p-value
I	Gender	−1.4801	1.0469	0.1574	0.2012	0.9975	0.8401
	Age	0.1092	0.0272	<0.01	0.0069	0.0222	0.7542
II	Gender	−0.9899	0.8940	0.2682	−0.0555	0.7486	0.9409
	Age	0.1011	0.0272	<0.01	0.0460	0.0208	0.0271

Note: Type I denotes the reaction with fever; Type II denotes the reaction with no fever; Est is the estimate of the parameter; and SE denotes the standard error estimate.

6. Application

For the illustration purpose, we applied the proposed method to a set of recurrent event data from some patients who may experience different febrile nonhemolytic transfusion reactions (FNHTRs) arising within 4–6 h post transfusion. The data were collected at five university teaching hospitals in Toronto coded A–E over three consecutive summers from 1996 to 1998 (Patterson et al., 2000). We considered a subset of the data which consist of 242 patients who were followed up during the 1997 summer. The occurrence of FNHTRs is temporary and thus it is reasonable to treat a reaction as a recurrent event. A total of 1201 transfusions were recorded and among them, there were 314 reaction episodes being observed and the mean number of reactions per patient was 1.3 ($sd = 1.8$). The FNHTR is characterized by fever, chill, rigor, hive and other symptoms. In the following analysis, for simplicity, we classified all kinds of reactions into two types: the reaction accompanied with fever (denoted by Type I reaction) and the reaction with no fever (denoted by Type II reaction). Based on the above classification, among the 314 observed reactions, there were 181 Type I reactions, 115 Type II reactions, and 18 reactions with missing types. Thus we got bivariate recurrent event data in the presence of missing event types.

Following Zhao et al. (2012), we defined $N^*_{i1}(t)$ and $N^*_{i2}(t)$ as the numbers of fever and no fever reactions that had occurred over interval $[0, t]$ for patient i, respectively. The covariate vector Z_i includes the gender of patients (1 if female, 0 if male) and the age of patients when entering the study (in years). The censoring time C_i is defined as the day of the last visit for patient i, and let $\tau = 165$ denote the maximum value of C_i's. In this dataset, the proportion of events with missing types is quite small (less than 6%). To illustrate our method, following the idea of Lu and Liang (2008), we further artificially deleted some event types for those recurrent events with known types according to the MAR mechanism. Specifically, the missing probability was chosen as $p = \exp(2 + 0.5 * \text{Gender} - 0.1 * \text{Age})/(1 + \exp(2 + 0.5 * \text{Gender} - 0.1 * \text{Age}))$, which leads to about 28% events with missing types. Our main goal is to estimate the covariate effects on the recurrence rate of the two types of transfusion reactions. In the interest of flexibility, both covariates are assumed to be type-specific.

For the analysis, we considered model (1) with $Z_{i1} = (Z_i', 0)'$ and $Z_{i2} = (0, Z_i')'$. Two methods were employed for comparison: WEE methods and CC methods. For the WEE methods, $\pi_{ik}(t; \gamma_0)$ was assumed to satisfy the following generalized logistic model

$$\pi_{ik}(t; \gamma_0) = \frac{\exp\{\gamma_0' W_{ik}\}}{\sum_{l=1}^{2} \exp\{\gamma_0' W_{il}\}}, \quad k = 1, 2$$

where $W_{i2} = (1, Z_i')'$ and $W_{i1} = 0$. The results are summarized in Table 3. These results indicate that by the WEE method, neither FNHTR rates seem to be correlated with the gender of the patients, and younger patients may experience lower risk of Type II platelet transfusion reactions. These results are consistent with those obtained by Zhao et al. (2012). On the other hand, the CC estimators for all the effects are significantly different from the WEE estimators, and the CC method would overestimate the effects of gender and age. In addition, the standard errors of the WEE estimators are smaller than those of the CC estimators as shown in the simulation.

For model checking, we first used the Hosmer–Lemeshow test (Hosmer and Lemeshow, 2000) to check the adequacy of the assumed logistic model. To calculate the test statistic, we ordered the fitted value $\hat{\pi}_{i2}(t; \hat{\gamma})$, and grouped them into 10 classes of roughly equal size. The test statistic value was about 12.54 with a p-value of 0.128, which indicates little evidence against the assumed logistic model. Finally, we apply the model checking techniques introduced in Section 4 to assess the adequacy of model (1) for the data. We calculated the statistics $\mathcal{L}_k(t, z)$ $(k = 1, 2)$, and obtained $\sup_{0\le t\le\tau, z} |\mathcal{L}_1(t, z)| = 0.890$ and $\sup_{0\le t\le\tau, z} |\mathcal{L}_2(t, z)| = 0.659$ with p-values of 0.324 and 0.316, respectively, based on 500 realizations of the corresponding statistics $\sup_{0\le t\le\tau, z} |\hat{\mathcal{L}}_1(t, z)|$ and $\sup_{0\le t\le\tau, z} |\hat{\mathcal{L}}_2(t, z)|$. These results suggest that model (1) fits the data adequately.

7. Concluding remarks

In this article, we proposed an additive rates model for multivariate recurrent event data with missing event categories under the MAR assumption. A weighted estimating equation method was developed for parameter estimation. Specifically, when an event category was missing, a weighted contribution was added to the estimating equation, with the weight

equal to the corresponding category-specific probability. The resulting estimators have explicit expressions and are easy to implement. The asymptotic properties of the proposed estimators were established. Simulation results showed that the proposed methods work well, and an application to the FNHTR data was provided.

Model (1) has the limitation that the linear predictor $\beta_0' Z_{ik}(t)$ needs to be constrained to ensure non-negativity for the right-hand side of (1). One may avoid this constraint by using a nonnegative link function, such as $d\mu_{0k}(t) + g(\beta_0' Z_{ik}(t))dt$. The proposed estimation procedure can be extended in a straightforward manner to deal with any regression function $g(\beta_0' Z_{ik}(t))$, where $g(\cdot)$ is a known link function. Furthermore, it would be worthwhile to study multivariate recurrent events with missing event categories under other competing models, such as an additive–multiplicative rates model (Liu et al., 2010) and a class of semiparametric transformation models (Lin et al., 2001; Zeng and Lin, 2006). In addition, Lin et al. (2013) developed fully nonparametric methods in which the missingness mechanism is completely unspecified. It would be desirable to develop similar nonparametric methods for an additive rates model.

In general, there are three types of assumptions on missingness: missing completely at random (MCAR), MAR and missing not at random (MNAR). The MAR assumption is common for statistical analysis with missing data and is reasonable in many applications (Little and Rubin, 2002), and MCAR is a special case of MAR. For MNAR, because missingness depends on missing data, some nonidentifiability problems arise (Tsiatis, 2006). Development of an estimation method to analyze multivariate recurrent event data with missing event categories under the MNAR assumption is challenging and merits future research.

Acknowledgments

The authors would like to thank the Editor, Professor Jae Chang Lee, the Associate Editor and the two reviewers for their constructive and insightful comments and suggestions that greatly improved the paper. Zhao's research was partly supported by the Research Grant Council of Hong Kong (Nos. 504011 and 503513), the National Natural Science Foundation of China (No. 11371299), and The Hong Kong Polytechnic University. Sun's research was partly supported by the National Natural Science Foundation of China Grant (Nos. 11231010 and 11171330), Key Laboratory of RCSDS, CAS (No. 2008DP173182), and BCMIIS. The authors thank Dr. Bruce J. Patterson for providing them with example data.

Appendix. Proofs of asymptotic results

Proof of Theorem 1. It can be checked that

$$\begin{aligned}\hat{\beta} - \beta_0 &= \hat{A}^{-1}\left[n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^{\tau}\{Z_{ik}(t) - \bar{Z}_k(t)\}\left\{dM_{ik}(t;\beta_0,\hat{\gamma}) + Y_{ik}(t)d\mu_{0k}(t)\right\}\right] \\ &= \hat{A}^{-1}n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^{\tau}\{Z_{ik}(t) - \bar{Z}_k(t)\}dM_{ik}(t;\beta_0,\hat{\gamma}), \end{aligned} \tag{A.1}$$

where

$$\hat{A} = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^{\tau} Y_{ik}(t)\{Z_{ik}(t) - \bar{Z}_k(t)\}^{\otimes 2}dt.$$

Using the strong law of large numbers (van der Vaart, 2000) and the strong consistency of $\hat{\gamma}$, we obtain

$$n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^{\tau}\{Z_{ik}(t) - \bar{Z}_k(t)\}dM_{ik}(t;\beta_0,\hat{\gamma}) \to 0, \quad a.s.$$

and $\hat{A}$ converges almost surely to A, which is nonsingular by condition (C4). Thus, it follows from (A.1) that $\hat{\beta}$ is strongly consistent to β_0.

To prove asymptotic normality of $\hat{\beta}$, write

$$n^{1/2}(\hat{\beta} - \beta_0) = \hat{A}^{-1}\sum_{k=1}^{K}\sum_{l=1}^{2} U_{kl}, \tag{A.2}$$

where

$$\begin{aligned} U_{k1} &= n^{-1/2}\sum_{i=1}^{n}\int_0^{\tau}\{Z_{ik}(t) - \bar{Z}_k(t)\}dM_{ik}(t), \\ U_{k2} &= n^{-1/2}\sum_{i=1}^{n}\int_0^{\tau}\{Z_{ik}(t) - \bar{Z}_k(t)\}\left\{dM_{ik}(t;\beta_0,\hat{\gamma}) - dM_{ik}(t)\right\}, \end{aligned}$$

and $M_{ik}(t) = M_{ik}(t; \beta_0, \gamma_0)$. Using the functional central limit theorem (Pollard, 1990), we have

$$U_{k1} = n^{-1/2} \sum_{i=1}^{n} \int_0^{\tau} \{Z_{ik}(t) - \bar{z}_k(t)\} dM_{ik}(t) + o_p(1). \tag{A.3}$$

In view of (4), by the Taylor expansion and some straightforward calculations, we have

$$\begin{aligned} U_{k2} &= n^{-1/2} \sum_{i=1}^{n} \int_0^{\tau} \{Z_{ik}(t) - \bar{Z}_k(t)\} \{\pi_{ik}(t; \hat{\gamma}) - \pi_{ik}(t; \gamma_0)\} dN_{i\cdot}^c(t) \\ &= n^{-1/2} \sum_{i=1}^{n} \Psi_k(\gamma_0) \Omega(\gamma_0)^{-1} \sum_{l=1}^{K} \Gamma_{il}(\gamma_0) + o_p(1), \end{aligned} \tag{A.4}$$

where $\Psi_k(\gamma)$, $\Omega(\gamma)$ and $\Gamma_{il}(\gamma)$ are as defined in Theorem 1. It follows from (A.2)–(A.4) and the strong consistency of $\hat{A}$ that

$$n^{1/2}(\hat{\beta} - \beta_0) = A^{-1} n^{-1/2} \sum_{i=1}^{n} \sum_{k=1}^{K} \Phi_{ik}(\beta_0, \gamma_0) + o_p(1), \tag{A.5}$$

where

$$\Phi_{ik}(\beta, \gamma) = \int_0^{\tau} \{Z_{ik}(t) - \bar{z}_k(t)\} dM_{ik}(t; \beta, \gamma) + \Psi_k(\gamma) \Omega(\gamma)^{-1} \sum_{l=1}^{K} \Gamma_{il}(\gamma).$$

Utilizing the multivariate central limit theorem, $n^{1/2}(\hat{\beta} - \beta_0)$ is asymptotically normal with mean zero and covariance matrix $A^{-1} \Sigma A^{-1}$, where Σ is given in Theorem 1.

Proof of Theorem 2. First write

$$\hat{\mu}_{0k}(t) - \mu_{0k}(t) = \{\hat{\mu}_{0k}(t) - \hat{\mu}_{0k}(t; \beta_0, \hat{\gamma})\} + \{\hat{\mu}_{0k}(t; \beta_0, \hat{\gamma}) - \hat{\mu}_{0k}(t; \beta_0, \gamma_0)\} + \{\hat{\mu}_{0k}(t; \beta_0, \gamma_0) - \mu_{0k}(t)\}. \tag{A.6}$$

It is easy to see that

$$\hat{\mu}_{0k}(t) - \hat{\mu}_{0k}(t; \beta_0, \hat{\gamma}) = -\int_0^t \bar{Z}_k(s)' ds (\hat{\beta} - \beta_0). \tag{A.7}$$

In addition, by the Taylor expansion, we have

$$\hat{\mu}_{0k}(t; \beta_0, \hat{\gamma}) - \hat{\mu}_{0k}(t; \beta_0, \gamma_0) = \int_0^t \frac{\sum_{i=1}^{n} \pi_{ik}(s; \gamma^*) \left[W_{ik}(s) - \sum_{l=1}^{K} \pi_{il}(s; \gamma^*) W_{il}(s) \right]' dN_{i\cdot}^c(s)}{\sum_{j=1}^{n} Y_{jk}(s)} \times (\hat{\gamma} - \gamma_0), \tag{A.8}$$

where γ^* lies between $\hat{\gamma}$ and γ_0. Under conditions (C1)–(C3), the integrals in (A.7) and (A.8) are bounded almost surely uniformly in $t \in [0, \tau]$. Hence using the consistency of $\hat{\beta}$ and $\hat{\gamma}$, we get that $\hat{\mu}_{0k}(t) - \hat{\mu}_{0k}(t; \beta_0, \hat{\gamma})$ and $\hat{\mu}_{0k}(t; \beta_0, \hat{\gamma}) - \hat{\mu}_{0k}(t; \beta_0, \gamma_0)$ converge almost surely to 0 uniformly in $t \in [0, \tau]$, respectively.

For the third term on the right-hand side of (A.6), after some algebraic manipulations, we obtain

$$\hat{\mu}_{0k}(t; \beta_0, \gamma_0) - \mu_{0k}(t) = \sum_{i=1}^{n} \int_0^t \frac{dM_{ik}(s)}{\sum_{j=1}^{n} Y_{jk}(s)}. \tag{A.9}$$

By the uniform strong law of large numbers (Pollard, 1990) and Lemma 1 of Lin et al. (2000), it can be seen that $\hat{\mu}_{0k}(t; \beta_0, \gamma_0) - \mu_{0k}(t)$ converges almost surely to 0 uniformly in $t \in [0, \tau]$. Thus, it follows from (A.6)–(A.9) that $\hat{\mu}_{0k}(t)$ converges almost surely to $\mu_{0k}(t)$ uniformly in $t \in [0, \tau]$.

Next, we show the weak convergence of $\hat{\mu}_{0k}(t)$. It follows from (A.5) and (A.7) that

$$n^{1/2}\{\hat{\mu}_{0k}(t) - \hat{\mu}_{0k}(t; \beta_0, \hat{\gamma})\} = H_k(t)' A^{-1} n^{-1/2} \sum_{i=1}^{n} \sum_{k=1}^{K} \Phi_{ik}(\beta_0, \gamma_0) + o_p(1) \tag{A.10}$$

uniformly in $t \in [0, \tau]$, where $H_k(t) = -\int_0^t \bar{z}_k(s) ds$. Based on (4) and (A.8), a straightforward calculation yields

$$n^{1/2}\{\hat{\mu}_{0k}(t; \beta_0, \hat{\gamma}) - \hat{\mu}_{0k}(t; \beta_0, \gamma_0)\} = Q_k(t; \gamma_0) \Omega(\gamma_0)^{-1} n^{-1/2} \sum_{i=1}^{n} \sum_{l=1}^{K} \Gamma_{il}(\gamma_0) + o_p(1) \tag{A.11}$$

uniformly in $t \in [0, \tau]$, where $Q_k(t; \gamma)$ is as defined in Theorem 2. In addition, by (A.9), we have that uniformly in $t \in [0, \tau]$,

$$n^{1/2}\{\hat{\mu}_{0k}(t; \beta_0, \gamma_0) - \mu_{0k}(t)\} = n^{-1/2} \sum_{i=1}^{n} \int_0^t \frac{dM_{ik}(s)}{\bar{y}_k(s)} + o_p(1), \tag{A.12}$$

where $\bar{y}_k(s)$ is the limit of $\bar{Y}_k(s)$. Thus, it follows from (A.6) and (A.10)–(A.12) that

$$n^{1/2}\{\hat{\mu}_{0k}(t) - \mu_{0k}(t)\} = n^{-1/2} \sum_{i=1}^{n} \phi_{ik}(t; \beta_0, \gamma_0) + o_p(1) \tag{A.13}$$

uniformly in $t \in [0, \tau]$, where $\phi_{ik}(t; \beta, \gamma)$ is as defined in Theorem 2. Because $\phi_{ik}(t; \beta_0, \gamma_0)$ are independent zero-mean random variables for each t, the multivariate central limit theorem implies that $n^{1/2}\{\hat{\mu}_{0k}(t) - \mu_{0k}(t)\}$ converges in finite-dimensional distributions to a zero-mean Gaussian process. Note that $H_k(t)$ and $Q_k(t; \gamma_0)$ are deterministic functions, and the third term of $\phi_{ik}(t; \beta_0, \gamma_0)$ can be written as sums of monotone functions of t. Thus, $n^{1/2}\{\hat{\mu}_{0k}(t) - \mu_{0k}(t)\}$ is tight (van der Vaart and Wellner, 1996), and converges weakly to a zero-mean Gaussian process whose covariance function at (s, t) is given by $\omega_k(s, t)$, which is defined in Theorem 2.

Proof of Theorem 3. First note that

$$\mathcal{L}_k(t, z) = n^{-1/2} \sum_{i=1}^{n} \int_0^t I(Z_{ik}(s) \le z) dM_{ik}(s) + R_{k1}(t, z) + R_{k2}(t, z) + R_{k3}(t, z), \tag{A.14}$$

where

$$R_{k1}(t, z) = n^{-1/2} \sum_{i=1}^{n} \int_0^t I(Z_{ik}(s) \le z)\{\pi_{ik}(s; \hat{\gamma}) - \pi_{ik}(s; \gamma_0)\} dN_{i\cdot}^c(s),$$

$$R_{k2}(t, z) = -n^{-1/2} \sum_{i=1}^{n} \int_0^t I(Z_{ik}(s) \le z) Y_{ik}(s) d\{\hat{\mu}_{0k}(s) - \mu_{0k}(s)\},$$

and

$$R_{k3}(t, z) = -n^{-1/2} \sum_{i=1}^{n} \int_0^t I(Z_{ik}(s) \le z) Y_{ik}(s) Z_{ik}(s)' ds(\hat{\beta} - \beta_0).$$

Similarly to (A.4), we obtain

$$R_{k1}(t, z) = n^{-1/2} \sum_{i=1}^{n} E_k(t, z) \Omega(\gamma_0)^{-1} \sum_{l=1}^{K} \Gamma_{il}(\gamma_0) + o_p(1), \tag{A.15}$$

where $E_k(t, z)$ is defined in Theorem 3. Using the uniform strong law of large numbers and (A.13), we have

$$\begin{aligned} R_{k2}(t, z) = & -n^{-1/2} \sum_{i=1}^{n} \Bigg\{ L_{2k}(t, z)' A^{-1} \sum_{l=1}^{K} \Phi_{il}(\beta_0, \gamma_0) + F_k(t, z) \Omega(\gamma_0)^{-1} \sum_{l=1}^{K} \Gamma_{il}(\gamma_0) \\ & + \int_0^t \frac{D_k(s, z)}{\bar{y}_k(s)} dM_{ik}(s) \Bigg\} + o_p(1), \end{aligned} \tag{A.16}$$

where

$$L_{2k}(t, z) = \int_0^t D_k(s, z) dH_k(s),$$

and $D_k(t, z)$ and $F_k(t, z)$ are defined in Theorem 3. In addition, by (A.5), we get

$$R_{k3}(t, z) = -n^{-1/2} \sum_{i=1}^{n} L_{1k}(t, z)' A^{-1} \sum_{l=1}^{K} \Phi_{il}(\beta_0, \gamma_0) + o_p(1), \tag{A.17}$$

where

$$L_{1k}(t, z) = E\left\{ \int_0^t Y_{ik}(s) I(Z_{ik}(s) \le z) Z_{ik}(s) ds \right\}.$$

Let $L_k(t,z) = L_{1k}(t,z) + L_{2k}(t,z)$. Then it follows from (A.14)–(A.17) that

$$\mathcal{L}_k(t,z) = n^{-1/2}\sum_{i=1}^{n}\sigma_{ik}(t,z) + o_p(1),$$

where $\sigma_{ik}(t,z)$ is defined in Theorem 3. By the multivariate central limit theorem, $\mathcal{L}_k(t,z)$ converges in finite-dimensional distributions to a zero-mean Gaussian process. By the same argument as the tightness of $n^{1/2}\{\hat{\mu}_{0k}(t) - \mu_{0k}(t)\}$, $\mathcal{L}_k(t,z)$ is tight. Thus, $\mathcal{L}_k(t,z)$ converges weakly to a zero-mean Gaussian process with covariance function at (t_1,z_1) and (t_2,z_2) equal to $E\{\sigma_{ik}(t_1,z_1)\sigma_{ik}(t_2,z_2)\}$. By the arguments of Lin et al. (2000), the limiting Gaussian process can be approximated by the zero-mean Gaussian process $\hat{\mathcal{L}}_k(t,z)$ given in (9).

References

Abu-Libdeh, H., Turnbull, B.W., Clark, L.C., 1990. Analysis of muti-type recurrent events in longitudinal studies: application to a skin cancer prevention trial. Biometrics 46, 1017–1034.

Cai, J., Schaubel, D.E., 2004. Marginal mean/rates models for multiple type recurrent event data. Lifetime Data Anal. 10, 121–138.

Chen, B.E., Cook, R.J., 2009. The analysis of multivariate recurrent events with partially missing event types. Lifetime Data Anal. 15, 41–58.

Chen, B.E., Cook, R.J., Lawless, J.F., Zhan, M., 2005. Statistical methods for multivariate interval-censored recurrent events. Stat. Med. 24, 671–691.

Chen, X., Wang, Q., Cai, J., Shankar, V., 2012. Semiparametric additive marginal regression models for multiple type recurrent events. Lifetime Data Anal. 18, 504–527.

Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression, second ed.. Wiley, New York.

Kalbfleisch, J.D., Prentice, R.L., 2002. The Statistical Analysis of Failure Time Data, second ed.. Wiley, New York.

Liang, K.Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.

Lin, F., Cai, J., Fine, J.P., Lai, H.J., 2013. Nonparametric estimation of the mean function for recurrent event data with missing event category. Biometrika 100, 727–740.

Lin, D.Y., Wei, L.J., Yang, I., Ying, Z., 2000. Semiparametric regression for the mean and rate functions of recurrent events. J. Roy. Statist. Soc. Ser. B 62, 711–730.

Lin, D.Y., Wei, L.J., Ying, Z., 1993. Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika 80, 557–572.

Lin, D.Y., Wei, L.J., Ying, Z., 2001. Semiparametric transformation models for point processes. J. Amer. Statist. Assoc. 96, 620–628.

Lin, D.Y., Ying, Z., 1994. Semiparametric analysis of the additive risk model. Biometrika 81, 61–71.

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data. Wiley, New York.

Liu, Y., Wu, Y., Cai, J., Zhou, H., 2010. Additive-multiplicative rates model for recurrent events. Lifetime Data Anal. 16, 353–373.

Lu, W., Liang, Y., 2008. Analysis of competing risks data with missing cause of failure under additive hazards model. Statist. Sinica 18, 219–234.

Patterson, B.J., Freedman, J., Blanchette, V., Sher, G., Pinkerton, P., Hannach, B., Meharchand, J., Lau, W., Boyce, N., Pinchefsky, E., Tasev, T., Pinchefsky, J., Poon, S., Shudman, L., Mack, P., Thomas, K., Blanchette, N., Greenspan, D., Panzarella, T., 2000. Effect of premedication guidelines and leukoreduction on the rate of febrile nonhaemolytic platelet transfusion. Transfus. Med. 10, 199–206.

Pollard, D., 1990. Empirical Processes: Theory and Applications. Institute of Mathematical Statistics, Hayward.

Schaubel, D.E., Cai, J., 2006a. Rate/mean regression for multiple-sequence recurrent event data with missing event category. Scand. J. Stat. 33, 191–207.

Schaubel, D.E., Cai, J., 2006b. Multiple imputation methods for recurrent event data with missing event category. Canad. J. Statist. 34, 677–692.

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

Sun, L., Zhu, L., Sun, J., 2009. Regression analysis of multivariate recurrent event data with time-varying covariate effects. J. Multivariate Anal. 100, 2214–2223.

Tsiatis, A.A., 2006. Semiparametric Theory and Missing Data. Springer, New York.

van der Vaart, A.W., 2000. Asymptotic Statistics. Cambridge: Melbourne.

van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes. Springer, New York.

Yin, G., Cai, J., 2004. Additive hazards model with multivariate failure time data. Biometrika 91, 801–818.

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

Zhao, X., Liu, L., Liu, Y., Xu, W., 2012. Analysis of multivariate recurrent event data with time-dependent covariates and informative censoring. Biom. J. 54, 585–599.

Zhu, L., Sun, J., Tong, X., Srivastava, D.K., 2010. Regression analysis of multivariate recurrent event data with a dependent terminal event. Lifetime Data Anal. 16, 478–490.