Published in final edited form as:

J Biopharm Stat. 2014 ; 24(2): 429–442. doi:10.1080/10543406.2013.860159.

Multivariate Recurrent Events in the Presence of Multivariate Informative Censoring with Applications to Bleeding and Transfusion Events in Myelodysplastic Syndrome

Donglin Zeng,

Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599

Joseph G. Ibrahim,

Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599

Ming-Hui Chen,

Department of Statistics, University of Connecticut, U-4120, Storrs, CT 06269

Kuolung Hu, and

Global Biostatistical Science, Amgen Inc., One Amgen Center Drive, Thousand Oaks, CA 91320

Catherine Jia

Global Biostatistical Science, Amgen Inc., One Amgen Center Drive, Thousand Oaks, CA 91320

Summary

We propose a general novel class of joint models to analyze recurrent events that has a wide variety of applications. The application of focus on this paper is to model the bleeding and transfusion events in Myelodysplastic Syndrome (MDS) studies, where patients may die or withdraw from the study early due to adverse events or other reasons, such as consent withdrawal or required alternative therapy during the study. The proposed model accommodates multiple recurrent events and multivariate informative censoring through a shared random effects model. The random effects model captures both within-subject and within-event dependence simultaneously. We construct the likelihood function for the semi-parametric joint model and develop an EM algorithm for inference. The computational burden does not increase with the number of types of recurrent events. We utilize the MDS clinical trial data to illustrate our proposed methodology. We also conduct a number of simulations to examine the performance of the proposed model.

Keywords

Bleeding events; EM algorithm; Semiparametric joint model; Simulation; Transfusions

1 Introduction

In many biomedical studies, patients may experience the same type of recurrent event repeatedly over time, such as bleeding, multiple infections and relapse. In such studies, the patients may also be subject to informative censoring, which can be death or administrative withdrawal due to adverse events. In our motivating case study of MDS, the interest lies in evaluating a treatment effect of an investigational product (IP) for reducing recurrent bleeding or transfusion events among lower risk MDS patients receiving disease modifying therapy. In addition, we also consider early subject discontinuation due to disease progression to acute myeloid leukemia (AML), death, adverse events, switching to alternative therapies, or other reasons.

A major motivation for our proposed methodology is that we have multivariate types of recurrent events as well as multivariate informative censoring. Since the censoring times are associated with the recurrent events, existing methods on recurrent events (e.g., Andersen and Gill, 1982; Prentice, Williams and Peterson, 1981; and Wei, Lin and Weissfeld, 1989) that are valid under noninformative censoring, may yield misleading results if applied in our setting, for example. Therefore, valid procedures which jointly take the recurrent events and informative censoring times into account in the analysis need to be developed. Marginal models have been proposed to analyze recurrent event data in the presence of a single terminal event (a univariate informative censoring time); see Cook and Lawless (1997), Ghosh and Lin (2000; 2002), and Chen and Cook (2004). To capture the recurrence given an individual's past event history, a more attractive approach is to adopt a joint modeling approach for both recurrent events and informative censoring times. Approaches along these lines include the shared frailty model by Wang, Qin and Chiang (2001), Huang and Wang (2004), Liu, Wolfe and Huang (2004), and Liu and Huang (2008), which model the recurrent event and informative censoring time separately but allow a common frailty shared by these two models. More general classes of transformation models and their theoretical properties have been more recently considered by Zeng and Lin (2009) and Zhu et al. (2011).

Recently, Zhu et al. (2010) generalize the above approaches to model multivariate recurrent events and one terminal event but their method relies on a very special frailty structure in recurrent events. Later, Zhu et al. (2011) consider a more general frailty structure and the frailties from each type of recurrent event are included into the model for the terminal event; the nonparametric maximum likelihood approach is used for inference. However, one challenging problem in implementing this approach is that one has to deal with high-dimensional numerical integrations whose dimensions increase with the number of recurrent events. This problem will become even more burdensome if dependent censoring time is also present in addition to the terminal event, as seen in our MDS application. Thus, in this paper, we present a joint model for handling multiple recurrent events and multiple informative censoring times. Specifically, we allow an event-specific frailty to account for the within-subject-within-event dependence while adopting another common frailty to be shared by all recurrent events and informative censoring times. Using this structure, our inference for the model parameters avoids high-dimensional numerical integration, which can be inaccurate when the number of event types is moderate to large. We propose a semi-parametric model, use a nonparametric maximum likelihood estimation (NPMLE) approach for inference, and develop a simple EM-algorithm for estimation of the NPMLE's.

The rest of the paper is organized as follows. In Section 2, we propose the joint models for multiple recurrent events in the presence of multiple types of informative censoring, discuss the inference procedure, and examine the asymptotic properties. In Section 3, we carry out an extensive simulation study to examine the empirical performance of the proposed joint model and inference procedure. In Section 4, we present a detailed analysis of a real dataset from a phase 2, randomized, double blind, placebo controlled MDS clinical trial. We conclude the paper with a brief discussion in Section 5.

2 Proposed Methodology

2.1 Joint models

We use joint models to analyze multiple recurrent events subject to multiple types of informative censoring. In these joint models, the subject-specific random effects will be used to represent the latent dependence between the recurrent events and the informative censoring times. Specifically, for recurrent events of type l , $l = 1, \dots, L$, we propose the following intensity model:

$$dA_{il}(t)=dA_l(t)\exp\{X_i(t)^T\beta_l+W_i(t)^Tb_{il}+Z_i(t)^T\xi_i\},$$

where $A_{il}(t)$ is the intensity function for subject i , $X_i(t)$ denotes external covariates associated with subject i , $Z_i(t)$ includes a constant term and is part of $X_i(t)$, $W_i(t)$ includes constant term and part of $X_i(t)$, $A_l(t)$ is an unknown baseline cumulative intensity function, b_{il} is a subject and event-specific random effect, and ξ_i is a subject random effect shared by all the events. Additionally, we assume that for each l , the b_{il} , $l = 1, \dots, L$ are iid $N(0, \Sigma_l)$ and the ξ_i 's are iid $N(0, \Psi)$ and independent of the b_{il} 's. For an informative censoring time of type k , $k = 1, \dots, K$, we propose the following proportional hazards model:

$$d\Lambda_{ik}(t)=d\Lambda_k(t)\exp\{\tilde{X}_i(t)^T\gamma_k+(\phi_k\circ\tilde{Z}_i(t))^T\xi_i\},$$

where $\lambda_{ik}(t) = d\Lambda_{ik}(t)/dt$ is the hazard function for the type k informative censoring time for subject i , $\lambda_k(t)$ is an unknown baseline hazard function and both γ_k and ϕ_k are unknown parameters. Here, $\phi_k \circ Z_i(t)$ denotes the component-wise product. Thus, the proposed joint model can simultaneously estimate the covariate effects on the recurrent events and informative censoring times. The random effect ξ_i captures the latent dependence between recurrent events and survival times, where the dependence is reflected in the parameter ϕ_k . Specifically, if the variance of ξ_i is zero, then the dependence among the recurrent events is completely due to the covariates; if $\phi_k = 0$, then the dependence between the recurrent events and the survival event k is captured by the covariates, while $\phi_k > 0$ or $\phi_k < 0$ reflects the possible positive correlation or negative correlation respectively, due to some unobserved factors.

In our MDS application, we will specifically choose b_{il} to be univariate, $X_i(t) = \tilde{X}_i(t)$, $W_i(t) = 1$, and $Z_i(t) = \tilde{Z}_i(t) = 1$. Moreover, in the MDS application, L denotes the number of types of recurrent events (bleeding and transfusion), i.e., $L = 2$, and the terminal event is death or AML, so $K = 1$. Thus, a 2×2 covariance matrix is assumed between the bleeding and transfusion events and ξ_i is the common random effect shared by the bleeding and transfusion events and the terminal event. In general, when $L > 2$, a compound symmetry covariance structure is assumed between recurrent events. We may further assume $\beta_1 = \dots = \beta_L$. Under these assumptions, the above model becomes

$$dA_{il}(t)=dA_l(t)\exp\{X_i(t)^T\beta+b_{il}+\xi_i\}, l=1, \dots, L,$$

and

$$d\Lambda_{ik}(t)=d\Lambda_k(t)\exp\{\tilde{X}_i(t)^T\gamma_k+\phi_k\xi_i\}, k=1, \dots, K.$$

2.2 Observed-data likelihood

The observed data consists of the recurrent events and survival endpoints:

$$(N_{il}(t \wedge Y_i), X_i(t), Y_i=\min(T_{i1}, \dots, T_{iK}, C_i), \Delta_i=(\Delta_{i1}, \dots, \Delta_{iK})),$$

where C_i is the noninformative censoring time (for example, administrative censoring) and Δ_{ik} is an indicator variable $I(T_{ik} = Y_i)$, that is, the observed endpoint is the k^{th} survival event.

For this data structure, and under our model assumptions, the observed-data likelihood function is proportional to

$$\prod_{i=1}^n \int \prod_{l=1}^L \int_{b_{il}} \left(\prod_t \left\{ A'_l(t) e^{X_i(t)^T \beta_l + W_i(t)^T b_{il} + Z_i(t)^T \xi_i} \right\}^{\delta N_{il}(t)} \right) \\ \times \exp \left\{ - \int_0^{Y_i} e^{X_i(t)^T \beta_l + W_i(t)^T b_{il} + Z_i(t)^T \xi_i} dA_l(t) \right\} \left[\prod_{k=1}^K \left\{ \Lambda'_k(Y_i) e^{\tilde{X}_i(Y_i)^T \gamma_k + (\phi_k \circ \tilde{Z}_i(Y_i))^T \xi_i} \right\}^{\Delta_{ik}} \right] \times \exp \left\{ - \int_0^{Y_i} e^{\tilde{X}_i(t)^T \gamma_k + i(\phi_k)} \right\}$$

where $\delta N_{il}(t)$ denotes the jumps of N_{il} at time t .

2.3 Inference procedure

The estimation and inference for all the parameters can be carried out by nonparametric maximum likelihood estimation. The EM algorithm is used to compute the maximum likelihood estimates treating b_i as missing data. In the M-step of the EM algorithm, we maximize the conditional expectation of the complete-data log-likelihood function given the observed data and the current parameter values. Note that the complete-data log-likelihood takes a similar form as the usual Cox partial log-likelihood function. Thus, the maximization for the regression parameters is equivalent to maximizing some modified expression of the Cox partial likelihood function while the estimation for the baseline cumulative hazard function is a Breslow-type estimator. Specifically, we solve the following estimating equations for updating the regression coefficients: for $l = 1, \dots, L$,

$$\sum_{i=1}^n \int \left[X_i(t) - \frac{\sum_{j=1}^n I(Y_j \geq t) X_j(t) \hat{E}[e^{X_j(t)^T \beta_l + W_j(t)^T b_{jl} + Z_j(t)^T \xi_j}]}{\sum_{j=1}^n I(Y_j \geq t) \hat{E}[e^{X_j(t)^T \beta_l + W_j(t)^T b_{jl} + Z_j(t)^T \xi_j}]} \right] dN_{il}(t) = 0;$$

for $k = 1, \dots, K$,

$$\sum_{i=1}^n \Delta_{ik} \left[\left(\frac{\tilde{X}_i(Y_i)}{\tilde{Z}_i(Y_i) \hat{E}[\xi_i]} \right) - \frac{\sum_{Y_j \geq Y_i} \hat{E} \left[\left(\frac{\tilde{X}_j(Y_i)}{\tilde{Z}_j(Y_i) \xi_j} \right) e^{\tilde{X}_j(Y_i)^T \gamma_k + (\phi_k \circ \tilde{Z}_j(Y_i))^T \xi_j} \right]}{\sum_{Y_j \geq Y_i} \hat{E}[e^{\tilde{X}_j(Y_i)^T \gamma_k + (\phi_k \circ \tilde{Z}_j(Y_i))^T \xi_j}]} \right] = 0.$$

The one-step Newton-Raphson iteration can be used to update the regression coefficients. The baseline functions A_l and Λ_k can be updated as

$$A_l(t) = \sum_{i=1}^n \int \frac{dN_{il}(t)}{\sum_{j=1}^n I(Y_j \geq t) \hat{E}[e^{X_j(t)^T \beta_l^{new} + W_j(t)^T b_{jl} + Z_j(t)^T \xi_j}]}, l=1, \dots, L,$$

and

$$\Lambda_k(t) = \sum_{i=1}^n \frac{I(Y_i \leq t) \Delta_i}{\sum_{Y_j \geq Y_i} \hat{E}[e^{\tilde{X}_j(Y_i)^T \gamma_k^{new} + (\phi_k \circ \tilde{Z}_j(Y_i))^T \xi_j}]}, k=1, \dots, K,$$

where the superscript “new” means that the parameter values are the ones from solving the previous equations. Finally, Σ_l can be updated as $n^{-1} \sum_{i=1}^n \hat{E}[b_{il} b_{il}^T]$ and Ψ is given by $n^{-1} \sum_{i=1}^n \hat{E}[\xi_i \xi_i^T]$.

In all these expressions, $\hat{E}[\cdot]$ denotes the conditional expectation of the missing data, i.e., b_{il} and ξ_i , given the observed data. Since this conditional density is proportional to

$$\prod_{l=1}^L \left[\prod_t \left\{ e^{W_i(t)^T b_{il} + Z_i(t)^T \xi_i} \right\}^{\delta N_{il}(t)} \exp \left\{ - \int_0^{Y_i} e^{X_i(t)^T \beta_l + W_i(t)^T b_{il} + Z_i(t)^T \xi_i} dA_l(t) \right\} \times \prod_{k=1}^K \left[\left\{ e^{(\phi_k \circ \tilde{Z}_i(Y_i))^T \xi_i} \right\}^{\Delta_{ik}} \exp \left\{ - \int_0^{Y_i} e^{\tilde{X}_i(t)^T \gamma_k + (\phi_k \circ \tilde{Z}_i(t))^T \xi_i} d\Lambda_k(t) \right\} \right] \times e^{-b_{il}^T \Sigma_l^{-1} b_{il} / 2} \right] \times e^{-\xi_i^T \Psi^{-1} \xi_i / 2},$$

the conditional expectation is usually calculated using numerical integration algorithms such as Gaussian Quadrature. An important point here is that although L can be large, the actual expectation to be calculated in the M-step involves ξ_i and at most one of the b_{il} 's. Therefore, the dimension of the numerical integration is at most two in our MDS application.

We iterate between the E-step and M-step until convergence. To estimate the asymptotic covariance matrix of the parameter estimates, one approach is to use the Louis formula (Louis, 1982) to estimate the observed information matrix for all the parameters including the β 's, γ 's, ϕ_i 's, Σ and all the jumps of A_l and Λ_k . The asymptotic covariance matrix is then the inverse of this matrix. Alternatively, when we are only interested in inference on the regression parameters, another approach is to calculate the profile log-likelihood function for these parameters in a neighborhood of their estimates and by profile likelihood theory, the inverse of the negative curvature of this profile log-likelihood function is a consistent estimator for the asymptotic covariance matrix for the parameters of interest. Specifically, to calculate the profile log-likelihood function, we can simply apply the same EM algorithm as before but hold the regression parameters fixed in the EM iterations.

2.4 Asymptotic properties

We provide the asymptotic properties for the NPMLE's. We impose the following conditions.

- (C1) The true parameter value for $\theta \equiv (\beta_l, \Sigma_l, l = 1, \dots, L, \gamma_k, \phi_k, k = 1, \dots, K, \Psi)$, denoted by θ_0 , belongs to the interior of a compact set in the domain. Additionally, A_l and Λ_k are continuously differentiable and bounded away from zero in $[0, \tau]$, where τ is the study duration.
- (C2) With probability one, $X_i(\cdot)$, $W_i(\cdot)$, $\tilde{X}_i(\cdot)$ and $\tilde{Z}_i(\cdot)$, $Z_i(\cdot)$ are left-continuous with uniformly bounded left- and right- derivatives in $[0, \tau]$.
- (C3) With probability one, $P(C_i - \tau | Z_i) > \delta_0 > 0$ for some constant δ_0 .
- (C4) If there exist $c(t)$ and v such that $c(t) + v^T X_i(t) = 0$ with probability 1, then $c(t) = 0$ and $v = 0$. The same condition holds for $\tilde{X}_i(t)$. In addition, there exists some $t \in [0, \tau]$ such that the linear space spanned by $\{w : w \in \text{supp}(W_i(t))\}$ is the whole space of b_{il} and the linear space spanned by $\{z : z \in \text{supp}(Z_i(t))\}$ is the whole space of ξ_i . The latter also holds for $\{Z_i(t)\}$ for some t .

In our MDS application, $W_i = Z_i = \tilde{Z}_i = 1$ and $X_i(t) = \tilde{X}_i(t)$ is time-independent. Then conditions (C2) and (C4) hold when $[1, X_i]$ are linearly independent with positive probability. Under these conditions, it is easy to see that conditions (D1)–(D8) in Zeng and Lin (2010) hold. Therefore, following their arguments of Section 10.1, we obtain the following theorem.

Theorem 1. *Under conditions (C1)–(C4), the NPMLE's for θ , denoted by $\hat{\theta}$, and $(\hat{A}_l, \hat{A}_k)$ satisfy*

$$\sqrt{n}(\hat{\theta} - \theta_0, \hat{A}_l - A_{l0}, \hat{A}_k - A_{k0}, l=1, \dots, L, k=1, \dots, K) \rightarrow_d \mathcal{G},$$

where $\mathcal{G}$ is a mean-zero and tight Gaussian process in $R^s \times BV[0, \tau]^{L+K}$ with $s = \dim(\theta_0)$.

Moreover, the asymptotic covariance matrix of $\sqrt{n}(\hat{\theta} - \theta_0)$ attains the semi-parametric efficiency bound.

Following Zeng and Lin (2010), we also obtain that the profile log-likelihood yields a consistent estimator of the asymptotic covariance matrix of $\hat{\theta}$.

3 Simulation Study

We conducted a set of extensive simulation studies to examine the small-sample performance of the proposed model. In the simulation studies, we let $W = Z = 1$. The covariates in the model are time-independent and consist of X_1 which is taken to have a Bernoulli distribution, and X_2 is taken to have a uniform distribution in $[0, 1]$. The baseline functions for the recurrent events are set to be constants which may vary for different types of events. We consider two types of survival events and the baseline hazard function for the first event time is a constant while the baseline hazard function for the second event time is linear in time. We choose different values for the β 's, γ 's and ϕ 's to reflect the different covariate effects on the recurrent events or survival events as well as the association between the recurrent events and survival events, where the actual values are given in Tables 1 and 2. In our simulation design, we take $L = 2, K = 2$ and $L = 4, K = 2$, and we consider sample sizes of $n = 400$ and $n = 800$. For $L = 2$ and $K = 2$, the average numbers of the two types of recurrent events per subject is 0.25 and 0.3, while the average censoring rate is around 59% and 54%, respectively. For $L = 4$ and $K = 2$, the average numbers of recurrent events are, respectively, 0.25, 0.39, 1.37 and 1.84, and the censoring rates are 37% and 64%.

For each simulated dataset, we applied the proposed methodology to estimate the parameters. The EM algorithm converges for every simulated data. Since the number of parameters to be estimated is large, we use the profile likelihood method to estimate the asymptotic covariance matrix of the estimates of the regression coefficients. In the profile likelihood method, the perturbation parameter in computing the second order numerical differences of the log-likelihood function is set to be $n^{-1/2}$. The results from 1,000 replicates are summarized in Tables 1 and 2. In the tables, β_{lk} denotes the coefficient of X_k in the l^{th} recurrent event and γ_{qk} is the coefficient of X_k in the q^{th} survival event. The column "Est" is the average value of the estimates from 1,000 replicates and the column "SD" is the standard deviation of the estimates. The column "ESE" reports the average of the estimated standard errors and the column "CP" is the coverage probability for 95% confidence intervals based on asymptotic normality, where the coverage for the variance estimates uses the Satterthwaite approximation in order to correct for the skewness in these parameter estimates. For the estimates of the baseline functions, we also report the average mean square errors in the tables.

The results indicate that the proposed method works well in real settings with small bias, yielding precise estimates of the variability and appropriate coverage probabilities. The performance improves with increasing sample sizes. The prediction of the baseline cumulative functions in both the recurrent events and survival events is accurate. The estimation for ϕ 's, the parameters capturing the dependence due to latent effects among all these endpoints, tends to have a wide confidence intervals. This phenomenon has been observed in some previous work (Zeng and Lin, 2009). It may be because that ϕ 's are equivalent to the scaled variance components of the frailties in survival endpoints so may not have sufficiently large information to be estimated well.

We further consider a simulation study with a time by covariate interaction. The models are similar to the first simulation study with $L = 2$ and $K = 2$ except that the baseline hazard functions are both constants and in both the recurrent event model and the survival model, the covariates are X and $X * t$ where X is generated from a uniform $[0, 1]$ distribution. In the simulations, the average numbers of the two types of recurrent events are 0.25 and 0.33 while the censoring rates are 80% and 24% for the two types of survival endpoints. The results from 1,000 replicates are given in Table 3. Similar conclusions as those of Tables 1 and 2 are obtained.

4 MDS Clinical Trial

The proposed method is applied to a phase 2, randomized, double blind, placebo controlled clinical trial in which lower risk MDS patients were treated with MDS disease modifying therapies. One of the major goals of the trial is to evaluate the IP (treatment) effect on reduction of platelet transfusion and bleeding events under this setting. It is known that the most common side effect of MDS disease modifying therapies is thrombocytopenia (platelet count $< 100 \times 10^9/L$). Low platelet counts will prevent the on-schedule of therapies administered and increase the risk of bleeding, and thus result in higher morbidity. Platelet transfusion interventions are often required therapeutically (given to patients who are actively bleeding) and prophylactically (to prevent future bleeding). Therefore, the incidence of platelet transfusion intervention and bleeding event occurrence are highly correlated. Thus, the composite endpoint of bleeding events and platelet transfusion events is currently being explored for future clinical program development. Informative censoring events are also included in this analysis. These events, which include death or requirement of alternative therapies, lead early study discontinuation and hence these termination events are accounted for in the analysis model appropriately.

The trial was conducted in multiple centers in the United States from 2008 to 2009. A total of 106 subjects were enrolled in the study, 38 (36%) subjects received placebo (standard of care) and 68 (64%) subjects received one of the two dose cohorts of the investigational product. Subjects were randomized and stratified by baseline platelet count $50 \times 10^9/L$. Two subjects were excluded from the analysis due to not receiving any investigational product. The median follow-up time was 133 days and during the follow-up, on average, patients experienced 1.6 bleeding and 4.8 transfusion events. Furthermore, potential informative censoring events were Acute Myeloid Leukemia (AML) or death (7 cases), or other informative censoring reasons (35 cases) including administrative decision (14 cases), adverse events (13 cases), and requirement for alternative therapy (8 cases). In addition to the investigational product received (yes/no), the other covariates included in the analysis include stratification factor (platelet count status at baseline) and the baseline disease risk status of patients. The latter is classified into higher-risk (IPSS INT-1, INT-2 and high) and lower-risk (IPSS low) groups based on a subject's disease characteristics. Table 4 shows some descriptive statistics for this trial, including platelet, risk, death rate, informative censoring, transfusion, and bleeding information.

Since the incidence of platelet transfusion intervention and bleeding event occurrence are highly correlated, we first exclude those events which are apparently caused by each other from the analysis. Specifically, any transfusion induced within one week of some bleeding event is considered as the consequence of the bleeding and, therefore is excluded. In addition, any transfusions within 7 days of each other are very likely to be due to administrative purposes and are therefore also considered as a single platelet transfusion event. Therefore, we believe that the remaining transfusion events are the patient's self-spontaneous events and are medical events not due to bleeding. In the data for our analysis, the patients, on average, experienced 1.42 bleeding events and 1.77 transfusions.

We fit our proposed model to analyze the data with two types of recurrent events ($L = 2$). Since there were only 7 AML or death cases, we combine both death and informative dropout into a single type of survival endpoint so $K = 1$. We start with the same initial values as in the simulations, and the perturbation parameter that is chosen in the variance estimation is $1/\sqrt{n}$, where n is the number of patients. The estimates, standard errors (SEs), and p-values from the joint analysis based on the proposed method and the separate analysis are given in Table 5. Table 5 shows that the treatment appears to reduce both the bleeding incidences by 16% ($1 - \exp(-0.177)$) or transfusions incidence by 51%. Moreover, there appears to be marginal evidence that the treatment would result in fewer deaths or early withdrawal. The patients with higher baseline platelet counts tend to experience less bleeding or transfusions, and the patients in the higher-risk group are likely to experience events and die or withdraw from the study. Although the estimate of ϕ does not really indicate strong evidence for the association between the recurrent events and the survival endpoint maybe due to small data size, the large variability of σ_{ξ}^2 and $\sigma_{b_i}^2$ reveals strong dependence between the two types of recurrent events and a strong dependence between all events in general. Even though ϕ is not significant at the 5% significance level, we still see from Table 5 that there are some differences in the p-values between the proposed joint analysis and the separate analysis. In particular, the treatment effects for bleeding events from the joint analysis and the separate analysis have the opposite signs, although they both are not significant, and the risk group is significantly associated with the survival endpoint in the joint analysis while it is not if we analyze the survival endpoint separately. Figure 1 shows the estimated cumulative intensity function and cumulative hazard function. It shows a steady constant risk for both bleeding events and transfusion events.

Since transfusion and bleeding events may be treated as the events concerning the same status of patients clinically, we also conducted an analysis by combining these two recurrent events into one composite endpoint. As before, we not only exclude any transfusions induced within one week of some bleeding events, but also exclude any transfusions within 7 days of a previous transfusion. The results from this composite analysis are given in Table 5, which also indicates that the treatment, although not statistically significant, tends to reduce the occurrence of bleeding/transfusion and the risk of death or withdrawal from the study. The strong evidence for the association between the recurrent events and the survival endpoint is still apparent.

5 Concluding Remarks

We have proposed a joint model for analyzing multiple types of recurrent events and multiple types of informative censoring times. The joint model accounts for within-event dependence and also between-event dependence through a decomposition of event-specific frailty and the shared frailty. A major advantage of this proposed model is that it allows any number of types of events without resorting to high-dimensional numerical integration due

to the number of events in the inference. We have applied the proposed model to analyze an MDS clinical trial.

Our joint models can be easily used for future prediction, for example, given one's recurrent event history, how likely he or she will experience one type of recurrent event or drop out after time t . The confidence intervals of the prediction can be constructed using the Delta method. Although the EM algorithm converges, it can be slow with a large number of events. An alternative approach is to use a penalized quadrature approach in Liu and Huang (2010). A substantial numerical comparison will be interesting.

Although we use the proportional intensity and proportional hazards model in the joint model, other transformation models can be easily incorporated into the modeling formulation. In addition, we can generalize our model to incorporate longitudinal biomarkers into the analysis, for example, the platelet biomarker is measured over time in the MDS study. Other complex features such as time-varying treatment effects can also be included in this joint model.

The detailed development of the implementation of the proposed algorithm, and the SAS code JointRT with an illustrative example including the SAS output and data sets are given in the Web based supplementary materials. The MDS data is confidential and, thus, it is not available for the public release.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank the Editor, an Associate Editor and two referees for their helpful comments and suggestions, which have led to an improvement of this article. We also thank Mr. Xin Zhou for helping to program the proposed algorithm in SAS. This research was partially supported by the NIH grant #GM 70335.

References

- Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*. 1982; 10:1100–1200.
- Chen BE, Cook RJ. Test for multivariate recurrent events in the presence of a terminal event. *Biostatistics*. 2004; 5:129–143. [PubMed: 14744832]
- Cook RJ, Lawless JF. Marginal analysis of recurrent events and a terminal event. *Statistics in Medicine*. 1997; 16:911–924. [PubMed: 9160488]
- Ghosh D, Lin DY. Nonparametric analysis of recurrent events and death. *Biometrics*. 2000; 56:554–562. [PubMed: 10877316]
- Ghosh D, Lin DY. Marginal regression models for recurrent and terminal events. *Statistica Sinica*. 2002; 12:663–688.
- Huang C, Wang M. Joint modeling and estimation for recurrent event processes and failure time data. *Journal of the American Statistical Association*. 2004; 99:1153–1165. [PubMed: 24068850]
- Liu L, Huang X. The use of Gaussian quadrature for estimation in frailty proportional hazards models. *Statistics in Medicine*. 2008; 27:2665–2683. [PubMed: 17910008]
- Liu L, Wolfe RA, Huang X. Shared frailty models for recurrent events and a terminal event. *Biometrics*. 2004; 60:747–756. [PubMed: 15339298]
- Louis T. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*. 1982; 44:226–233.
- Prentice RL, Williams BJ, Peterson AV. On the regression analysis of multivariate failure time data. *Biometrika*. 1981; 68:373–379.

- Wang M, Qin J, Chiang C. Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association*. 2001; 96:1057–1065.
- Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*. 1989; 84:1065–1073.
- Zeng D, Lin DY. Semiparametric Transformation Models with Random Effects for Joint Analysis of Recurrent and Terminal Events. *Biometrics*. 2009; 65:746–752. [PubMed: 18945267]
- Zeng D, Lin DY. A General Asymptotic Theory for Maximum Likelihood Estimation in Semiparametric Regression Models With Censored Data. *Statistica Sinica*. 2010; 20:871–910. [PubMed: 20577580]
- Zhu L, Sun J, Tong X, Srivastava DK. Regression analysis of multivariate recurrent event data with a dependent terminal event. *Lifetime Data Analysis*. 2010; 16:478–490. [PubMed: 20217474]
- Zhu L, Sun J, Srivastava DK, Tong X, Leisenring W, Zhang H, Robison LL. Semiparametric transformation models for joint analysis of multivariate recurrent and terminal events. *Statistics in Medicine*. 2011; 30:3010–3023. [PubMed: 21786280]

Figure 1.

Estimated baseline functions in the analysis of MDS data: (a) the solid curve is the estimated cumulative intensity function for bleeding events and the dashed curve is the estimated cumulative intensity function for transfusion events; (b) the solid curve is the estimated cumulative hazards function for survival endpoint.

Table 1

Simulation results with time-independent covariates for $L = 2$ and $K = 2$

Parameter	$n = 400$					$n = 800$				
	True	Est	SD	ESE	CP	Est	SD	ESE	CP	CP
β_{11}	-0.5	-0.511	0.251	0.267	0.97	-0.507	0.178	0.182	0.96	0.96
β_{12}	0.5	0.504	0.410	0.413	0.94	0.497	0.288	0.270	0.93	0.93
β_{21}	0	-0.015	0.221	0.232	0.96	-0.003	0.158	0.159	0.95	0.95
β_{22}	0.5	0.528	0.393	0.373	0.94	0.522	0.267	0.245	0.93	0.93
γ_{11}	0.4	0.415	0.169	0.170	0.95	0.403	0.113	0.116	0.96	0.96
γ_{12}	-0.4	-0.414	0.270	0.289	0.95	-0.398	0.203	0.197	0.95	0.95
γ_{21}	0.5	0.508	0.169	0.171	0.95	0.507	0.111	0.114	0.96	0.96
γ_{22}	-0.5	-0.504	0.277	0.286	0.96	-0.520	0.193	0.192	0.95	0.95
ϕ_1	0.2	0.145	0.690	0.711	1.00	0.175	0.477	0.437	0.99	0.99
ϕ_2	0.5	0.480	0.586	0.627	0.99	0.449	0.389	0.385	0.95	0.95
σ_{ξ}^2	0.25	0.302	0.111	0.258	0.98	0.303	0.083	0.174	0.97	0.97
$\sigma_{b_1}^2$	0.5	0.497	0.182	0.423	0.99	0.516	0.134	0.276	0.98	0.98
$\sigma_{b_2}^2$	0.5	0.487	0.183	0.379	0.98	0.487	0.125	0.247	0.99	0.99

Baseline	$n = 400$					$n = 800$				
	True	Est	SD	MSE ($\times 10^{-3}$)	CP	Est	SD	MSE ($\times 10^{-3}$)	CP	CP
$A_1(\tau/4)$	0.1	0.100	0.029	0.854	0.98	0.098	0.021	0.447	0.97	0.97
$A_1(3\tau/4)$	0.3	0.299	0.089	7.882	0.98	0.296	0.066	4.333	0.97	0.97
$A_2(\tau/4)$	0.1	0.100	0.028	0.805	0.98	0.098	0.020	0.392	0.97	0.97
$A_2(3\tau/4)$	0.3	0.301	0.088	7.716	0.98	0.294	0.059	3.500	0.97	0.97
$\Lambda_1(\tau/4)$	0.25	0.244	0.050	2.519	0.98	0.246	0.035	1.240	0.97	0.97
$\Lambda_1(3\tau/4)$	0.75	0.755	0.153	23.305	0.98	0.749	0.100	10.044	0.97	0.97

$\Lambda_2(\pi/4)$	0.125	0.123	0.029	0.845	0.125	0.020	0.404
$\Lambda_2(3\pi/4)$	1.125	1.141	0.242	58.753	1.143	0.156	24.757

Table 2

Simulation results with time-independent covariates for $L = 4$ and $K = 2$

Parameter	$n = 400$					$n = 800$				
	True	Est	SD	ESE	CP	Est	SD	ESE	CP	
β_{11}	-0.5	-0.524	0.256	0.275	0.97	-0.504	0.183	0.186	0.95	
β_{12}	0.5	0.496	0.413	0.450	0.96	0.499	0.293	0.301	0.96	
β_{21}	0	-0.018	0.208	0.213	0.96	-0.002	0.143	0.145	0.94	
β_{22}	1.0	1.013	0.361	0.369	0.95	1.015	0.251	0.248	0.94	
β_{31}	0.6	0.599	0.161	0.148	0.92	0.600	0.111	0.101	0.92	
β_{32}	0.8	0.818	0.270	0.259	0.94	0.805	0.195	0.174	0.92	
β_{41}	1.0	0.995	0.149	0.140	0.93	0.991	0.105	0.094	0.91	
β_{42}	0	0.000	0.263	0.245	0.93	0.008	0.181	0.165	0.92	
γ_{11}	0.4	0.408	0.137	0.135	0.95	0.401	0.091	0.093	0.95	
γ_{12}	0.4	0.407	0.233	0.232	0.95	0.408	0.160	0.159	0.95	
γ_{21}	0.5	0.498	0.202	0.202	0.96	0.499	0.133	0.137	0.96	
γ_{22}	0	0.014	0.326	0.339	0.96	-0.014	0.228	0.229	0.95	
φ_1	0.2	0.204	0.331	0.321	0.98	0.199	0.220	0.212	0.96	
φ_2	1.0	1.038	0.483	0.483	0.96	1.010	0.304	0.315	0.96	
σ_ξ^2	0.25	0.248	0.064	0.081	0.97	0.251	0.046	0.055	0.976	
$\sigma_{b_1}^2$	0.5	0.549	0.195	0.395	0.97	0.563	0.135	0.251	0.97	
$\sigma_{b_2}^2$	0.5	0.516	0.177	0.274	0.96	0.519	0.121	0.177	0.97	
$\sigma_{b_3}^2$	0.8	0.787	0.155	0.172	0.97	0.793	0.113	0.117	0.96	
$\sigma_{b_4}^2$	0.8	0.786	0.151	0.157	0.97	0.790	0.104	0.107	0.96	
Parameter	$n = 400$					$n = 800$				
	True	Est	SD	ESE	CP	Est	SD	ESE	CP	

Baseline	True	Est	SD	MSE	Est	SD	MSE
$A_1(\tau/2)$	0.4	0.404	0.119	0.014	0.391	0.082	0.0067
$A_2(\tau/2)$	0.4	0.406	0.109	0.012	0.396	0.074	0.0055
$A_3(\tau/2)$	1.0	1.014	0.207	0.043	1.003	0.141	0.020
$A_4(\tau/2)$	1.6	1.624	0.292	0.086	1.613	0.205	0.042
$\Lambda_1(\tau/4)$	0.5	0.502	0.082	0.0067	0.498	0.054	0.0029
$\Lambda_1(3\tau/4)$	1.5	1.536	0.351	0.125	1.501	0.220	0.048
$\Lambda_2(\tau/4)$	0.25	0.252	0.062	0.0038	0.254	0.042	0.0018
$\Lambda_2(3\tau/4)$	2.25	2.328	0.758	0.581	2.326	0.505	0.261

Table 3

Simulation results with time-dependent covariates for $L = 2$ and $K = 2$

Parameter	True	$n = 400$				$n = 800$			
		Est	SD	ESE	CP	Est	SD	ESE	CP
β_{11}	-0.5	-0.520	0.598	0.627	0.96	-0.507	0.408	0.417	0.96
β_{12}	0.5	0.579	2.326	2.352	0.95	0.554	1.471	1.478	0.94
β_{21}	0	-0.015	0.513	0.556	0.96	0.008	0.359	0.372	0.96
β_{22}	0.5	0.533	2.265	2.343	0.96	0.515	1.327	1.308	0.95
γ_{11}	-0.4	-0.420	0.639	0.637	0.94	-0.387	0.447	0.429	0.94
γ_{12}	0.4	0.413	2.265	2.343	0.97	0.355	1.568	1.494	0.94
γ_{21}	-0.5	-0.499	0.310	0.322	0.96	-0.501	0.218	0.221	0.95
γ_{22}	0.5	0.468	1.093	1.196	0.96	0.493	0.766	0.787	0.96
ϕ_1	0.2	0.162	0.841	0.969	1.00	0.184	0.613	0.544	0.99
ϕ_2	0.5	0.477	0.421	0.975	0.98	0.456	0.331	0.329	0.94
σ_{ξ}^2	0.25	0.302	0.120	0.250	0.96	0.297	0.081	0.164	0.97
$\sigma_{b_1}^2$	0.5	0.496	0.201	0.412	0.98	0.504	0.139	0.260	0.98
$\sigma_{b_2}^2$	0.5	0.460	0.177	0.355	0.99	0.482	0.131	0.231	0.99

Baseline	True	$n = 400$				$n = 800$			
		Est	SD	MSE ($\times 10^{-1}$)		Est	SD	MSE ($\times 10^{-1}$)	
$A_1(\tau/4)$	0.25	0.254	0.073	0.053		0.250	0.049	0.024	
$A_1(3\tau/4)$	0.75	0.791	0.325	1.071		0.763	0.218	0.479	
$A_2(\tau/4)$	0.25	0.254	0.066	0.044		0.249	0.046	0.022	
$A_2(3\tau/4)$	0.75	0.794	0.322	1.057		0.762	0.198	0.393	
$\Lambda_1(\tau/4)$	0.25	0.245	0.067	0.045		0.245	0.048	0.023	
$\Lambda_1(3\tau/4)$	0.75	0.783	0.306	0.948		0.766	0.210	0.443	

$\Lambda_2(\tau/4)$	1.0	1.007	0.133	0.178	1.004	0.097	0.094
$\Lambda_2(3\tau/4)$	3.0	3.177	0.753	5.976	3.087	0.521	2.793

Table 4

Descriptive Information of MDS Study

	Control arm (<i>n</i> = 38)	Treatment arm (<i>n</i> = 68)
High baseline platelet(%)	47.4%	50.0%
High risk (%)	84.2%	79.4%
Death/AML rate (%)	7.9%	5.9%
Informative censoring (%)	47.4%	25.0%
# Transfusion/patient	6.105	4.029
# Bleeding/patient	1.684	1.559

Table 5

Results from analysis of MDS study with multiple and composite recurrent events and informative censoring or death

Covariate	Two Recurrent Events			
	Joint Analysis		Separate Analysis	
	Estimate	SE	P-value	Estimate
Bleeding event				
Treatment	-0.177	0.466	0.704	0.115
Baseline platelet ($50 \times 10^9/L$)	-2.080	0.586	<0.001	-2.135
Risk group (higher)	0.578	0.368	0.117	-0.361
Transfusion event				
Treatment	-0.709	0.504	0.160	-0.271
Baseline platelet ($50 \times 10^9/L$)	-2.254	0.628	<0.001	-1.490
Risk group (higher)	2.173	0.394	<0.001	2.232
Informative censoring or death				
Treatment	-0.611	0.345	0.077	-0.583
Baseline platelet ($50 \times 10^9/L$)	0.370	0.323	0.253	0.357
Risk group (higher)	0.636	0.277	0.022	0.631
Variance components				
ϕ	0.228	0.242	0.346	
σ_{ξ}^2	1.440	0.712	0.043	
$\sigma_{b_1}^2$	1.188	0.571	0.038	
$\sigma_{b_2}^2$	0.708	0.468	0.130	
Covariate	One Composite Recurrent Event			
	Joint Analysis		Separate Analysis	
	Estimate	SE	P-value	Estimate
Bleeding or transfusion event				
Treatment	-0.259	0.342	0.450	-0.098
				0.112
				0.380

Baseline platelet ($< 50 \times 10^9/L$)	-2.349	0.419	<0.001	-1.775	0.171	<0.001
Risk group (higher)	1.201	0.343	0.001	0.483	0.258	0.061
	Informative censoring or death					
Treatment	-0.593	0.340	0.081	-0.583	0.313	0.063
Baseline platelet ($< 50 \times 10^9/L$)	0.355	0.308	0.250	0.357	0.329	0.279
Risk group (higher)	0.639	0.269	0.018	0.631	0.504	0.211
	Variance components					
ϕ	0.292	0.281	0.299			
σ_{ξ}^2	1.137	0.829	0.171			
σ_b^2	0.842	0.740	0.255			