Lifetime Data Anal (2010) 16:478–490
DOI 10.1007/s10985-010-9158-9

Regression analysis of multivariate recurrent event data with a dependent terminal event

Liang Zhu · Jianguo Sun · Xingwei Tong · Deo Kumar Srivastava

Received: 6 May 2009 / Accepted: 25 February 2010 / Published online: 10 March 2010

Abstract Recurrent event data occur in many clinical and observational studies (Cook and Lawless, Analysis of recurrent event data, 2007) and in these situations, there may exist a terminal event such as death that is related to the recurrent event of interest (Ghosh and Lin, Biometrics 56:554–562, 2000; Wang et al., J Am Stat Assoc 96:1057–1065, 2001; Huang and Wang, J Am Stat Assoc 99:1153–1165, 2004; Ye et al., Biometrics 63:78–87, 2007). In addition, sometimes there may exist more than one type of recurrent events, that is, one faces multivariate recurrent event data with some dependent terminal event (Chen and Cook, Biostatistics 5:129–143, 2004). It is apparent that for the analysis of such data, one has to take into account the dependence both among different types of recurrent events and between the recurrent and terminal events. In this paper, we propose a joint modeling approach for regression analysis of the data and both finite and asymptotic properties of the resulting estimates of unknown parameters are established. The methodology is applied to a set of bivariate recurrent event data arising from a study of leukemia patients.

Keywords Joint modeling · Multivariate analysis · Regression analysis · Survival analysis

L. Zhu · D. K. Srivastava
Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, USA

J. Sun (✉)
Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65203, USA
e-mail: sunj@missouri.edu

X. Tong
School of Mathematical Sciences, Beijing Normal University, Beijing, People's Republic of China

1 Introduction

Recurrent event data usually refer to the data in which the event of interest can occur more than once and they arise in many fields such as clinical and longitudinal studies, reliability experiments and sociological studies (Andersen et al. 1993; Cook and Lawless 1996, 2007; Cook et al. 1996). Examples of such events include hospitalizations, infections, acute myocardial infections, and tumor metastases. Sometimes two or more different but related types of recurrent events may occur together and in this case one faces multivariate recurrent event data. Many authors have discussed the analysis of recurrent event data. For example, for univariate recurrent event data, Andersen and Gill (1982) and Prentice et al. (1981) developed some intensity-based methods, while Lawless and Nadeau (1995), Lin et al. (2000, 2001) proposed some marginal mean and rate-based approaches. The authors who investigated the analysis of multivariate recurrent event data include Cai and Schaubel (2004), Clegg et al. (1999), Schaubel and Cai (2006) and Spiekerman and Lin (1998). In particular, Cook and Lawless (2007) provided a comprehensive review of the existing literature about recurrent events.

For most of the methods mentioned above, they apply only to situations where the follow-up period is independent of the underlying recurrent process, which is the case if, for example, the follow-up stops at a prespecified study end time point. In some instances, however, this may not be true and the follow-up period may be determined by a terminal event that is dependent of the underlying recurrent process. A typical example is that the recurrent event is related to death, which would terminate the event process, such that the high recurrence rate of the events may indicate the increasing of the risk of death. For such situations, sometimes we say that there exists a dependent terminal event (Chen and Cook 2004; Ghosh and Lin 2002). It is apparent that one has to take this into account in the analysis.

In the presence of dependent terminal events, several approaches have been developed for the analysis of univariate recurrent event data. For example, Wang et al. (2001), Huang and Wang (2004) and Ye et al. (2007) developed frailty-based joint modeling procedures that model the recurrent event process and the terminal event process together. Ghosh and Lin (2000, 2002) also studied the same problem by relying on marginal models instead of the frailty model. Among others, Chen and Cook (2004) discussed multivariate recurrent event data with a dependent terminal event and considered the treatment comparison problem. In this article, we consider regression analysis of multivariate recurrent event data, for which there does not seem to exist an established method. For the situation, one has to not only consider the dependence among different types of recurrent events, but also deal with the association between recurrent and terminal events.

The remainder of the article is organized as follows. After introducing some notation and assumptions, we propose in Sect. 2 a joint modeling approach for regression analysis of multivariate recurrent event data in the presence of a dependent terminal event. In particular, an estimating equation is developed for estimation of unknown parameters and the asymptotic properties of the resulting estimates are established. The approach is similar to those given in Huang and Wang (2004) and Wang et al. (2001). Section 3 gives some results obtained from a simulation study conducted for

the assessment of the finite sample properties of the proposed estimates. In Sect. 4, we apply the proposed methodology to a set of bivariate recurrent event data arising from a study of leukemia patients and Sect. 5 contains some concluding remarks.

2 A joint modeling analysis approach

2.1 Notation and assumptions

Consider a recurrent event study that involves n independent subjects and each subject may experience K types of recurrent events. For subject i, let $N_{ik}(t)$ denote the number of the kth type event occurring up to time t and z_i a $p \times 1$ vector of time-independent covariates, $i = 1, \ldots, n$, $k = 1, \ldots, K$. In the following, we assume that there exists a terminal event that may be associated with $N_{ik}(t)$ and could occur for each subject. For convenience, we will assume that the event is death and use D_i to denote the death time. In practice, there also usually exists a censoring time and it will be denoted by C_i. Define $y_i = \min(D_i, C_i, \tau)$ and $\delta_i = I(D_i \le y_i)$, where τ is the stop time of the study. Then the observed data consist of $\{ y_i, \delta_i, z_i, N_{ik}(t) ; i = 1, \ldots, n, k = 1, \ldots, K \}$. The main goal is to make inferences about the underlying recurrent event processes.

For regression analysis, we will first assume that for subject i, there exists a nonnegative-valued latent variable v_{ik} with $E(v_{ik}|z_i) = 1$ and given v_{ik} and z_i, $N_{ik}(t)$ is a nonstationary Poisson process with intensity

$$\lambda_k(t; z_i, v_{ik}) = v_{ik}\lambda_{0k}(t)\, e^{\beta' z_i}, \quad 0 \le t \le \tau, k = 1, \cdots, K. \tag{1}$$

Here β is a $p \times 1$ vector of regression parameters, $\lambda_{0k}(t)$ is an unknown baseline intensity function. Also we will assume that $N_{ik}(\cdot)$, D_i and C_i are mutually independent given v_{ik} and z_i. In terms of D_i, it will be supposed that its distribution depends on v_{ik} and z_i in an arbitrary way and more comments on this will be given below. Note that model (1) with or without v_{ik} has been used by many authors in modeling recurrent event data. For example, a number of inference procedures have been developed for regression analysis of recurrent event data under model (1) without the latent variable (Andersen et al. 1993; Cook and Lawless 2007).

2.2 Estimation of unknown parameters

For inference about model (1), let m_{ik} denote the total number of the kth type recurrent events observed before y_i for subject i, $i = 1, \ldots, n$. As pointed out in Wang et al. (2001), for subject i and conditional on $(z_i, y_i, v_{ik}, m_{ik})$, the observed kth type recurrent event times, say $t_{ik1}, \ldots, t_{ikm_{ik}}$, are the order statistics of a set of i.i.d. random variables with the density function

$$f_k(t) = \frac{v_{ik}\lambda_{0k}(t)e^{\beta' z_i}}{v_{ik}\Lambda_{0k}(\tau)e^{\beta' z_i}} = \frac{\lambda_{0k}(t)}{\Lambda_{0k}(\tau)}, \quad 0 \le t \le \tau$$

assuming that $v_{ik} > 0$, where $\Lambda_{0k}(t) = \int_0^t \lambda_{0k}(s)ds$. It follows that the conditional likelihood function of the observed data given $(z_i, y_i, v_{ik}, m_{ik})$ is proportional to

$$L_c = \prod_{i=1}^{n} \prod_{k=1}^{K} \prod_{j=1}^{m_{ik}} \frac{f_k(t_{ikj})}{F_k(y_i)},$$

which is independent of the unobserved latent variable v_{ik}, where F_k denotes the cumulative distribution function corresponding to f_k. The maximization of the conditional likelihood above gives the product-limit estimate

$$\hat{F}_k(t) = \prod_{s_{k(l)} > t} \left(1 - \frac{d_{k(l)}}{R_{k(l)}}\right)$$

for $F_k(t)$. Here for each k, the $s_{k(l)}$'s are the ordered and distinct values of the kth type event times t_{ikj}'s, $d_{k(l)}$ denotes the number of the kth type events occurring at $s_{k(l)}$, and $R_{k(l)}$ denotes the total number of the kth type events with the event time and the terminating time satisfying $\{t_{ikj} \le s_{k(l)} \le y_i\}$.

Define $\Lambda_0(\tau) = (\Lambda_{01}(\tau), \ldots, \Lambda_{0K}(\tau))'$ and for $k = 1, \ldots, K$, let e_k denote the K-dimensional vector of zeros except its kth entry equal to one. Also let $\bar{z}_{ik} = (e_k', z_i')'$ and $\gamma = (\ln(\Lambda_0(\tau))', \beta')'$. To estimate γ, notice that $E[m_{ik}|z_i, y_i, v_{ik}] = v_{ik} e^{\beta' z_i} \Lambda_{0k}(y_i)$, which gives

$$\begin{aligned} &E\left[m_{ik}\Lambda_{0k}^{-1}(y_i)|z_i, y_i\right] \\ &= E\left[E[m_{ik}|z_i, y_i, v_{ik}]\Lambda_{0k}^{-1}(y_i)|z_i, y_i\right] = e^{\beta' z_i}\Lambda_{0k}(\tau) = e^{\gamma'\bar{z}_{ik}}. \end{aligned}$$

This suggests a natural class of unbiased estimating equations

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\bar{z}_{ik}\left\{m_{ik}F_k^{-1}(y_i) - e^{\gamma'\bar{z}_{ik}}\right\} = 0 \tag{2}$$

for estimation of γ.

Let $\hat{\gamma}$ denote the estimate of γ given by the solution to the Eq. 2 and $\hat{\beta}$ and $\hat{\Lambda}_0(\tau)$ the resulting estimates of β and $\Lambda_0(\tau)$, respectively. Define

$$\hat{Q}_k(u) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_{ik}} I(t_{ikj} \le u)$$

and

$$\hat{R}_k(u) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_{ik}} I(t_{ikj} \le u \le y_i)$$

for $k = 1, \ldots, K$. We show in the Appendix that under some regularity conditions and as $n \to \infty$, we have

$$\sqrt{n}\left(\hat{\gamma} - \gamma_0\right) \xrightarrow{d} N(0, A),$$

where γ_0 denotes the true value of γ and the covariance matrix A is defined in the Appendix and can be consistently estimated by $\hat{A} = \hat{\psi}^{-1}\hat{\Sigma}\left(\hat{\psi}'\right)^{-1}$. Here

$$\hat{\psi} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\bar{z}_{ik}\bar{z}'_{ik}e^{\hat{\gamma}'\bar{z}_{ik}}$$

and

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left[\sum_{k=1}^{K}\hat{e}_{ik}\right]\left[\sum_{k=1}^{K}\hat{e}_{ik}\right]'\right\}$$

with

$$\hat{e}_{ik} = \left[-\frac{1}{n}\sum_{j=1}^{n}\frac{\bar{z}_{jk}m_{jk}\hat{b}_{ik}(y_j)}{\hat{F}_k(y_j)} + \bar{z}_{ik}\left\{\frac{m_{jk}}{\hat{F}_k(y_i)} - e^{\hat{\gamma}'\bar{z}_{ik}}\right\}\right]$$

and

$$\hat{b}_{ik}(t) = \sum_{j=1}^{m_{ik}}\left\{\int_t^{\tau}\frac{I(t_{ikj} \le u \le y_i)d\hat{Q}_k(u))}{\hat{R}_k^2(u)} - \frac{I(t < t_{ikj} \le \tau)}{\hat{R}_k(t_{ikj})}\right\}.$$

for $t \in [0, \tau]$. Thus it follows that one can approximate the distribution of $\sqrt{n}\left(\hat{\beta} - \beta_0\right)$ by the multivariate normal distribution with mean zero and the covariance matrix that can be consistently estimated by the lower-right $p \times p$ block of $\hat{A}$ given above, where β_0 denotes the true value of β.

In addition to estimation of β, sometimes one may be also interested in estimating $\Lambda_{0k}(t)$, which represents the baseline mean function of the underlying recurrent event process under the Poisson assumption. For this, a natural estimate is given by

$$\hat{\Lambda}_{0k}(t) = \hat{F}_k(t)\,\hat{\Lambda}_{0k}(\tau). \tag{3}$$

By following Wang et al. (2001), one can show that for a given t, $\sqrt{n}\left\{\hat{\Lambda}_{0k}(t) - \Lambda_{0k}(t)\right\}$ converges weakly to the normal distribution with mean zero and the variance that can be estimated by

$$\hat{\Lambda}_{0k}^2(t)\left\{\hat{g}_{ik} + \hat{b}_{ik}(t)\right\}^2, \tag{4}$$

where $\hat{g}_{ik}$ denotes the kth element of the vector

$$\left\{ \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{K} \bar{z}_{jk} \bar{z}'_{jk} e^{\hat{\gamma}' \bar{z}_{jk}} \right\}^{-1} e_i .$$

Note that model (1) assumes that covariates and their effects are the same for different types of recurrent events and there is no type-specific covariates without loss of generality. If there exist type-specific covariates and/or covariate effects may differ from type to type, one can still apply the methodology developed above by defining a large and new vector of covariates. For example, in the case that one prefers model (1) with β and z_i replaced by β_k and z_{ik}, respectively, we can define new $\beta^* = (\beta_1', \ldots, \beta_K')'$ and $z_i^* = (0, \ldots, z_{ik}(\cdot)', \ldots, 0)'$. For estimation of β^* and thus the β_k's, one can apply the approach given above under model (1) with β and z_i replaced by β^* and z_i^*, respectively.

3 A simulation study

A simulation study was performed to assess the finite-sample performance of the estimation procedure proposed in the previous section. In the study, we considered situations where there exist two types of recurrent events ($K = 2$) and assumed that z_i took value 0 or 1 with probability 0.5. The frailty variables v_{i1} and v_{i2} were generated from the bivariate gamma distribution with both mean and variance equal to 1 and their correlation being 0.5. For the death time, it was supposed that given z_i, v_{i1} and v_{i2}, the hazard function of D_i has the form

$$\lambda^D(t; z_i, v_{i1}, v_{i2}) = \frac{(v_{i1} + v_{i2})\, e^{\alpha' z_i} t}{60}.$$

Furthermore, we generated the censoring time C_i from the uniform distribution over $(0, 15)$ and took $\tau = 10$.

For the generation of the recurrent event process $N_{ik}(t)$, we considered two situations. For the first case, which will be referred to the Poisson case, we assumed that given v_{i1} and v_{i2}, $N_{ik}(t)$ was the nonstationary Poisson processes with the intensity given by model (1) and took $\lambda_{01}(t) = 1.5/(t+1)$ and $\lambda_{02}(t) = 1/2$. For the second case, which will be referred to as the Non-Poisson case, we first generated the number of observed events m_{ik} by taking $m_{ik} = \left[v_{ik}\, e^{\beta' z_i}\, \mu_{0k}(y_i)\, \epsilon_i \right]$ with $\mu_{01}(t) = 3t/4$, $\mu_{02}(t) = t/2$ and $\epsilon_i \sim \Gamma(4, 1/4)$. Given m_{ik}, the recurrent event times $t_{ik1}, \ldots, t_{ikm_{ik}}$ were generated as the order statistics of m_{ik} random variables from the uniform distribution over $(0, y_i)$. The results given below are based on 1000 replications.

Tables 1 and 2 present the results obtained for estimation of β with the true value being 0, 0.5 or 1.0 and the sample size $n = 100$ or 200. The former corresponds to the Poisson case for $N_{ik}(t)$ and the latter is for the Non-Poisson case. The results include the estimated bias of the estimate given by the average of the point estimates

Table 1 Estimation of β for the Poisson case

			$\alpha = 0.5$					$\alpha = 2$			
n	β	BIAS*	BIAS	SSD	ESE	CP	BIAS*	BIAS	SSD	ESE	CP
100	0.0	−.045	.008	.234	.227	.941	−.191	.009	.255	.244	.950
	0.5	−.047	−.008	.218	.217	.944	−.205	.006	.244	.228	.934
	1.0	−.045	.004	.219	.208	.934	−.214	.006	.231	.218	.924
200	0.0	−.043	−.002	.160	.161	.958	−.197	.003	.186	.173	.941
	0.5	−.050	.007	.157	.153	.950	−.207	.002	.165	.163	.959
	1.0	−.048	.005	.145	.147	.957	−.224	−.001	.154	.154	.947

BIAS* represents the bias given by the approach of Cai and Schaubel (2004)

Table 2 Estimation of β for the Non-Poisson case

			$\alpha = 0.5$					$\alpha = 2$			
n	β	BIAS*	BIAS	SSD	ESE	CP	BIAS*	BIAS	SSD	ESE	CP
100	0.0	−.075	.004	.217	.214	.946	−.277	−.019	.233	.227	.941
	0.5	−.071	.030	.210	.211	.953	−.275	.014	.231	.220	.937
	1.0	−.066	.029	.220	.208	.935	−.276	.030	.222	.214	.930
200	0.0	−.066	.005	.155	.153	.947	−.273	−.019	.160	.161	.955
	0.5	−.060	.016	.148	.150	.954	−.274	.007	.153	.154	.954
	1.0	−.058	.030	.149	.149	.945	−.289	.021	.154	.152	.941

BIAS* represents the bias given by the approach of Cai and Schaubel (2004)

minus the true value (BIAS), the sample standard deviation of the estimates (SSD), the average of the estimated standard errors (ESE), and the 95% empirical coverage probability (CP). For comparison, we also applied the approach of Cai and Schaubel (2004), which ignored the dependent death time, to the simulated data, and calculated and included them in the tables the bias of the obtained estimates of β (BIAS*). These results suggest that the proposed estimate $\hat{\beta}$ seems to be unbiased and the given variance estimate seems to be reasonable as it is close the the sample variance. Also as expected, the approach that ignores the dependent death time tends to give biased estimates of the covariate effect.

In the simulation study, we also investigated the performance of the estimate of $\Lambda_{0k}(t)$ given in (3). Figure 1 presents the obtained results corresponding to the cases considered in Table 1 with $\alpha = 0.5$. Specifically, we plotted the true $\Lambda_{0k}(t)$ and the 95% pointwise confidence bands based on the sample standard deviation (solid lines for both). It also includes the averages of the estimates $\hat{\Lambda}_{0k}(t)$ and the 95% pointwise confidence bands based on the averages of the estimated standard deviations given in (4) (dashed lines for both). It can be seen that both the estimate (3) and the variance estimate (4) seem to perform well. Simulation studies of different parameter configurations yielded evidence of similar good frequency properties of the estimates.

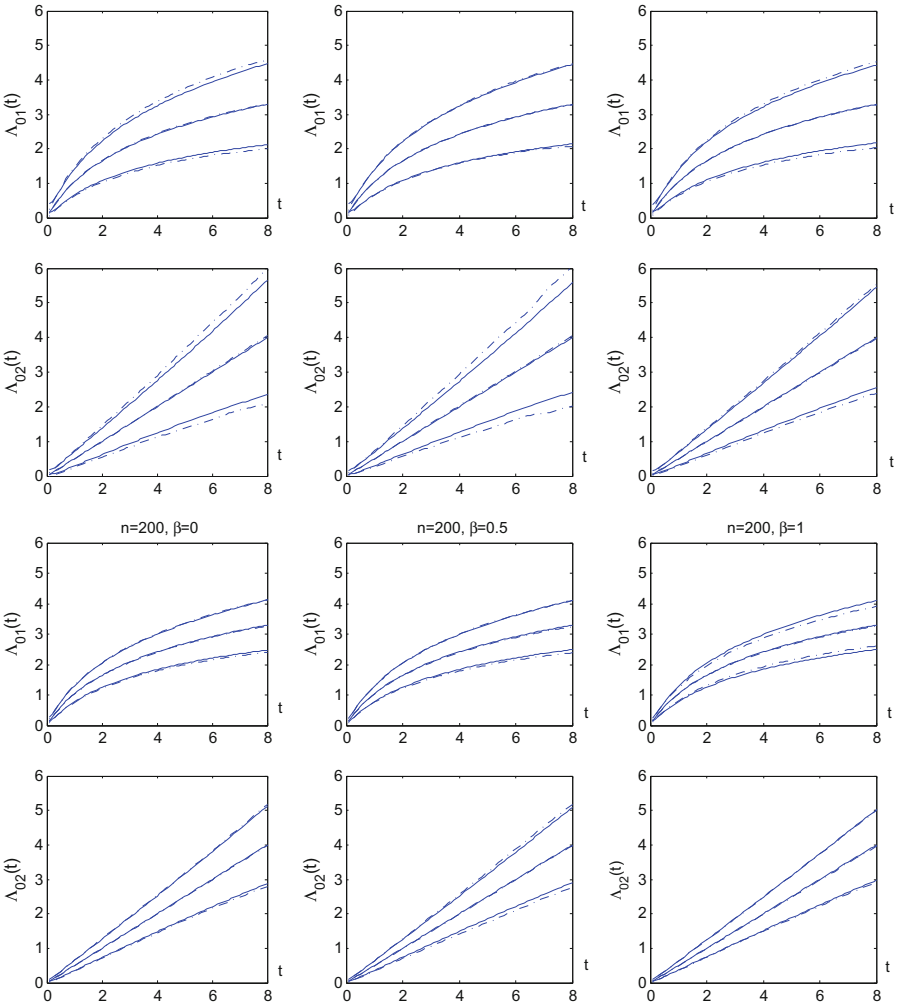

Fig. 1 Estimates of $\Lambda_{01}(t)$ and $\Lambda_{02}(t)$ with the pointwise 95% confidence bands. *Solid lines* are true $\Lambda_{01}(t)$ and $\Lambda_{02}(t)$ and sample confidence bands; *Dashed lines* are estimated functions and confidence bands

4 An application

Now we apply the methodology developed in the previous sections to a set of recurrent event data arising from an ongoing clinical trial on the patients with acute myeloid leukemia (AML). The data set consists of 201 subjects who were consecutively treated for their AML from October 2002 through October 2008. During the chemotherapy course, the patients may experience repeated bacterial, viral or fungal infections, some recurrent events, and one of the objectives of the study was to investigate these infection rates and their relationship with various predictor variables. In the analysis below, we will classify all infections into two types, the bacterial infection,

Table 3 Frequencies of the two types of infections in the AML study

Type of infection	0	1	2	≥ 3
Bacterial	55	62	35	49
Fungal or viral	139	48	10	4

Table 4 Estimates of covariate effects for the AML study

	Cai and Schaubel's method			Proposed method		
Covariate	Estimate	SE	p value	Estimate	SE	p value
Gender	−0.096	0.114	0.398	−0.061	0.120	0.615
Risk	−0.072	0.082	0.381	−0.106	0.079	0.179
Dose	0.061	0.118	0.607	0.098	0.123	0.427
Race	0.026	0.124	0.836	0.026	0.132	0.844
WBC	−0.0006	0.0007	0.418	−0.0009	0.0008	0.241
Age	−0.021	0.010	0.045	−0.020	0.012	0.090

which is very common for leukemia patients, and the fungal or viral infection. Table 3 provides a summary of the frequency of these two recurrent events among the 201 patients. During the study, 38 patients experienced the relapse, transplant or death, which will be treated as the terminal event in the following. The median follow up time is 160 days.

For the study, there are six predictor variables or baseline covariates of interest. They are gender (male = 0; female = 1), the leukemia risk level (low = 0; standard = 1; high = 2), the dose of cytarabine given for the first course of chemotherapy (standard = 0; high = 1), and race (white = 0; others = 1) along with the white blood count (WBC) and the age both at the diagnosis of AML. Table 4 presents the analysis results obtained by applying the estimation procedure proposed above to the data here. They include the estimated effect of each covariate on the occurrence rate of the two types of infections, the estimated standard error (SE), and the p-value for testing no significant covariate effect. These results indicate that none of the covariates seems to have significant effect on the infection rate. For comparison, we also applied the method given in Cai and Schaubel (2004), which does not take into account the dependent terminal event, to the data and included the results also in Table 4. It can be seen that it gave similar conclusions to those obtained by the proposed approach for most of covariates except the age and suggested that the infection rate may be negatively related to the patient age.

To compare the infection rates between the two types of infections, we calculated and presented in Fig. 2 the estimates of the cumulative baseline intensities given in (3) along with the pointwise 95% confidence bands based on the estimate given in (4). It indicated that the bacterial infections seem to have significantly higher occurrence rate than the fungal or viral infections.

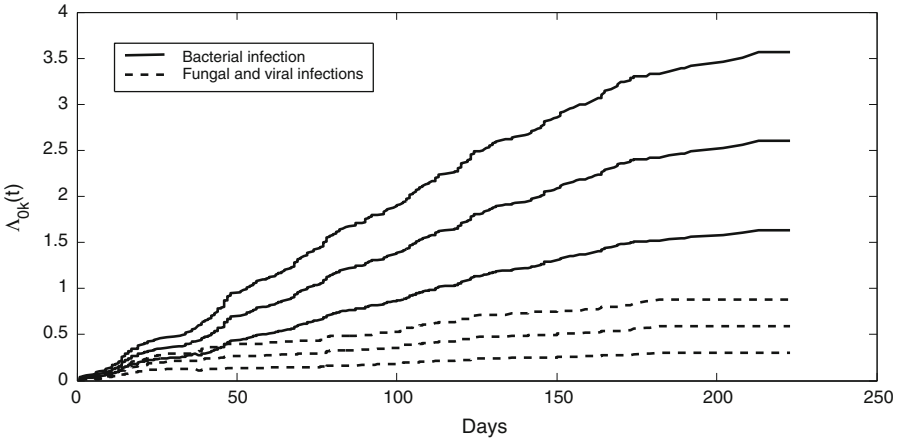

Fig. 2 Estimates of $\Lambda_{01}(t)$ and $\Lambda_{02}(t)$ with the pointwise 95% confidence bands

5 Concluding remarks

In the preceding sections, we presented a joint modeling approach for regression analysis of multivariate recurrent event data in the presences of a dependent terminal event. More specifically, we investigated situations where the relationship between the underlying recurrent event process and the death or terminal process can be described by some latent variables. The joint modeling approach proposed can be seen as a generalization of the methods given by Wang et al. (2001) and Huang and Wang (2004) for univariate recurrent event data. The simulation study indicated that the approach works reasonably well for practical situations.

One advantage of the joint modeling approach proposed above is that it leaves the structures about both the relationship between the recurrent event and the terminal event and the effect of covariates on the terminal event arbitrary. Also it is worth noting that although the Poisson process assumption was employed in the development of the estimation procedure, the asymptotic properties of the resulting estimates $\hat{\gamma}$ do not depend on the assumption. In other words, the inference procedure is still valid without the assumption and has the robustness property. A limitation of the approach developed above is that it only applies to situations where covariates are time-independent. For the case with time-dependent covariates, some new estimation procedures need to be developed.

In the discussion above, it has been assumed that one is only interested in the effect of covariates on recurrent event processes and the estimation procedure presented does not require the specification of the form of covariate effects on the terminal event. That is, the covariate effects on the terminal event were left arbitrary. On the other hand, sometimes one may be also interested in directly modeling and estimating covariate effects on the terminal event or death time. In these situations, of course, one needs first to specify a regression model such as the proportional hazards model

$$\lambda^D(t; z_i, v_i) = v_i \lambda_0^D(t)\, e^{\alpha' z_i} \tag{5}$$

for the death time D, where $v_i = \frac{1}{K}\sum_{k=1}^{K} v_{ik}$, $\lambda_0^D(t)$ is an unknown baseline hazard function, and α denotes the regression parameter. Under model (5) and given $\hat{\beta}$ and $\hat{\Lambda}_{0k}$, by following the partial likelihood approach, one can naturally estimate α by using the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n}\delta_i\left\{z_i - \frac{\sum_{j=1}^{n} z_j\hat{v}_j e^{\alpha' z_j} I(y_j \geq y_i)}{\sum_{j=1}^{n}\hat{v}_j e^{\alpha' z_j} I(y_j \geq y_i)}\right\} = 0,$$

where

$$\hat{v}_i = \frac{1}{K}\sum_{k=1}^{K}\frac{m_{ik}}{e^{\hat{\beta}' z_i}\,\hat{\Lambda}_{0k}(y_i)}.$$

Let $\hat{\alpha}$ denote the estimate of α defined above. One can further estimate the baseline cumulative hazard function $\Lambda_0^D(t) = \int_0^t \lambda_0^D(s)\,ds$ by

$$\hat{\Lambda}_0^D(t) = \int_0^t \frac{d\sum_{i=1}^{n}\delta_i I(y_i \leq s)}{\sum_{i=1}^{n}\hat{v}_i\, e^{\hat{\alpha}' z_i} I(y_i \geq s)}.$$

Acknowledgments The authors wish to thank Dr. Rubnitz for sharing the AML data used in the paper. Also they want to thank the Guest Editors, Drs. Cook and Andersen, and a reviewer for their many helpful comments and suggestions.

Appendix: Asymptotic normality of $\hat{\gamma}$

In this appendix, we will sketch the proof of the asymptotic normality of the estimate $\hat{\gamma}$. To see this, note that the left side of the estimating Eq. 2 can be expressed as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\bar{z}_{ik}\left\{\frac{m_{ik}}{\hat{F}_k(y_i)} - \frac{m_{ik}}{F_k(y_i)}\right\} + \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\bar{z}_{ik}\left\{\frac{m_{ik}}{F_k(y_i)} - e^{\gamma'\bar{z}_{ik}}\right\}\\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left[E\left\{-\frac{\bar{z}_k m_k b_{ik}(y)}{F_k(y)}\right\} + \bar{z}_{ik}\left\{\frac{m_{ik}}{F_k(y_i)} - e^{\gamma'\bar{z}_{ik}}\right\}\right] + o_p\left(n^{-1/2}\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} e_{ik} + o_p\left(n^{-1/2}\right).
\end{aligned}$$

In the above,

$$e_{ik} = E\left\{-\frac{\bar{z}_k m_k b_{ik}(y)}{F_k(y)}\right\} + \bar{z}_{ik}\left\{\frac{m_{ik}}{F_k(y_i)} - e^{\gamma'\bar{z}_{ik}}\right\}$$

and

$$b_{ik}(t) = \sum_{j=1}^{m_{ik}} \left\{ \int_t^\tau \frac{I\left(t_{ikj} \le u \le y_i\right) dQ_k(u)}{R_k^2(u)} - \frac{I\left(t < t_{ikj} \le \tau\right)}{R_k(t_{ikj})} \right\}$$

with $G_k(u) = E\{v_k I(y \ge u)\}$, $Q_k(u) = \int_0^u G_k(v) d\Lambda_{0k}(v)$ and $R_k(u) = G_k(u)\,\Lambda_{0k}(u)$.

Then by using the Taylor series expansion, we can obtain that

$$\sqrt{n}\left(\hat{\gamma} - \gamma_0\right) = n^{-1/2}\,\psi^{-1} \sum_{i=1}^{n} e_i + o_p(1),$$

where $e_i = \sum_{k=1}^{K} e_{ik}$ and $\psi = E\left[-\sum_{k=1}^{K} \partial e_{ik}/\partial\gamma\right]$. It thus follows from the central limit theorem that $\sqrt{n}\,(\hat{\gamma} - \gamma_0)$ converges weakly to the multivariate normal distribution with mean 0 and the variance-covariance matrix $A = \psi^{-1}\,\Sigma\,(\psi')^{-1}$, where Σ represents the variance-covariance matrix of the e_i's.

References

Andersen PK, Borgan O, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer-Verlag, New York

Andersen PK, Gill RD (1982) Cox's regression model for counting processes: a large sample study. Ann Stat 10:1100–1120

Cai J, Schaubel DE (2004) Marginal means/rates models for multiple type recurrent event data. Lifetime Data Anal 10:121–138

Chen BE, Cook RJ (2004) Tests for multivariate recurrent events in the presence of a terminal event. Biostatistics 5:129–143

Clegg LX, Cai J, Sen PK (1999) A marginal mixed baseline hazards model for multivariate failure time data. Biometrics 55:805–812

Cook RJ, Lawless JF (1996) Interim monitoring of longitudinal comparative studies with recurrent event responses. Biometrics 52:1311–1323

Cook RJ, Lawless JF (2007) Analysis of recurrent event data. Springer-Verlag, New York

Cook RJ, Lawless JF, Nadeau JC (1996) Robust tests for treatment comparisons based on recurrent event responses. Biometrics 52:557–571

Ghosh D, Lin DY (2000) Nonparametric analysis of recurrent events and death. Biometrics 56:554–562

Ghosh D, Lin DY (2002) Marginal regression models for recurrent and terminal events. Statistica Sinica 12:663–688

Huang CY, Wang MC (2004) Joint modeling and estimation for recurrent event processes and failure time data. J Am Stat Assoc 99:1153–1165

Lawless JF, Nadeau C (1995) Some simple robust methods for the analysis of recurrent events. Technometrics 37:158–168

Lin DY, Wei LJ, Yang I, Ying Z (2000) Semiparametric regression for the mean and rate function of recurrent events. J R Stat Soc, Series B 69:711–730

Lin DY, Wei LJ, Ying Z (2001) Semiparametric transformation models for point processes. J Am Stat Assoc 96:620–628

Prentice RL, Williams BJ, Peterson AV (1981) On the regression analysis of multivariate failure time data. Biometrika 68:373–379

Schaubel DE, Cai J (2006) Rate/mean regression for multiple-sequence recurrent event data with missing event category. Scand J Stat 33:191–207

Spiekerman CF, Lin DY (1998) Marginal regression models for multivariate failure time data. J Am Stat Assoc 93:1164–1175

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

Ye Y, Kalbfleisch JD, Schaubel DE (2007) Semiparametric analysis of correlated recurrent and terminal events. Biometrics 63:78–87