Semiparametric Transformation Models With Random Effects for Recurrent Events

Donglin ZENG and D. Y. LIN

In this article we study a class of semiparametric transformation models with random effects for the intensity function of the counting process. These models provide considerable flexibility in formulating the effects of possibly time-dependent covariates on the developments of recurrent events while accounting for the dependence of the recurrent event times within the same subject. We show that the nonparametric maximum likelihood estimators (NPMLEs) for the parameters of these models are consistent and asymptotically normal. The limiting covariance matrices for the estimators of the regression parameters achieve the semiparametric efficiency bounds and can be consistently estimated. The limiting covariance function for the estimator of any smooth functional of the cumulative intensity function also can be consistently estimated. We develop a simple and stable EM algorithm to compute the NPMLEs as well as the variance and covariance estimators. Simulation studies demonstrate that the proposed methods perform well in practical situations. Two medical studies are provided for illustrations.

KEY WORDS: Box–Cox transformation; Counting process; EM algorithm; Intensity function; Nonparametric likelihood; Semiparametric efficiency.

1. INTRODUCTION

Recurrent-events data are commonly encountered in scientific studies because each study subject may repeatedly experience a certain phenomenon. Medical examples of recurrent events are multiple infection episodes and tumor recurrences. Other examples include repeated breakdowns of a certain machinery in reliability testing and repeated purchases of a certain product in marketing research. In such studies, investigators are interested in evaluating the effects of covariates (e.g., treatment assignments and demographic characteristics) on the recurrent event times and in predicting the developments of future events given past event histories. The statistical analysis is complicated by the presence of censoring (due to subject withdrawals and study termination), as well as the potential dependence of the recurrent event times within the same subject.

It is natural and convenient to represent the recurrent event times as a counting process. The most popular counting-process model is the proportional intensity model studied by Andersen and Gill (1982). Let $N^*(t)$ denote the number of events that the subject has experienced by time t, and let $\mathbf{X}(t)$ be a vector of possibly time-dependent covariates. The proportional intensity model specifies that the intensity function for $N^*(t)$ associated with $\mathbf{X}$ takes the form

$$\lambda(t|\mathbf{X}) = \lambda_0(t)e^{\boldsymbol{\beta}^T\mathbf{X}(t)}, \tag{1}$$

where $\lambda_0(\cdot)$ is an unspecified baseline intensity function and $\boldsymbol{\beta}$ is a vector of unknown regression parameters. The maximum partial likelihood estimator for $\boldsymbol{\beta}$ and the corresponding Aalen–Breslow estimator for $\Lambda_0(t) \equiv \int_0^t \lambda_0(u)\,du$ are consistent and asymptotically normal (Andersen and Gill 1982).

Under model (1), the occurrence of an event is independent of any earlier events that occurred to the same subject unless $\mathbf{X}(t)$ includes the past history. It is difficult to model correctly the intraclass correlation through time-dependent covariates. Furthermore, it is not appropriate to use such time-dependent covariates when assessing treatment effect in randomized clinical trials, because the inclusion of postrandomization response in the model would attenuate the estimator of treatment effect.

A useful approach to accommodating the dependence of the recurrent event times within the same subject is to incorporate a random effect or frailty ξ into (1),

$$\lambda(t|\mathbf{X};\xi) = \xi\lambda_0(t)e^{\boldsymbol{\beta}^T\mathbf{X}(t)} \tag{2}$$

(e.g., Nielsen, Gill, Andersen, and Sorenson 1992; Oakes 1992). This frailty may represent the intraclass correlation in lieu of or in addition to time-dependent covariates. The presence of frailty poses considerable challenges in statistical inference. To date, rigorous asymptotic theory has been established only for the special case of gamma frailty without covariates (Murphy 1994, 1995). Unfortunately, gamma frailty induces a very restrictive form of dependence. Furthermore, the proportional intensity assumption (i.e., the multiplicative relationship between the baseline intensity function and the exponential regression function) may not be satisfied in applications.

In this article we propose a broad class of intensity models with random effects that accommodates nonproportional intensity and allows various frailty distributions. Specifically, the cumulative intensity function for $N^*(t)$ takes the form

$$\Lambda(t|\mathbf{X},\mathbf{Z};\mathbf{b}) = G\left(\int_0^t \lambda(s)e^{\boldsymbol{\beta}^T\mathbf{X}(s)+\mathbf{b}^T\mathbf{Z}(s)}\,ds\right), \tag{3}$$

where $\boldsymbol{\beta}$ is a set of unknown regression parameters, $\mathbf{b}$ is a set of random effects with density function $\psi(\mathbf{b};\boldsymbol{\gamma})$ indexed by parameters $\boldsymbol{\gamma}$, $\mathbf{X}$ and $\mathbf{Z}$ are the covariate processes associated with the fixed and random effects, $\lambda(\cdot)$ is an arbitrary positive function, and $G(\cdot)$ is a thrice continuously differentiable and strictly increasing transformation function with $G(0) = 0$ and $G(\infty) = \infty$. Note that $\mathbf{X}(t)$ and $\mathbf{Z}(t)$ are possibly time-dependent and may include covariates derived from the event history before time t.

There are considerable flexibilities in choosing the transformation G and the distribution of the random effects $\psi(\mathbf{b};\boldsymbol{\gamma})$. In particular, one may specify the multivariate normal distribution for ψ, which, unlike the gamma distribution, has an unrestricted

Donglin Zeng is Assistant Professor (E-mail: *dzeng@bios.unc.edu*) and D. Y. Lin is Dennis Gillings Distinguished Professor (E-mail: *lin@bios.unc.edu*), Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599. This research was supported by the National Institutes of Health. The authors thank the editor, an associate editor, and three referees for their reviews.

Journal of the American Statistical Association
March 2007, Vol. 102, No. 477, Theory and Methods
DOI 10.1198/016214506000001239

covariance structure. Various of covariate effects can be formulated through the class of Box–Cox transformations,

$$G(x) = \begin{cases} \{(1+x)^{\rho} - 1\}/\rho, & \rho > 0 \\ \log(1+x), & \rho = 0. \end{cases}$$

If $\rho = 1$, then (3) reduces to the proportional intensity model with random effects, in which case the effects of covariates on the intensity function are constant over time. For $\rho > 1$, the covariate effects increase over time; for $\rho < 1$, the covariate effects decrease over time. Another useful class of transformations is

$$G(x) = \begin{cases} \log(1+rx)/r, & r > 0 \\ x, & r = 0. \end{cases}$$

For $r > 0$, the covariate effects always decrease over time, with a higher rate of decrease for larger r. Misspecification of the transformation would lead to incorrect characterization of the covariate effects over time, as well as biased prediction of the occurrence of events over time.

If an important covariate is omitted from the proportional hazards model, then the resulting model will be a transformation model. Various transformations can be derived by postulating different distributions for the omitted covariate. Kosorok, Lee, and Fine (2004) studied such transformation models. The class of models given in (3) extends the transformation models for univariate survival data to recurrent events. These models not only involve transformations, but also contain random effects characterizing the dependence of recurrent event times within the same subject. The characterization of the intraclass correlation through random effects adds considerable complexity to the statistical inference.

It is more challenging, both theoretically and computationally, to deal with the class of models given in (3) rather than with model (2). We show that the nonparametric maximum likelihood estimators for the parameters of (3) are consistent, asymptotically normal, and asymptotically efficient by appealing to modern empirical process theory (van der Vaart and Wellner 1996) and semiparametric efficiency theory (Bickel, Klaasen, Ritov, and Wellner 1993). In addition, we develop a simple EM algorithm to compute the maximum likelihood estimators (MLEs) and their variances and covariances. Finally, we demonstrate through simulation studies and real examples that the proposed methods work well in practical situations.

2. INFERENCE PROCEDURES

2.1 Nonparametric Maximum Likelihood Estimation

Recurrent-event times are commonly subject to right censoring. Let C denote the censoring time. For a random sample of n subjects, the data consist of $\{\mathbf{X}_i(\cdot), \mathbf{Z}_i(\cdot), N_i(\cdot), Y_i(\cdot)\}$, $i = 1, \ldots, n$, where $N_i(t) = N_i^*(t \wedge C_i)$, $Y_i(t) = I(C_i \geq t)$, and $I(\cdot)$ is the indicator function. We wish to use these data to make inference about $\boldsymbol{\theta}$, which is a $d \times 1$ vector of parametric components in model (3), and $\Lambda(t) \equiv \int_0^t \lambda(s)\, ds$, the nonparametric component.

Let τ denote the duration of the study. We assume that the conditional probability of $C > t$ given $\{\mathbf{X}(s), \mathbf{Z}(s), N^*(s); s \in [0, \tau]\}$ and $\mathbf{b}$ depends only on $\{\mathbf{X}(s), \mathbf{Z}(s); s \leq t\}$ and is noninformative about $(\Lambda, \boldsymbol{\theta})$. In addition, we assume that the conditional distribution of $\{\mathbf{X}(t), \mathbf{Z}(t)\}$ given $\{\mathbf{X}(s), \mathbf{Z}(s), N(s), Y(s); s < t\}$ is noninformative about $(\Lambda, \boldsymbol{\theta})$. The first assumption pertains to the assumption of coarsening at random and reduces to the two conditions of Andersen, Borgan, Grill, and Keiding (1993, p. 665) in the absence of covariates. The second assumption, which implies that no information on the parameters can be extracted from the covariates process, is required in any regression analysis.

Under the foregoing assumptions, the log-likelihood function concerning the parameters $(\Lambda, \boldsymbol{\theta})$ is

$$\sum_{i=1}^{n} \log \int_{\mathbf{b}} \prod_{t \leq \tau} \left\{ Y_i(t)\lambda(t) e^{\boldsymbol{\beta}^T \mathbf{X}_i(t) + \mathbf{b}^T \mathbf{Z}_i(t)} \times G'\left(\int_0^t Y_i(s) e^{\boldsymbol{\beta}^T \mathbf{X}_i(s) + \mathbf{b}^T \mathbf{Z}_i(s)}\, d\Lambda(s) \right) \right\}^{\Delta N_i(t)} \times \exp\left\{ -G\left(\int_0^{\tau} Y_i(t) e^{\boldsymbol{\beta}^T \mathbf{X}_i(t) + \mathbf{b}^T \mathbf{Z}_i(t)}\, d\Lambda(t) \right) \right\} \psi(\mathbf{b}; \boldsymbol{\gamma})\, d\mathbf{b},$$

where G' denotes the derivative of G and $\Delta N_i(t)$ denotes the jump of N_i at t. The maximum of this function does not exist if $\Lambda(\cdot)$ is restricted to be absolutely continuous. Thus we allow $\Lambda(\cdot)$ to be any increasing right-continuous function and replace $\lambda(t)$ with the jump size of Λ at time t, denoted by $\Lambda\{t\}$. We then maximize the modified log-likelihood function

$$l_n(\Lambda, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int_{\mathbf{b}} \prod_{t \leq \tau} \left\{ Y_i(t) \Lambda\{t\} e^{\boldsymbol{\beta}^T \mathbf{X}_i(t) + \mathbf{b}^T \mathbf{Z}_i(t)} \times G'\left(\int_0^t Y_i(s) e^{\boldsymbol{\beta}^T \mathbf{X}_i(s) + \mathbf{b}^T \mathbf{Z}_i(s)} d\Lambda(s) \right) \right\}^{\Delta N_i(t)} \times \exp\left\{ -G\left(\int_0^{\tau} Y_i(t) e^{\boldsymbol{\beta}^T \mathbf{X}_i(t) + \mathbf{b}^T \mathbf{Z}_i(t)}\, d\Lambda(t) \right) \right\} \times \psi(\mathbf{b}; \boldsymbol{\gamma})\, d\mathbf{b} \tag{4}$$

over Λ and $\boldsymbol{\theta}$, treating $\Lambda(\cdot)$ as a step function with jumps at the observed event times T_{ij} $(i = 1, \ldots, n; j = 1, \ldots, n_i)$, where n_i is the number of observed events on the ith subject. This maximization is equivalent to maximizing (4) over $\boldsymbol{\theta}$ and $\Lambda\{T_{ij}\}$ $(i = 1, \ldots, n; j = 1, \ldots, n_i)$. The existence of the maximum is shown in Appendix A. The resulting nonparametric MLEs (NPMLEs) for Λ and $\boldsymbol{\theta}$ are denoted by $\widehat{\Lambda}_n$ and $\widehat{\boldsymbol{\theta}}_n$.

2.2 Asymptotic Results for Known Transformations

We describe the asymptotic properties of the NPMLEs when the transformation G is completely specified, in which case $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$. Denote the true values of $\Lambda(t)$ and $\boldsymbol{\theta}$ by $\Lambda_0(t)$ and $\boldsymbol{\theta}_0$. We impose the following conditions:

- Condition 1. The function $\Lambda_0(\cdot)$ is strictly increasing and continuously differentiable with $\Lambda_0'(t) > 0$, and $\boldsymbol{\theta}_0$ lies in the interior of a known compact set in the domain of $\boldsymbol{\theta}$.
- Condition 2. With probability 1, $\mathbf{X}(\cdot)$ and $\mathbf{Z}(\cdot)$ have bounded total variations in $[0, \tau]$. In addition, the following identifiability conditions hold: (a) If there exists a vector $\boldsymbol{\mu}$ and a deterministic function $\alpha_0(t)$ such that $\alpha_0(t) + \boldsymbol{\mu}^T \mathbf{X}(t) = 0$ with probability 1, then $\boldsymbol{\mu} = \mathbf{0}$ and $\alpha_0(t) = 0$; (b) $P\{\mathbf{Z}(t)^T \mathbf{Z}(t)$ is full rank for some $t \in [0, \tau]\} > 0$; (c) $\psi(\mathbf{b}; \boldsymbol{\gamma}) = \psi(\mathbf{b}; \boldsymbol{\gamma}_0)$ for almost every $\mathbf{b}$ implies that $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$; (d) if $\psi'(\mathbf{b}; \boldsymbol{\gamma}_0)^T \mathbf{v}_{\gamma} = 0$ almost everywhere for $\mathbf{b}$, where $\psi'(\mathbf{b}; \boldsymbol{\gamma})$ is the derivative of ψ with respect to $\boldsymbol{\gamma}$, then $\mathbf{v}_{\gamma} = \mathbf{0}$.

- Condition 3. With probability 1, there exists a positive constant δ such that $P\{C \geq \tau | \mathbf{X}(t), \mathbf{Z}(t); t \in [0, \tau]\} > \delta$.
- Condition 4. There exist a constant α and a function $\mu(\cdot)$ such that for any $0 < x_1 < x_2 < \cdots < x_m < y < \tau$,

$$\prod_{i=1}^{m} \{(1 + x_i)G'(x_i)\} \leq \mu(m)(1 + y)^{-\alpha} \exp\{G(y)\}$$

and $E[\log \mu(N^*(\tau))] < \infty$.
- Condition 5. For any finite K, K_1, and K_2,

$$\sup_{\boldsymbol{\gamma}} \left(\int_{\mathbf{b}} e^{K|\mathbf{b}|} \left| \psi^{(m)}(\mathbf{b}; \boldsymbol{\gamma}) \right| d\mathbf{b} + E\left[\log \int e^{\{K_1 + K_2 N^*(\tau)\}|\mathbf{b}|} \left| \psi^{(m)}(\mathbf{b}; \boldsymbol{\gamma}) \right| d\mathbf{b} \right] \right) < \infty$$

for $m = 0, 1, 2$, where $\psi^{(m)}$ is the mth derivative of ψ with respect to $\boldsymbol{\gamma}$.

Remark 1. Parts (a) and (b) of Condition 2 are standard assumptions for (random-effects) regression models. Part (c) means that the parameterization of the random-effects distribution is unique, whereas (d) means that the information matrix for the random-effects distribution is nonsingular. This condition is trivially satisfied by all commonly used distributions. Condition 3 implies that there is a positive probability for the events to be observed in the time interval $[0, \tau]$. Condition 4 covers all commonly used transformations. In particular, the classes of Box–Cox transformations and logarithmic transformations mentioned in Section 1 satisfy this condition because of the inequalities

$$\prod_{i=1}^{m} (1 + x_i)^{\rho} \leq (1 + y)^m \leq [2m/\rho]! \rho^{[2m/\rho]} e^{1/\rho} (1 + y)^{-1/\rho} e^{[(1+y)^{\rho} - 1]/\rho}$$

and

$$\prod_{i=1}^{m} \frac{(1 + x_i)}{1 + r x_i} \leq \frac{\{\max(1/r, 1)\}^m}{\min(r, 1)^{1/r}} (1 + y)^{-1/r} (1 + ry)^{1/r}.$$

It can be shown that the transformations given by Kosorok et al. (2004), which are derived as $\int e^{-\xi x} dF(\xi)$ with $F(\xi)$ following the inverse-Gaussian, gamma, and other distributions, also satisfy Condition 4. Condition 5 is satisfied by the normal, mixture-normal, and other distributions with tails less heavy than exponential. Thus any combination of the aforementioned transformations and random-effects distributions would satisfy all of the conditions.

The consistency of the NPMLEs $(\widehat{\Lambda}_n, \widehat{\boldsymbol{\theta}}_n)$ is stated as follows.

Theorem 1. Under Conditions 1–5, $|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0| \to 0$ and $\|\widehat{\Lambda}_n - \Lambda_0\|_{l^\infty[0,\tau]} \to 0$ almost surely, where $\|\cdot\|_{l^\infty[0,\tau]}$ is the supremum norm in the interval $[0, \tau]$.

The proofs of this theorem and others are given in Appendix A.

To describe the asymptotic distribution, we define $\mathcal{Q} = \{q(t) : q(t) \in BV[0, \tau], \|q\|_{BV[0,\tau]} \leq 1\}$, where $BV[0, \tau]$ denotes the set of functions with bounded total variations and $\|q\|_{BV[0,\tau]}$ denotes the total variation of $q(t)$ in $[0, \tau]$. Then $\widehat{\Lambda}_n(t)$ can be considered as a bounded linear functional in $l^\infty(\mathcal{Q})$ by the definition $\widehat{\Lambda}_n(q) = \int_0^\tau q(t)\, d\widehat{\Lambda}_n(t)$. Thus $(\widehat{\Lambda}_n - \Lambda_0, \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ is treated as a random element in the metric space $l^\infty(\mathcal{Q}) \times \mathcal{R}^d$.

Theorem 2. Suppose that Conditions 1–5 hold. Then $\sqrt{n} \times (\widehat{\Lambda}_n - \Lambda_0, \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges weakly to a mean-0 Gaussian process in the metric space $l^\infty(\mathcal{Q}) \times \mathcal{R}^d$. In addition, the asymptotic covariance matrix of $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ attains the semiparametric efficiency bound.

Theorem 2 implies that $\sqrt{n}(\widehat{\Lambda}_n - \Lambda_0)$ and $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ are asymptotically normal. Estimating their asymptotic variances is useful. These statistics can be expressed in the form $\sqrt{n}\int_0^\tau q(t)\, d(\widehat{\Lambda}_n - \Lambda_0) + \sqrt{n}\mathbf{u}^T(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$, where $q(t) \in \mathcal{Q}$ and $\mathbf{u} \in \mathcal{R}^d$. Denote the random element in the limiting distribution by $(\mathcal{B}, \mathbf{V}) \in l^\infty(\mathcal{Q}) \times \mathcal{R}^d$. Then the random variable $\sqrt{n}\int_0^\tau q(t)\, d(\widehat{\Lambda}_n - \Lambda_0) + \sqrt{n}\mathbf{u}^T(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ is asymptotically normal with mean 0 and variance $\mathrm{var}(\mathcal{B}[q] + \mathbf{u}^T\mathbf{V})$, and this normal approximation is uniform in q and $\mathbf{u}$. To estimate the variance, we view (4) a parametric log-likelihood with $\Lambda\{T_{ij}\}$ $(i = 1, \ldots, n; j = 1, \ldots, n_i)$ and $\boldsymbol{\theta}$ the parameters. Then, according to the parametric likelihood theory, the asymptotic covariance matrix for these parameters can be estimated by the inverse of the observed information matrix. The observed information matrix is equal to the negative Hessian matrix of (4) at $\widehat{\Lambda}_n\{T_{ij}\}$ $(i = 1, \ldots, n; j = 1, \ldots, n_i)$ and $\widehat{\boldsymbol{\theta}}_n$, which is denoted by $n\mathbf{I}_n$. Thus, the asymptotic variance of $\sqrt{n}\int_0^\tau q(t)\, d(\widehat{\Lambda}_n - \Lambda_0) + \sqrt{n}\mathbf{u}^T(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ is equal to that of $\sqrt{n}\sum_{i=1}^n \sum_{j=1}^{n_i} q(T_{ij})\widehat{\Lambda}_n\{T_{ij}\} + \sqrt{n}\mathbf{u}^T(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$, which is estimated by $\widehat{V}_n \equiv (\mathbf{q}^T, \mathbf{u}^T)\mathbf{I}_n^{-1}(\mathbf{q}^T, \mathbf{u}^T)^T$, where $\mathbf{q}$ consists of $q(T_{ij})$ $(i = 1, \ldots, n; j = 1, \ldots, n_i)$. The following theorem justifies the proposed variance estimator.

Theorem 3. Under Conditions 1–5, $\sup_{q \in \mathcal{Q}, |\mathbf{u}| \leq 1} |\widehat{V}_n - \mathrm{var}(\mathcal{B}[q] + \mathbf{u}^T\mathbf{V})| \to 0$ in probability.

The results in the foregoing theorems allow us to make inference about $\boldsymbol{\theta}$ and Λ, and in fact about any Hadamard-differentiable functional of Λ and $\boldsymbol{\theta}$. In particular, the conditional survival function for the second recurrence time T_2 given that the first recurrence time T_1 is equal to t_1 for a subject with covariate values $\mathbf{x}$ and $\mathbf{z}$ is estimated consistently by

$$\int_{\mathbf{b}} \exp\left\{ G\left(\int_0^{t_1} e^{\widehat{\boldsymbol{\beta}}_n^T \mathbf{x}(s) + \mathbf{b}^T \mathbf{z}(s)}\, d\widehat{\Lambda}_n(s) \right) - G\left(\int_0^{t} e^{\widehat{\boldsymbol{\beta}}_n^T \mathbf{x}(s) + \mathbf{b}^T \mathbf{z}(s)}\, d\widehat{\Lambda}_n(s) \right) \right\} \psi(\mathbf{b}; \widehat{\boldsymbol{\gamma}}_n)\, d\mathbf{b}$$

for $t > t_1$. The variance of this estimator can be consistently estimated according to Theorem 3 and the functional δ-method.

Remark 2. An alternative approach to estimating the asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}_n$ is with the profile log-likelihood function (Murphy and van der Vaart 2000), in which the negative second-order numerical difference of the profile log-likelihood function at $\widehat{\boldsymbol{\theta}}_n$ is used to estimate the inverse covariance matrix. This approach, however, does not provide variance estimation for $\widehat{\Lambda}_n$, which is an important limitation because prediction of recurrent events is highly desirable under transformation models.

2.3 Asymptotic Results for Unknown Transformation Parameters

Suppose that the transformation G belongs to a one-parameter family, $\{G_\eta : \eta \in (a_0, b_0)\}$, where a_0 and b_0 are finite bounds for η. In this case, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \eta)^T$. We assume that Conditions 1, 3, and 5 hold and that Condition 4 holds for α and $\mu(\cdot)$ regardless of the value of η. In addition, we impose a stronger version of Condition 2:

- Condition 2′. With probability 1, $\mathbf{X}(\cdot)$ and $\mathbf{Z}(\cdot)$ have bounded total variations in $[0, \tau]$. In addition, the following identifiability conditions hold: (a) if there exists a vector $\boldsymbol{\mu}$ and a constant α_0 such that $\alpha_0 + \boldsymbol{\mu}^T\mathbf{X}(0) = 0$ with probability 1, then $\boldsymbol{\mu} = \mathbf{0}$ and $\alpha_0 = 0$; (b) $P\{\mathbf{Z}(0)^T\mathbf{Z}(0)$ is full rank$\} > 0$; (c) $\psi(\mathbf{b}; \boldsymbol{\gamma}) = \psi(\mathbf{b}; \boldsymbol{\gamma}_0)$ for almost every $\mathbf{b}$ implies that $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$; and (d) if $\psi'(\mathbf{b}; \boldsymbol{\gamma}_0)^T\mathbf{v}_\gamma = 0$ almost everywhere for $\mathbf{b}$, then $\mathbf{v}_\gamma = \mathbf{0}$.

When covariates are time-independent, Condition 2 and Condition 2′ are the same. The linear independence of covariates at time zero was also imposed by Kosorok et al. (2004). We require some smoothness conditions for the transformation with respect to its parameter η:

- Condition 6′. G_η is twice-continuously differentiable with respect to η; $G'_\eta(0) = 1$, $\dot{G}'_\eta(0) = 0$, and $\dot{G}''_\eta(0) < 0$, where $\dot{G}_\eta$ denotes the derivative of G with respect to η.

Remark 3. Condition 6′ is similar to, but slightly weaker than the last condition of (D1) of Kosorok et al. (2004). It is easy to show that this condition holds for the two families of transformations mentioned in Section 1.

Theorem 4. Under Conditions 1, 2′, 3, 4, 5, and 6′, the conclusions of Theorems 1–3 hold.

2.4 Numerical Methods

The maximization of (4) can be achieved through various optimization algorithms, such as those implemented in MATLAB. In Appendix B we describe a simple and reliable EM algorithm (Dempster, Laird, and Rubin 1977) for calculating the NPMLEs and their variance estimators for known transformations. In the E-step, conditional expectations are evaluated through numerical integration, such as the Gaussian-quadrature approximation for normal random effects. In the M-step, we are able to reduce the number of equations to be solved to $(d + 1)$ by taking advantage of a recursive formula for $\widehat{\Lambda}_n$. To calculate the variance estimators, it suffices to evaluate the observed information matrix for $\widehat{\boldsymbol{\theta}}_n$ and $\widehat{\Lambda}_n\{T_{ij}\}$ $(i = 1, \ldots, n; j = 1, \ldots, n_i)$. This is accomplished through the well-known formula of Louis (1982).

When the transformation parameter is unknown, we estimate it along with other model parameters using the MATLAB optimization algorithm *fminunc*. The variances and covariances can be estimated by inverting the observed information matrix for all of the parameters in the extended model or by using the profile likelihood method.

3. SIMULATION STUDIES

We conducted extensive simulation studies to assess the performance of the proposed methods. We generated recurrent event times from the counting process with cumulative intensity $\Lambda(t|X_1, X_2; b) = G(\Lambda(t)e^{-.5X_1+X_2+b})$, where X_1 is Bernoulli with success probability .5, $X_2 = X_1 + \epsilon I(|\epsilon| < 1) + I(|\epsilon| \geq 1)$, ϵ is standard normal, b is normal with mean 0 and variance σ^2, $\Lambda(t) = \alpha \log(1 + t)$, and $G(x) = \{(1 + x)^\rho - 1\}/\rho$ or $\log(1 + rx)/r$. We generated censoring times from the uniform $[2, 6]$ distribution and set τ to 4. We chose $\rho = 1$ and .5 and $r = .5$ and 1, and set the corresponding values of (α, σ^2) to $(.2, 1)$, $(.2, 2)$, $(.5, 4)$, and $(.5, 4)$, which yield, on average, 1.05, .98, 1.81, and 1.24 observed events per subject.

We used the proposed EM algorithm to calculate the NPMLEs and used the inverse of the observed information matrix to estimate the variances. We set the initial value of $\boldsymbol{\beta}$ to $\mathbf{0}$ and that of σ^2 to 1. In addition, we set the initial values for the jump sizes of Λ to $1/n$. For each combination of the simulation parameters, we generated 1,000 replicates of data. It took less than 2 days on an IBM BladeCenter HS20 machine to complete all of the simulation studies given in Table 1. No convergence problem was encountered in any simulation run.

Because Λ is nonnegative, we used the log-transformation in constructing its confidence interval. Specifically, the 95% confidence interval for $\Lambda(t)$ is $\widehat{\Lambda}_n(t) \exp[\pm 1.96 \times \widehat{\text{se}}\{\log(\widehat{\Lambda}_n(t))\}]$, where $\widehat{\text{se}}$ denotes the estimated standard error. In addition, we used the Satterthwaite approximation (e.g., Burdick and Graybill 1992) to construct the 95% confidence interval for σ^2: $(\nu\widehat{\sigma}_n^2/\chi^2_{\nu,.975}, \nu\widehat{\sigma}_n^2/\chi^2_{\nu,.025})$, where $\nu = 2\{\widehat{\sigma}_n^2/\widehat{\text{se}}(\widehat{\sigma}_n^2)\}^2$ and $\chi^2_{\nu,\alpha}$ is the α-quantile of the chi-squared distribution with ν degrees of freedom.

As is evident in Table 1, the NPMLEs for $\boldsymbol{\beta}_0$ and $\Lambda_0(\cdot)$ are virtually unbiased, the variance estimators accurately reflect the true variations, and the confidence intervals achieve proper coverages. As expected, the asymptotic normal approximation for $\widehat{\sigma}_n^2$ is not very accurate in small samples, although the proposed inference procedures appear to perform reasonably well, at least for $n = 400$.

To assess the consequences of misspecifying the transformation, we generated recurrent events from the foregoing proportional odds model but fit the data with the proportional intensity model. The biases of the estimators for β_1 and β_2 are approximately .25 and $-.5$. The predicted numbers of events are also biased, especially for large t.

In the second set of simulation studies, we generated data in the same manner as in the first set but estimated the transformation parameters along with other parameters and accounted for extra variation in the variance estimation. The results are summarized in Table 2. As expected, it is difficult to estimate the transformation parameters with good precision in moderate samples. Treating the transformation parameters as unknown versus known appears to have more impact on the estimation of the variance components than on that of the other parameters.

4. EXAMPLES

We first consider a randomized clinical trial conducted to assess the efficacy of rhDNase, a highly purified recombinant enzyme, in reducing exacerbations of respiratory symptoms for patients with cystic fibrosis (Therneau and Hamilton 1997). A total of 321 patients were assigned to rhDNase and 324 were assigned to placebo. Most patients were followed for approximately 170 days. By the end of follow-up, 65 patients in the rhDNase group experienced 1 exacerbation and 39 experienced

Table 1. Summary Statistics for the Simulation Studies With Fixed Transformations

			$n = 200$				$n = 400$			
$G(x)$	Parameter	True value	Bias	SE	SEE	CP	Bias	SE	SEE	CP
$\rho = 1$	β_1	$-.5$	$-.010$	.315	.300	.937	$-.005$	.219	.212	.932
	β_2	1.0	.015	.206	.204	.941	$-.002$	.144	.141	.943
	σ^2	1.0	$-.029$	.245	.253	.965	$-.009$	.160	.176	.960
	$\Lambda(\tau/4)$	.139	.001	.034	.033	.932	.002	.024	.023	.947
	$\Lambda(\tau/2)$	.220	.002	.052	.050	.940	.003	.036	.036	.950
	$\Lambda(\tau)$	.322	.003	.076	.072	.933	.004	.052	.052	.946
$\rho = .5$	β_1	$-.5$	.007	.382	.387	.956	$-.007$	.277	.274	.939
	β_2	1.0	$-.001$	.255	.251	.956	.004	.180	.178	.943
	σ^2	2.0	$-.010$	.486	.467	.972	$-.048$	.350	.332	.938
	$\Lambda(\tau/4)$	.139	.008	.043	.042	.938	.004	.031	.029	.932
	$\Lambda(\tau/2)$	.220	.011	.065	.065	.939	.007	.048	.045	.936
	$\Lambda(\tau)$	.322	.005	.095	.093	.946	.008	.069	.065	.940
$r = .5$	β_1	$-.5$	$-.026$	.456	.454	.949	.007	.317	.322	.948
	β_2	1.0	.017	.278	.287	.957	$-.008$	.204	.203	.950
	σ^2	4.0	$-.134$	.787	.773	.958	$-.056$	.543	.554	.962
	$\Lambda(\tau/4)$	.347	.020	.115	.110	.944	.010	.078	.077	.953
	$\Lambda(\tau/2)$	.549	.028	.180	.170	.940	.015	.123	.119	.947
	$\Lambda(\tau)$	.805	.039	.261	.247	.944	.021	.180	.173	.938
$r = 1$	β_1	$-.5$	$-.004$	.510	.492	.954	$-.002$	.343	.353	.955
	β_2	1.0	.006	.317	.312	.942	$-.001$	.221	.223	.957
	σ^2	4.0	$-.287$	.962	.951	.966	$-.094$	.706	.699	.953
	$\Lambda(\tau/4)$	.347	.025	.128	.122	.940	.010	.088	.085	.946
	$\Lambda(\tau/2)$	.549	.038	.201	.199	.937	.014	.137	.131	.945
	$\Lambda(\tau)$	.805	.054	.291	.272	.932	.019	.201	.189	.941

NOTE: Bias and SE correspond to the bias and standard error of the NPMLE; SEE, to the mean of the standard error estimator; and CP, to the coverage probability of the 95% confidence interval.

at least 2 exacerbations; in the placebo group, 97 patients experienced 1 exacerbation and 42 experienced at least 2 exacerbations.

The solid step curve in Figure 1 is the ratio of the nonparametric estimates of mean frequencies of exacerbations between the placebo and rhDNase groups. In view of the decreasing

Table 2. Summary Statistics for the Simulation Studies With Estimated Transformations

			$n = 200$				$n = 400$			
$G(x)$	Parameter	True value	Bias	SE	SEE	CP	Bias	SE	SEE	CP
$\rho = 1$	β_1	$-.5$	.014	.285	.305	.956	.015	.199	.216	.960
	β_2	1.0	$-.001$	.221	.246	.955	$-.009$	.160	.171	.952
	σ^2	1.0	$-.007$	.390	.372	.899	$-.001$	.280	.261	.891
	ρ	1.0	.100	.465	.402	.856	.052	.286	.251	.848
	$\Lambda(\tau/4)$	.139	$-.001$	.034	.035	.952	.001	.025	.025	.938
	$\Lambda(\tau/2)$	.220	$-.003$	.049	.051	.956	$-.001$	.035	.036	.943
	$\Lambda(\tau)$	.322	$-.004$	.074	.073	.956	$-.002$	.049	.051	.955
$\rho = .5$	β_1	$-.5$	.014	.379	.391	.959	.007	.279	.280	.949
	β_2	1.0	$-.011$	.279	.281	.941	$-.004$	.193	.201	.954
	σ^2	2.0	$-.009$	.720	.722	.945	$-.002$	.514	.954	.940
	ρ	.5	.066	.260	.261	.907	.020	.159	.165	.926
	$\Lambda(\tau/4)$	.139	.007	.045	.045	.934	.002	.031	.033	.953
	$\Lambda(\tau/2)$	.220	.010	.065	.065	.940	.003	.046	.047	.942
	$\Lambda(\tau)$	.322	.009	.093	.092	.954	.005	.065	.065	.946
$r = .5$	β_1	$-.5$	.015	.440	.450	.956	.014	.315	.320	.956
	β_2	1.0	$-.021$	.299	.298	.941	$-.018$	.217	.212	.940
	σ^2	4.0	$-.232$	1.078	1.095	.960	$-.167$	.776	.788	.956
	r	.5	$-.030$	.121	.129	.965	$-.015$	.093	.093	.926
	$\Lambda(\tau/4)$	.347	.016	.111	.107	.946	.011	.078	.076	.954
	$\Lambda(\tau/2)$	.549	.015	.172	.165	.941	.013	.120	.117	.949
	$\Lambda(\tau)$	.805	.018	.256	.244	.944	.015	.179	.174	.946
$r = 1$	β_1	$-.5$	.038	.490	.479	.953	.015	.345	.348	.958
	β_2	1.0	$-.070$	.332	.333	.924	$-.037$	.248	.244	.929
	σ^2	4.0	$-.552$	1.352	1.442	.960	$-.333$	1.067	1.127	.962
	r	1.0	$-.132$	.293	.329	.967	$-.072$	.240	.253	.964
	$\Lambda(\tau/4)$	.347	.021	.117	.116	.964	.010	.085	.083	.951
	$\Lambda(\tau/2)$	.549	.015	.185	.184	.967	.010	.139	.133	.953
	$\Lambda(\tau)$	.805	.001	.286	.285	.961	.005	.220	.210	.943

NOTE: See the note to Table 1.

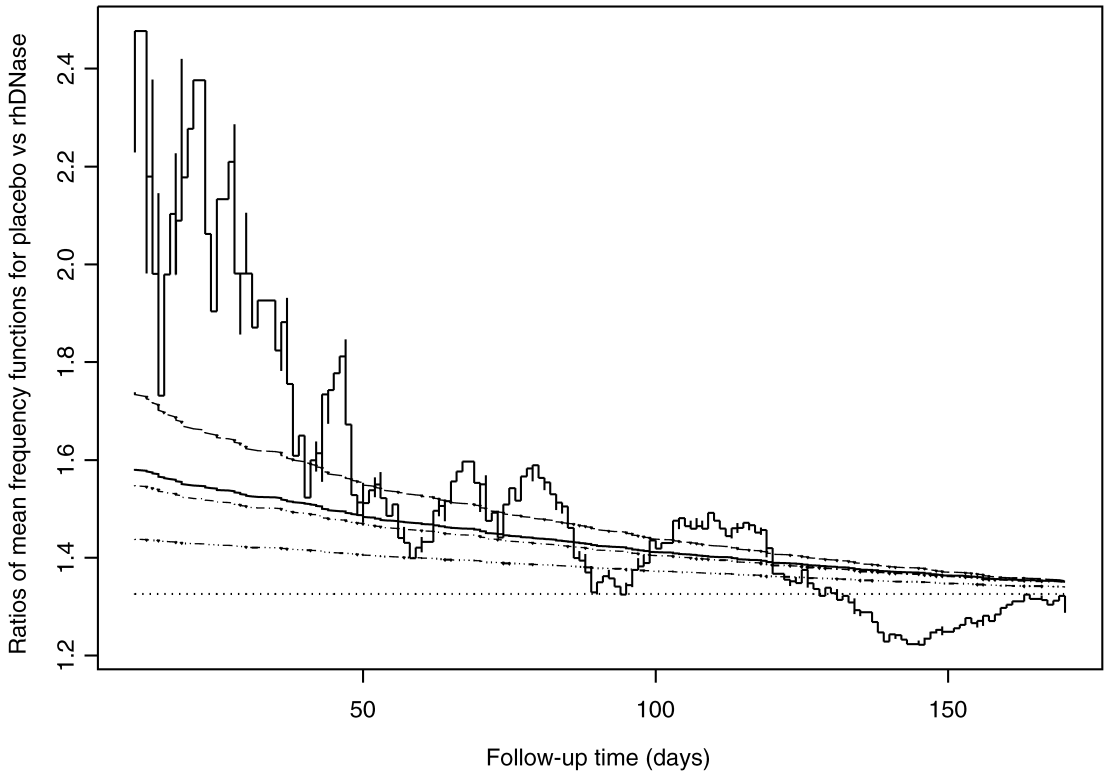

Figure 1. Ratio of the Mean Frequency Functions for the Placebo Group versus the rhDNase Group. The solid step curve pertains to the nonparametric estimates. The other curves from the bottom up pertain to the model-based estimates for r = 0, .5, 1, 1.181, and 2. The values of r = 0 and 1 correspond to the proportional intensity and proportional odds models.

trend over time, we fit the transformation intensity models with a normal random effect and with the treatment indicator and baseline level of forced expiratory volume per second (FEV_1) as the covariates. Figure 2 shows the log-likelihood values under the Box–Cox and logarithmic transformations, whereas Figure 1 plots the ratios of the predicted frequencies of exacerbations between the placebo and rhDNase groups under various transformations. Table 3 summarizes the estimation results. The standard error estimates in the last column accounts for the variation due to estimation of the transformation parameter.

By the Akaike information criterion (AIC) (Akaike 1985), which is tantamount to the log-likelihood value in this setting, one would select the logarithmic transformation with $r = 1.181$. Because $r = 1$ is fairly close to 1.181 (in terms not only of numerical value, but also of the corresponding log-likelihood value), and corresponds to the well-known proportional odds model for survival data (Bennett 1983; Pettitt 1984), it is reasonable to choose $r = 1$ for simplicity of interpretation. The choice of $r = 0$ (i.e., $\rho = 1$), which corresponds to the proportional intensity model, would clearly contradict the fact (shown in Fig. 1) that the treatment effect diminishes over time rather than staying constant, whereas the choice of $r = 1$ reflects this time trend. As shown in Table 3, the prediction error is much lower under the proportional odds model than under the proportional intensity model.

Figure 3 displays the estimated conditional survival probabilities for the time to the second exacerbation given that the first exacerbation occurs on days 120 for patients with specific characteristics under the proportional odds and proportional intensity models. The two models lead to quite different predictions.

As a second example, we use the recurrent infection data from the chronic granulomatous disease (CGD) study described by Lin, Wei, Yang, and Ying (2000). This study enrolled 128 patients with CGD, of whom 63 received gamma interferon and 65 received placebo. Data were collected on all serious pyogenic infections occurring up to cessation of follow-up, 400 days for most patients. During the study, 14 patients receiving gamma interferon and 30 receiving placebo experienced at least 1 infection. In the gamma interferon group, four patients had two infections and one had a third infection. In the placebo group, five patients had two infections, four patients had three infections, and three patients had four or more infections.

Under the Andersen–Gill proportional intensity model without frailty, the estimated effect of gamma interferon is at -1.097, with (estimated) model-based SE of .261. The robust SE (Lin et al. 2000), which accounts for the dependence of recurrent infections, is .311. To model the intraclass correlation, we add a time-dependent covariate indicating the occurrence of infections within the last 60 days. Then the estimated treatment effect is reduced to $-.989$, with model-based and robust SEs of .266 and .294. The difference between the two SEs seem to suggest that the time-dependent covariate may not adequately capture the intraclass correlation. Consequently, we consider random-effects models.

We fit the proportional intensity models with normal and gamma random effects in S–PLUS. The covariates include the treatment indicator and age. The estimated treatment effects are

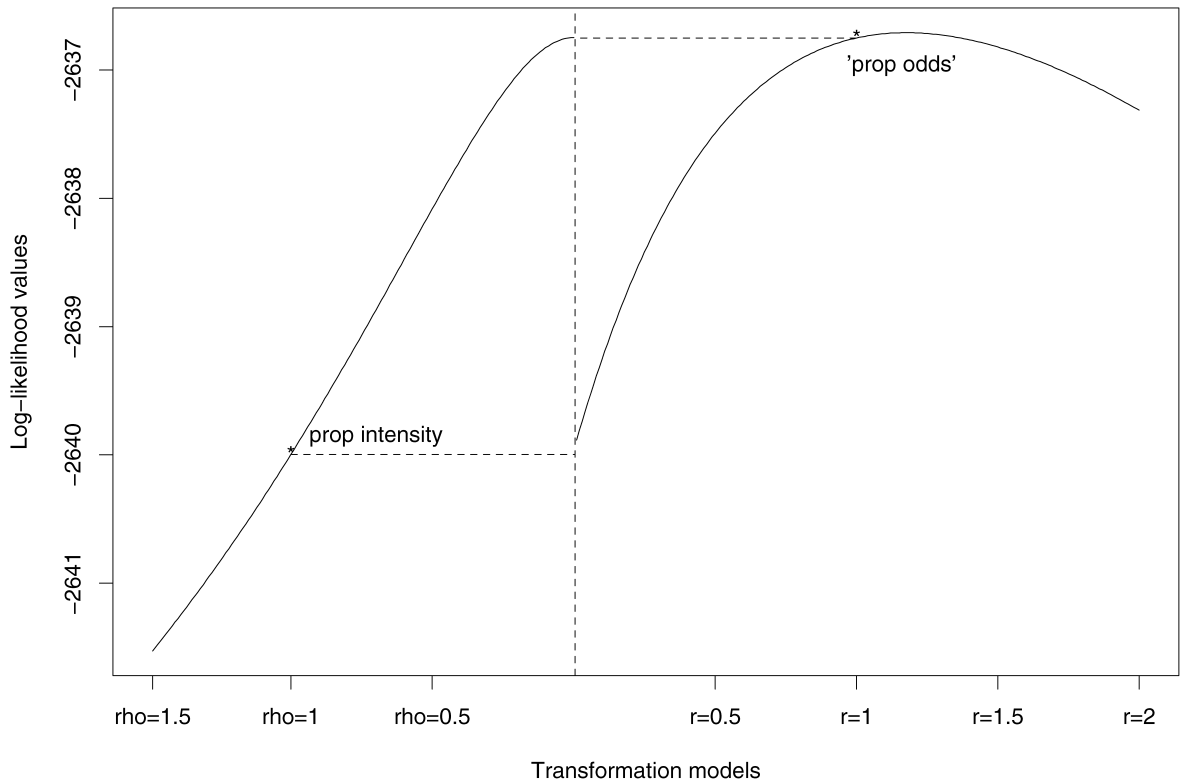

Figure 2. The Log-Likelihood Values Under Various Transformation Models for the Cystic Fibrosis Study. The left panel pertains to the Box–Cox transformations $G(x) = \{(1+x)^{\rho} - 1\}/\rho$; the right panel, to the logarithmic transformations $G(x) = \log(1+rx)/r$. The horizontal dashed lines indicate that the right limit of the Box–Cox family (i.e., $\rho = 0$) pertains to the proportional odds model (i.e., $r = 1$ in the logarithmic family) and the left limit of the logarithmic family (i.e., $r = 0$) pertains to the proportional intensity model (i.e., $\rho = 1$ in the Box–Cox family). The curves are based on grids of size .001.

-1.018 with a SE of .298 and -1.051 with a SE of .308 under the normal and gamma random-effects distributions. Although the estimates of regression parameters are very similar under these two random-effects distributions, the predictions of recurrent events can be very different. For a placebo subject age 2 years, the predicted cumulative frequencies at days 100, 200, and 300 are .375, .739, and 1.510 under the normal random effect versus .459, .908, and 1.865 under the gamma random effect; for a placebo patient age 7 years, the corresponding predicted values are .072, .142, and .300 under the normal random effect, compared with .050, .099, and .210 under the gamma random effect.

Table 4 summarizes the results for the normal random-effect models with the Box–Cox and logarithmic transformations. The model with $r = .347$ yields the highest likelihood and smallest prediction error.

Table 3. Estimation Results for the Cystic Fibrosis Study

	$G(x) = \{(1+x)^{\rho} - 1\}/\rho$			
	ρ *held fixed*			ρ *estimated*
ρ	2	1	.5	$.013_{(.231)}$
Treatment	$-.216_{(.099)}$	$-.280_{(.123)}$	$-.341_{(.143)}$	$-.444_{(.186)}$
FEV_1	$-.013_{(.002)}$	$-.017_{(.003)}$	$-.020_{(.003)}$	$-.025_{(.005)}$
σ^2	$.258_{(.079)}$	$.439_{(.126)}$	$.643_{(.182)}$	$.998_{(.359)}$
Log-likelihood	$-2{,}642.7$	$-2{,}640.0$	$-2{,}638.0$	$-2{,}636.7$
MSE ($\times 10^{-4}$)	2.203	1.673	1.511	1.371
	$G(x) = \log(1+rx)/r$			
	r held fixed			*r estimated*
r	.5	1	2	$1.181_{(.642)}$
Treatment	$-.365_{(.151)}$	$-.449_{(.176)}$	$-.602_{(.223)}$	$-.477_{(.211)}$
FEV_1	$-.021_{(.003)}$	$-.025_{(.004)}$	$-.033_{(.005)}$	$-.027_{(.006)}$
σ^2	$.728_{(.212)}$	$1.005_{(.304)}$	$1.597_{(.524)}$	$1.113_{(.502)}$
Log-likelihood	$-2{,}637.5$	$-2{,}636.7$	$-2{,}637.3$	$-2{,}636.6$
MSE ($\times 10^{-4}$)	1.471	1.400	1.237	1.344

NOTE: The standard error estimates are given in parentheses. MSE is the mean squared error comparing the nonparametric and model-based estimates of cumulative frequency functions.

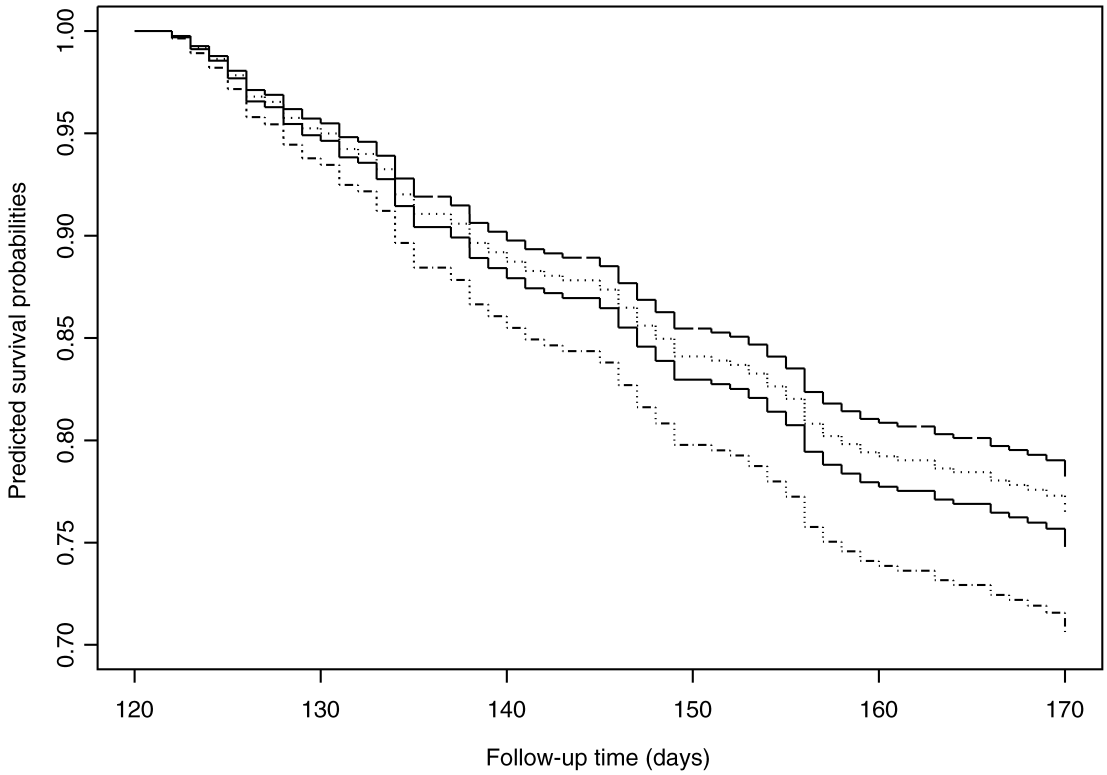

Figure 3. Estimated Conditional Survival Probabilities for the Cystic Fibrosis Patients. The dashed and dotted curves pertain to treated patients with FEV_1 of 16 under the proportional odds model and proportional hazards model. The solid and dashed–dotted curves pertain to untreated patients with FEV_1 of 16 under the proportional odds model and proportional hazards model.

5. DISCUSSION

We have developed a general asymptotic theory, together with a simple and stable numerical algorithm, for the maximum likelihood estimation in a broad class of transformation intensity models with random effects. There is considerable flexibility in choosing the transformation and random-effects distribution. Therneau and Grambsch (2000) proposed fitting the proportional intensity model with normal random effects by the penalized partial likelihood method, but provided no formal theoretical justifications. The current version of S–PLUS allows one to fit the proportional intensity model with normal or gamma random effects. We have developed a MATLAB program to fit the class of models studied in this article.

In our data examples, we used the AIC to determine the best transformation. Other criteria for model selection include the Bayes information criterion (BIC) and likelihood-based cross-validation (van der Laan, Dudoit, and Keles 2004). But none of these criteria takes into consideration the complexity of the transformation. For example, if the transformation $G(x) = \log\{1 + \log(1 + x)\}$ provides a slightly better fit than $G(x) = \log(1 + x)$, one may still prefer the latter transformation

Table 4. Estimation Results for the CGD Study

	$G(x) = \{(1+x)^\rho - 1\}/\rho$			
	ρ *held fixed*			ρ *estimated*
ρ	2	1	.5	$.334_{(.402)}$
Treatment	$-.840_{(.251)}$	$-1.067_{(.311)}$	$-1.282_{(.367)}$	$-1.387_{(.485)}$
Age	$-.026_{(.013)}$	$-.032_{(.016)}$	$-.038_{(.020)}$	$-.041_{(.022)}$
σ^2	$.328_{(.188)}$	$.593_{(.308)}$	$.944_{(.467)}$	$1.141_{(.788)}$
Log-likelihood	−397.14	−396.35	−395.88	−395.82
MSE($\times 10^{-3}$)	1.724	1.449	1.341	1.343
	$G(x) = \log(1+rx)/r$			
	r held fixed			*r estimated*
r	.5	1	2	$.347_{(.393)}$
Treatment	$-1.387_{(.398)}$	$-1.659_{(.474)}$	$-2.137_{(.621)}$	$-1.297_{(.445)}$
FEV_1	$-.041_{(.021)}$	$-.047_{(.025)}$	$-.058_{(.032)}$	$-.038_{(.021)}$
σ^2	$1.166_{(.592)}$	$1.662_{(.887)}$	$2.762_{(1.610)}$	$1.004_{(.659)}$
Log-likelihood	−395.76	−396.39	−398.09	−395.70
MSE($\times 10^{-3}$)	1.324	1.596	2.500	1.293

NOTE: See the note to Table 3.

because of its simplicity and interpretability. Thus incorporating the complexity of the transformation into the model selection process is an interesting open problem.

Our work can be viewed as a generalization of the transformation models for univariate survival data studied by Kosorok et al. (2004) to correlated recurrent event times. Our conditions on transformation are similar to those authors' conditions D1 and D2; however, we do not assume their concavity of the transformation or restrictive condition E1. Due to the lack of concavity, presence of random effects, and unbounded number of events, our technical developments are more delicate, especially in proving the consistency, model identifiability, and invertibility of the information matrix.

In some applications, one is interested in testing whether the variance of a random effect is zero. Then the hypothesized parameter value lies on the boundary of the parameter space, which violates Condition 1. However, the conclusions of Theorems 1–3 continue to hold if the log-likelihood function (4) allows an extended definition beyond the boundary and after the extension the log-likelihood function still satisfies the regularity conditions. We note in Remark A.1 in Appendix A that the extension is indeed valid for gamma frailty. We believe that in general the likelihood ratio statistic for testing zero variance has a mixture of chi-squared distributions asymptotically. This conjecture is supported by our simulation studies; a formal derivation is underway.

Following the arguments of Kosorok et al. (2004), we can show that the information operator in a neighborhood of the true parameter value is invertible. This fact, together with the uniqueness of the Kullback–Leibler maximizer, allows us to make inference under misspecified transformations. The details are omitted here.

A number of authors, including Pepe and Cai (1993), Lawless and Nadeau (1995), Lawless, Nadeau, and Cook (1997), and Lin et al. (2000), have advocated the proportional mean/rate model, under which

$$E\{dN^*(t)|\mathbf{X}(t)\} = e^{\boldsymbol{\beta}^T\mathbf{X}(t)}\,d\mu_0(t),$$

where $\mu_0(t)$ is an unknown continuous function. For time-independent covariates, Lin, Wei, and Ying (2001) proposed a class of transformation models for the mean function

$$E\{N^*(t)|\mathbf{X}\} = G\big(\mu_0(t)e^{\boldsymbol{\beta}^T\mathbf{X}}\big),$$

where $G(\cdot)$ is a transformation function. These marginal models cannot be used to make predictions about future recurrences based on individual event histories shown in Figure 3. Furthermore, the existing estimators for these models are not efficient, especially outside of the proportional mean/rate model. For the proportional intensity/rate model studied in Section 3, the efficiencies of the maximum partial likelihood estimators of β_1 and β_2 relative to the NPMLEs were found to be approximately .9 and .85 under $\sigma^2 = 1$. The efficiency loss becomes more substantial as σ^2 increases.

In some applications, censoring depends on the underlying recurrent event process even after conditioning on the covariates in the model. One possible way to adjust for such dependent censoring is to postulate a proportional hazards model or, more generally, a transformation model for time to dependent censoring that shares the random effects of (3). The joint model for recurrent events and dependent censoring can be estimated by the nonparametric maximum likelihood method, and the resultant estimators are consistent, asymptotically normal and asymptotically efficient. We will communicate these results in a separate report.

For the proportional rate model, Miloslavski, Keles, and van der Laan (2004) proposed adjusting for dependent censoring through the inverse-probability-of-censoring weighting. This approach requires that censoring depend only on the observed (possibly time-dependent) covariates, some of which may not be included in the model for recurrent events. The estimating equation may be numerically unstable under heavy censoring. Computing the efficient estimator is difficult due to the implicit nature of the efficient influence function. Generalization of this approach beyond the proportional rate model is unclear.

Counting process models characterize the effects of covariates on the development of recurrent events in a parsimonious and efficient manner. These models require that the recurrent events be of the same nature. In some medical applications, the second recurrent event may be biologically different from the first recurrence. Then it is more appropriate to model each recurrence separately. Several methods are available for doing so (see Wei, Lin, and Weissfeld 1989; Prentice, Williams, and Peterson 1981; Schaubel and Cai 2004). It would be worthwhile to study random-effects models for such data.

APPENDIX A: PROOF OF ASYMPTOTIC RESULTS

Here we sketch the proofs of Theorems 1–4. Detailed proofs are available in a separate technical report.

A.1 Proof of Theorem 1

Let $O(1)$ denote some positive constant and define $M = \sup_t\{\sup_{\mathbf{X},\boldsymbol{\beta}}|\boldsymbol{\beta}^T\mathbf{X}(t)| + \sup_{\mathbf{Z}}|\mathbf{Z}(t)|\}$. Write $Q(t,\mathbf{O},\mathbf{b};\Lambda,\boldsymbol{\beta}) = \int_0^t Y(s)e^{\boldsymbol{\beta}^T\mathbf{X}(s)+\mathbf{b}^T\mathbf{Z}(s)}\,d\Lambda$, where $\mathbf{O}$ represents the data.

Step 1. We prove the existence of NPMLEs. Under Condition 4, the integrand in the ith term of $l_n(\Lambda,\boldsymbol{\theta})$ is dominated by

$$\begin{aligned}&O(1)e^{M|\mathbf{b}|N_i(\tau)}\mu(N_i(\tau))\\&\quad\times\prod_{t\le\tau}\big[\Lambda\{t\}\{1+Q(t,\mathbf{O}_i,\mathbf{b};\Lambda,\boldsymbol{\beta})\}^{-1}\big]^{\Delta N_i(t)}\\&\quad\times\{1+Q(\tau,\mathbf{O}_i,\mathbf{b};\Lambda,\boldsymbol{\beta})\}^{-\alpha}\psi(\mathbf{b};\boldsymbol{\gamma}).\end{aligned}$$

Thus $l_n(\Lambda,\boldsymbol{\theta})$ attains the maximum for finite Λ values, so the NPMLEs exist.

Step 2. We show that $\sup_n\widehat{\Lambda}_n(\tau) < \infty$ with probability 1. By setting the derivative of $l_n(\Lambda,\boldsymbol{\theta})$ with respect to $\Lambda\{T_{ij}\}$ to 0, we obtain $\widehat{\Lambda}_n(t) = n^{-1}\int_0^t\sum_{i=1}^n dN_i(s)/|\phi_n(s;\widehat{\Lambda}_n,\widehat{\boldsymbol{\theta}}_n)|$ with

$$\phi_n(s;\Lambda,\boldsymbol{\theta}) = \frac{1}{n}\sum_{k=1}^n\frac{\int_{\mathbf{b}}R_1(\mathbf{b},\mathbf{O}_k;\Lambda,\boldsymbol{\theta})R_2(s,\mathbf{b},\mathbf{O}_k;\Lambda,\boldsymbol{\theta})\psi(\mathbf{b};\boldsymbol{\gamma})\,d\mathbf{b}}{\int_{\mathbf{b}}R_1(\mathbf{b},\mathbf{O}_k;\Lambda,\boldsymbol{\theta})\psi(\mathbf{b};\boldsymbol{\gamma})\,d\mathbf{b}}, \quad (A.1)$$

where

$$\begin{aligned}&R_1(\mathbf{b},\mathbf{O};\Lambda,\boldsymbol{\theta})\\&\quad=\prod_{t\le\tau}\big\{Y(t)e^{\boldsymbol{\beta}^T\mathbf{X}(t)+\mathbf{b}^T\mathbf{Z}(t)}G'(Q(t,\mathbf{O},\mathbf{b};\Lambda,\boldsymbol{\beta}))\big\}^{\Delta N(t)}\\&\qquad\times e^{-G(Q(\tau,\mathbf{O},\mathbf{b};\Lambda,\boldsymbol{\beta}))}\end{aligned}$$

and

$$R_2(s,\mathbf{b},\mathbf{O};\Lambda,\boldsymbol{\theta}) = -\left\{\int I(t\ge s)\frac{G''(Q(t,\mathbf{O},\mathbf{b};\Lambda,\boldsymbol{\beta}))}{G'(Q(t,\mathbf{O},\mathbf{b};\Lambda,\boldsymbol{\beta}))}\,dN(t) - G'(Q(\tau,\mathbf{O},\mathbf{b};\Lambda,\boldsymbol{\beta}))\right\}Y(s)e^{\boldsymbol{\beta}^T\mathbf{X}(s)+\mathbf{b}^T\mathbf{Z}(s)}.$$

Replacing $\widehat{\Lambda}_n$ and $\widehat{\boldsymbol{\theta}}_n$ on the right side of the equation for $\widehat{\Lambda}_n$ by Λ_0 and $\boldsymbol{\theta}_0$, we obtain a similar function denoted by $\widetilde{\Lambda}_n(t)$. It follows from the Glivenko–Cantelli theorem that $\widetilde{\Lambda}_n$ converges uniformly to Λ_0 almost surely.

Clearly, $n^{-1}\{l_n(\widehat{\Lambda}_n,\widehat{\boldsymbol{\theta}}_n) - l_n(\widetilde{\Lambda}_n,\boldsymbol{\theta}_0)\}\ge 0$. We show that if $\widehat{\Lambda}_n(\tau)$ diverges, then the left side of this inequality cannot be nonnegative when n is large. By Condition 5 and the fact that $e^{-|x|}(1+y)\le(1+e^x y)\le e^{|x|}(1+y)$,

$$l_n(\widehat{\Lambda}_n,\widehat{\boldsymbol{\theta}}_n)\le O(1)+\sum_{i=1}^n\int_0^\tau Y_i(t)\log\frac{\widehat{\Lambda}_n\{t\}}{1+\int_0^t Y_i(s)\,d\widehat{\Lambda}_n(s)}\,dN_i(t) - \sum_{i=1}^n\alpha\log\left\{1+\int_0^\tau Y_i(t)\,d\widehat{\Lambda}_n(t)\right\}.$$

Therefore,

$$0\le O(1)+\frac{1}{n}\sum_{i=1}^n\int_0^\tau Y_i(t)\log\frac{n\widehat{\Lambda}_n\{t\}}{1+\int_0^t Y_i(s)\,d\widehat{\Lambda}_n(s)}\,dN_i(t) - \frac{\alpha}{n}\sum_{i=1}^n\log\left\{1+\int_0^\tau Y_i(t)\,d\widehat{\Lambda}_n(t)\right\}. \quad (A.2)$$

The right side of (A.2) will diverge to $-\infty$ if $\widehat{\Lambda}_n(\tau)$ diverges. The proof of this claim is based on the partitioning idea of Murphy (1994). Specifically, we construct a sequence $t_0=\tau>t_1>t_2>\cdots>t_{m0}=0$ and denote $I_q=[t_{q+1},t_q)$. The right side of (A.2) can be bounded by

$$\begin{aligned}O(1)\le\Bigg(&-n^{-1}\sum_{i=1}^n\frac{\alpha}{2}I(Y_i(t_0)>0)\log\{1+\widehat{\Lambda}_n(\tau)\}\\&-\left\{n^{-1}\sum_{i=1}^n\frac{\alpha}{2}I(Y_i(t_0)>0)\right.\\&\left.-n^{-1}\sum_{i=1}^n I\big(Y_i(t_0)=0,Y_i(t_1)>0\big)\int_{t\in I_0}dN_i^*\right\}\\&\times\log\{1+\widehat{\Lambda}_n(t_0)\}\\&-\sum_{q=1}^{m_0-1}\left[n^{-1}\sum_{i=1}^n I\big(Y_i(t_{q-1})=0,Y_i(t_q)>0\big)\left\{\alpha+\int_{t\in I_q}dN_i^*\right\}\right.\\&\left.-n^{-1}\sum_{i=1}^n I\big(Y_i(t_q)=0,Y_i(t_{q+1})>0\big)\int_{t\in I_q}dN_i^*\right]\\&\times\{1+\log\widehat{\Lambda}_n(t_q)\}\Bigg).\end{aligned}$$

The t_q's are chosen such that the coefficients in front of $\log\widehat{\Lambda}_n(t_q)$ are all negative when n is large enough. Thus the corresponding terms cannot diverge to ∞. However, if $\widehat{\Lambda}_n(\tau)\to\infty$, then the right side diverges to $-\infty$. We conclude that $\limsup_n\widehat{\Lambda}_n(\tau)<\infty$. It then follows from Helly's selection theorem that along a subsequence, $\widehat{\Lambda}_n\to\Lambda^*$ and $\widehat{\boldsymbol{\theta}}_n\to\boldsymbol{\theta}^*$.

Step 3. We show that $\Lambda^*=\Lambda_0$ and $\boldsymbol{\theta}^*=\boldsymbol{\theta}_0$. By the Glivenko–Cantelli theorem, $\phi_n(s;\widehat{\Lambda}_n,\widehat{\boldsymbol{\theta}}_n)$ given in (A.1) converges uniformly to a continuously differentiable function $\phi^*(s;\Lambda^*,\boldsymbol{\theta}^*)$. We show that $\min_{s\in[0,\tau]}|\phi^*(s;\Lambda^*,\boldsymbol{\theta}^*)|\ge 2\epsilon_0$ for some positive constant by contradiction. If this inequality does not hold, then $\phi^*(s_0;\Lambda^*,\boldsymbol{\theta}^*)=0$ for some $s_0\in[0,\tau]$. For any $\epsilon>0$,

$$\widehat{\Lambda}_n(\tau)\ge\int_0^\tau\frac{\sum_{i=1}^n I(C_i\ge s)\,dN_i(s)/n}{|\phi_n(s;\widehat{\Lambda}_n,\widehat{\boldsymbol{\theta}}_n)|+\epsilon}\to E\left[\int_0^\tau\frac{I(C\ge s)\,dN(s)}{|\phi^*(s;\Lambda^*,\boldsymbol{\theta}^*)|+\epsilon}\right].$$

Letting ϵ decrease to 0, we obtain $E[\int_0^\tau I(C\ge s)\,dN(s)/|\phi^*(s;\Lambda^*,\boldsymbol{\theta}^*)|]<\infty$. However, $|\phi^*(s;\Lambda^*,\boldsymbol{\theta}^*)|=|\phi^*(s;\Lambda^*,\boldsymbol{\theta}^*)-\phi^*(s_0;\Lambda^*,\boldsymbol{\theta}^*)|\le c_1|s-s_0|$ for some constant c_1, and $\int_0^\tau|s-s_0|^{-1}E[I(C\ge s)\,dN(s)]=\infty$. This is a contradiction. Thus, when n is large enough, $|\phi_n(t;\widehat{\Lambda}_n,\widehat{\boldsymbol{\theta}}_n)|>\epsilon_0>0$ for some constant ϵ_0.

The lower bound of $|\phi_n|$ implies that $\widehat{\Lambda}_n(t)$ is absolutely continuous with respect to $\widetilde{\Lambda}_n(t)$ and $d\widehat{\Lambda}_n/d\widetilde{\Lambda}_n$ converges to a bounded measurable function $\zeta(t)$; that is, $\Lambda^*(t)=\int_0^t\zeta(s)\,d\Lambda_0(t)$. Thus $\Lambda^*(t)$ is absolutely continuous with respect to the Lebesgue measure, and we denote its derivative by $\lambda^*(t)$. Furthermore, $\zeta(t)=\lambda^*(t)/\lambda_0(t)$. Finally, because $(\widehat{\Lambda}_n,\widehat{\boldsymbol{\theta}}_n)$ maximizes $l_n(\Lambda,\boldsymbol{\theta})$,

$$\frac{1}{n}\sum_{i=1}^n\left[\int_0^\tau\log\frac{\widehat{\Lambda}_n\{t\}}{\widetilde{\Lambda}_n(t)}\,dN_i(t) + \log\frac{\int_{\mathbf{b}}R_1(\mathbf{b},\mathbf{O}_i;\widehat{\Lambda}_n,\widehat{\boldsymbol{\theta}}_n)\psi(\mathbf{b};\widehat{\boldsymbol{\gamma}}_n)\,d\mathbf{b}}{\int_{\mathbf{b}}R_1(\mathbf{b},\mathbf{O}_i;\widetilde{\Lambda}_n,\boldsymbol{\theta}_0)\psi(\mathbf{b};\boldsymbol{\gamma}_0)\,d\mathbf{b}}\right]\ge 0.$$

We take the limits on both sides. By the Glivenko–Cantelli theorem and the fact that $\widehat{\Lambda}_n\{t\}/\widetilde{\Lambda}\{t\}$ converges uniformly to $\lambda^*(t)/\lambda_0(t)$, we conclude that the Kullback–Leibler information between the density indexed by $(\Lambda^*,\boldsymbol{\theta}^*)$ and the true density is negative; therefore, the two densities are equal almost surely. We show in Section A.4 that this finding implies that $\Lambda^*=\Lambda_0$ and $\boldsymbol{\theta}^*=\boldsymbol{\theta}_0$. Thus we have proven that $\widehat{\boldsymbol{\theta}}_n\to\boldsymbol{\theta}_0$ and $\widehat{\Lambda}_n(t)\to\Lambda_0(t)$ almost surely. The latter convergence can be strengthened to uniform convergence in $t\in[0,\tau]$ by the continuity of Λ_0.

Remark A.1. Inequality (A.2) is the key to the derivation of the boundedness of $\widehat{\Lambda}_n$ and thus to the proof of consistency. This inequality holds naturally for the classical gamma frailty model, under which $G(x)=x$ and $\exp\{\mathbf{b}^T\mathbf{Z}_i(t)\}$ is replaced by a gamma variable ξ_i. The reason for this is as follows. Suppose that ν is the variance of ξ_i, which is assumed to be bounded by $\nu_{\max}$. By integrating out the frailty, we obtain

$$l_n(\Lambda,\boldsymbol{\theta})\le\sum_{i=1}^n O(1)\log\frac{\Gamma(N_i(\tau)+1/\nu_{\max})\nu_{\max}^{N_i(\tau)}}{\Gamma(1/\nu_{\max})} + \sum_{i=1}^n\int_0^\tau Y_i(t)\log\Lambda\{t\}\,dN_i(t) - \sum_{i=1}^n\left(\frac{1}{\nu}+N_i(\tau)\right)\log\left\{1+\nu\int_0^\tau Y_i(s)\,d\Lambda(s)\right\}.$$

Thus it follows from the inequality

$$\begin{aligned}&(1+\nu x)^{N_i(\tau)+1/\nu}\\&\quad\ge(N_i(\tau)+1)^{-1-N_i(\tau)}\\&\qquad\times\min\big\{(1+x)^{N_i(\tau)}(1+\nu_{\max}x)^{1/\nu_{\max}},(1+x)^{N_i(\tau)+1}\big\}\end{aligned}$$

that (A.2) holds. Therefore, our consistency result holds for the gamma frailty model, even though the gamma distribution does not satisfy

Condition 5. We further note that the foregoing arguments allow $\nu = 0$, that is, zero variance of frailty.

A.2 Proof of Theorem 2

Let $\mathcal{P}_n$ be the empirical measure determined by n iid observations, and let $\mathcal{P}$ be its expectation. Also let $\mathcal{G}_n$ be the empirical process given by $\sqrt{n}(\mathcal{P}_n - \mathcal{P})$. In addition, let $l(\Lambda, \boldsymbol{\theta})$ be the log-likelihood function from a single subject and define its derivative with respect to Λ as

$$l_\Lambda(\Lambda, \boldsymbol{\theta})[\Delta\Lambda] = \lim_{\epsilon \to 0} \frac{l(\Lambda + \epsilon\Delta\Lambda, \boldsymbol{\theta}) - l(\Lambda, \boldsymbol{\theta})}{\epsilon}.$$

Also define

$$l_{\Lambda\Lambda}(\Lambda, \boldsymbol{\theta})[\Delta_1\Lambda, \Delta_2\Lambda] = \lim_{\epsilon \to 0} \frac{l_\Lambda(\Lambda + \epsilon\Delta_2\Lambda, \boldsymbol{\theta})[\Delta_1\Lambda] - l_\Lambda(\Lambda, \boldsymbol{\theta})[\Delta_1\Lambda]}{\epsilon}.$$

Similarly, let $l_{\boldsymbol{\theta}}(\Lambda, \boldsymbol{\theta})$ be the score vector for $\boldsymbol{\theta}$, and let $l_{\boldsymbol{\theta}\boldsymbol{\theta}}(\Lambda, \boldsymbol{\theta})$ be the Hessian matrix of $l(\Lambda, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. The proof is based on the expansion of the score functions for both Λ and $\boldsymbol{\theta}$ and relies on theorem 3.3.1 of van der Vaart and Wellner (1996).

A key step is to verify the invertibility of the information operator for $(\Lambda_0, \boldsymbol{\theta}_0)$. The information operator, denoted by $-\dot{\mathcal{W}}$, is the sum of an invertible linear operator and a compact operator. It suffices to show that if the information matrix along the submodel $(\Lambda_0 + \delta \int q\, d\Lambda_0, \boldsymbol{\theta}_0 + \delta\mathbf{v})$ is zero (i.e., the score function along this submodel is zero), then $\mathbf{v} = \mathbf{0}$ and $q = 0$. This is verified in Section A.4.

In light of theorem 3.3.1 of van der Vaart and Wellner (1996), we conclude that in the metric space $l^\infty(\mathcal{Q} \times \mathcal{O})$, $\sqrt{n}(\widehat{\Lambda}_n - \Lambda_0, \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges weakly to a mean-0 Gaussian process. In addition,

$$\sqrt{n}\dot{\mathcal{W}}\begin{pmatrix} \widehat{\Lambda}_n - \Lambda_0 \\ \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \end{pmatrix}\left[\begin{pmatrix} q \\ \mathbf{v} \end{pmatrix}\right] = \mathcal{G}_n\left\{l_\Lambda\left[\int q\, d\Lambda_0\right] + l_{\boldsymbol{\theta}}^T\mathbf{v}\right\} + o_p(1).$$

We can choose a finite number of q's such that $\dot{\mathcal{W}}(q, \mathbf{v}) = \mathbf{v}$. Thus $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T\mathbf{v} = \mathcal{G}_n\{l_\Lambda[\int q\, d\Lambda_0] + l_{\boldsymbol{\theta}}^T\mathbf{v}\} + o_p(1)$. We conclude that $\widehat{\boldsymbol{\theta}}_n$ is an asymptotically linear estimator for $\boldsymbol{\theta}_0$ and that its influence function is on the space spanned by the score functions. Thus $\widehat{\boldsymbol{\theta}}_n$ is an asymptotically efficient estimator.

Because the density for the random-effects model takes a generic form, the foregoing proof applies to the gamma-frailty model as well. This is also true of the proof for Theorem 3.

A.3 Proof of Theorem 3

The results in Section A.2 imply that

$$-\mathcal{P}\begin{pmatrix} l_{\Lambda\Lambda}(\Lambda_0, \boldsymbol{\theta}_0) & l_{\Lambda\boldsymbol{\theta}}(\Lambda_0, \boldsymbol{\theta}_0) \\ l_{\boldsymbol{\theta}\Lambda}(\Lambda_0, \boldsymbol{\theta}_0) & l_{\boldsymbol{\theta}\boldsymbol{\theta}}(\Lambda_0, \boldsymbol{\theta}_0) \end{pmatrix} \times \left[\begin{pmatrix} \sqrt{n}(\widehat{\Lambda}_n - \Lambda_0) \\ \sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \end{pmatrix}, \begin{pmatrix} \int_0^t q\, d\Lambda_0 \\ \mathbf{v} \end{pmatrix}\right] = \mathcal{G}_n\begin{pmatrix} l_\Lambda(\Lambda_0, \boldsymbol{\theta}_0)[\int_0^t q\, d\Lambda_0] \\ l_{\boldsymbol{\theta}}^T(\Lambda_0, \boldsymbol{\theta}_0)\mathbf{v} \end{pmatrix} + o_p(1)$$

uniformly for q with bounded variation and $\mathbf{v}$ with bounded norm. We define a function $\widetilde{\Lambda}(t)$ as a step function with jumps at the observed event times T_{ij}, with the jump size at T_{ij} equal to $\Lambda_0(T_{ij}) - \max_{T_{kl} < T_{ij}} \Lambda_0(T_{kl})$. Clearly, $\widetilde{\Lambda}(T_{ij}) = \Lambda_0(T_{ij})$. For any bounded vector $\{p_{ij}, i = 1, \ldots, n, j = 1, \ldots, n_i\}$ and bounded vector $\mathbf{v} \in \mathcal{R}^d$, we define a step function $p(t)$ such that it jumps only at T_{ij} and $p(T_{ij}) = p_{ij}$. Let $\boldsymbol{\Delta}$ denote the vector consisting of $p_{ij}\widehat{\Lambda}_n\{T_{ij}\}$. By the definition of $\mathbf{I}_n$, we obtain

$$(\boldsymbol{\Delta}^T, \mathbf{v}^T)\mathbf{I}_n\begin{pmatrix} \boldsymbol{\Delta} \\ \mathbf{v} \end{pmatrix} = -\mathcal{P}_n\begin{pmatrix} l_{\Lambda\Lambda}(\widehat{\Lambda}_n, \widehat{\boldsymbol{\theta}}_n) & l_{\Lambda\boldsymbol{\theta}}(\widehat{\Lambda}_n, \widehat{\boldsymbol{\theta}}_n) \\ l_{\boldsymbol{\theta}\Lambda}(\widehat{\Lambda}_n, \widehat{\boldsymbol{\theta}}_n) & l_{\boldsymbol{\theta}\boldsymbol{\theta}}(\widehat{\Lambda}_n, \widehat{\boldsymbol{\theta}}_n) \end{pmatrix} \times \left[\begin{pmatrix} \int_0^t p\, d\widehat{\Lambda}_n \\ \mathbf{v} \end{pmatrix}, \begin{pmatrix} \int_0^t p\, d\widehat{\Lambda}_n \\ \mathbf{v} \end{pmatrix}\right].$$

The right side of the foregoing equation converges to the information operator. Thus $\mathbf{I}_n$ is positive definite for large n.

In contrast,

$$-\sqrt{n}\begin{pmatrix} \widehat{\Lambda}_n\{T_{ij}\} - \widetilde{\Lambda}\{T_{ij}\} \\ \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \end{pmatrix}\mathbf{I}_n\begin{pmatrix} \boldsymbol{\Delta} \\ \mathbf{v} \end{pmatrix} = \mathcal{G}_n\left\{l_\Lambda(\Lambda_0, \boldsymbol{\theta}_0)\left[\int_0^t p\, d\widehat{\Lambda}_n\right] + l_{\boldsymbol{\theta}}^T(\Lambda_0, \boldsymbol{\theta}_0)\mathbf{v}\right\} + o_p(1). \quad \text{(A.3)}$$

Because $\mathbf{I}_n$ is invertible, for any $\widetilde{\mathbf{v}}$ and bounded sequence $\{q_{ij}\}_{i=1,\ldots,n, j=1,\ldots,n_i}$, we can choose $\{p_{ij}\}_{i=1,\ldots,n, j=1,\ldots,n_i}$ and $\mathbf{v}$ such that $\mathbf{I}_n\begin{pmatrix} \boldsymbol{\Delta} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \mathbf{q} \\ \widetilde{\mathbf{v}} \end{pmatrix}$, where $\mathbf{q}$ is the vector consisting of q_{ij}. With such choices, (A.3) implies that $\sum_{i=1}^n \sum_{j=1}^{n_i} \sqrt{n}(\widehat{\Lambda}_n\{T_{ij}\} - \widetilde{\Lambda}\{T_{ij}\})q_{ij} + \sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T\widetilde{\mathbf{v}}$ converges to a normal distribution with covariance

$$\begin{aligned} &\mathcal{P}\left[\left\{l_\Lambda(\Lambda_0, \boldsymbol{\theta}_0)\left[\int_0^t p\, d\widehat{\Lambda}_n\right] + l_{\boldsymbol{\theta}}^T(\Lambda_0, \boldsymbol{\theta}_0)\mathbf{v}\right\}\right. \\ &\qquad \left.\times \left\{l_\Lambda(\Lambda_0, \boldsymbol{\theta}_0)\left[\int_0^t p\, d\widehat{\Lambda}_n\right] + l_{\boldsymbol{\theta}}^T(\Lambda_0, \boldsymbol{\theta}_0)\mathbf{v}\right\}^T\right] \\ &= -\mathcal{P}\begin{pmatrix} l_{\Lambda\Lambda}(\Lambda_0, \boldsymbol{\theta}_0) & l_{\Lambda\boldsymbol{\theta}}(\Lambda_0, \boldsymbol{\theta}_0) \\ l_{\boldsymbol{\theta}\Lambda}(\Lambda_0, \boldsymbol{\theta}_0) & l_{\boldsymbol{\theta}\boldsymbol{\theta}}(\Lambda_0, \boldsymbol{\theta}_0) \end{pmatrix} \\ &\qquad \times \left[\begin{pmatrix} \int_0^t p\, d\widehat{\Lambda}_n \\ \mathbf{v} \end{pmatrix}, \begin{pmatrix} \int_0^t p\, d\widehat{\Lambda}_n \\ \mathbf{v} \end{pmatrix}\right]. \end{aligned}$$

The right-side of this equation is the limit of $(\boldsymbol{\Delta}^T, \mathbf{v}^T)\mathbf{I}_n\begin{pmatrix} \boldsymbol{\Delta} \\ \mathbf{v} \end{pmatrix}$. Thus for any vector $\mathbf{v}$ and any bounded function $q(t)$ such that $q(T_{ij}) = q_{ij}$, the asymptotic variance of $\sqrt{n}\int_0^\tau q(t)\, d(\widehat{\Lambda}_n - \Lambda_0) + \sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T\widetilde{\mathbf{v}}$ can be consistently estimated by $(\mathbf{q}^T, \widetilde{\mathbf{v}}^T)\mathbf{I}_n^{-1}\begin{pmatrix} \mathbf{q} \\ \widetilde{\mathbf{v}} \end{pmatrix}$.

A.4 Two Identifiability Results

Proposition A.1. The equation

$$\begin{aligned} &\int_{\mathbf{b}} \prod_{t \le \tau} \left\{Y(t)\lambda(t)e^{\boldsymbol{\beta}^T\mathbf{X}(t) + \mathbf{b}^T\mathbf{Z}(t)}G'(Q(t, \mathbf{O}, \mathbf{b}; \Lambda, \boldsymbol{\beta}))\right\}^{\Delta N(t)} \\ &\qquad \times \exp\left\{-G(Q(\tau, \mathbf{O}, \mathbf{b}; \Lambda, \boldsymbol{\beta}))\right\}\psi(\mathbf{b}; \boldsymbol{\gamma})\, d\mathbf{b} \\ &= \int_{\mathbf{b}} \prod_{t \le \tau} \left\{Y(t)\lambda_0(t)e^{\boldsymbol{\beta}_0{}^T\mathbf{X}(t) + \mathbf{b}^T\mathbf{Z}(t)}G'(Q(t, \mathbf{O}, \mathbf{b}; \Lambda_0, \boldsymbol{\beta}_0))\right\}^{\Delta N(t)} \\ &\qquad \times \exp\left\{-G(Q(\tau, \mathbf{O}, \mathbf{b}; \Lambda_0, \boldsymbol{\beta}_0))\right\}\psi(\mathbf{b}; \boldsymbol{\gamma}_0)\, d\mathbf{b} \end{aligned} \quad \text{(A.4)}$$

implies that $\Lambda = \Lambda_0$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Proof. We consider only the case where $Y(\tau) = 1$. For any different $t_1, \ldots, t_m, t_1', \ldots, t_k' \in [0, \tau]$, we consider the event that the recurrences occur at these times. We integrate $t_1, \ldots, t_m$ from 0 to $t_1, \ldots, t_m$ while integrating $t_1', \ldots, t_k'$ from 0 to τ in (A.4). In the equalities thus obtained, we let t_i have multiplicity k_i and further multiply by $\prod_{i=1}^m (\mathrm{i}s_i)^{k_i!}/k_i!$. After summing over $k_i = 0, 1, 2, \ldots$, we notice that the joint distribution of $\{G(\int_0^{t_i} e^{\boldsymbol{\beta}^T\mathbf{X}(s) + \mathbf{b}_1^T\mathbf{Z}(s)}\, d\Lambda)\}_{i=1}^m$ is the same as that of $\{G(\int_0^{t_i} e^{\boldsymbol{\beta}_0^T\mathbf{X}(s) + \mathbf{b}_2^T\mathbf{Z}(s)}\, d\Lambda_0)\}_{i=1}^m$, where $\mathbf{b}_1 \sim \psi(\mathbf{b}; \boldsymbol{\gamma})$ and $\mathbf{b}_2 \sim \psi(\mathbf{b}; \boldsymbol{\gamma}_0)$. It follows that $\{\log\lambda(t_i) + \boldsymbol{\beta}^T\mathbf{X}(t_i) + \mathbf{b}_1^T\mathbf{Z}(t_i)\}_{i=1}^m$ and $\{\log\lambda_0(t_i) + \boldsymbol{\beta}_0^T\mathbf{X}(t_i) + \mathbf{b}_2^T\mathbf{Z}(t_i)\}_{i=1}^m$ have the same joint distribution. Because $\int_{\mathbf{b}} \mathbf{b}\psi(\mathbf{b}; \boldsymbol{\gamma})\, d\mathbf{b} = \int_{\mathbf{b}} \mathbf{b}\psi(\mathbf{b}; \boldsymbol{\gamma}_0)\, d\mathbf{b} = 0$, we have $\log\lambda(t) + \boldsymbol{\beta}^T\mathbf{X}(t) = \log\lambda_0(t) + \boldsymbol{\beta}_0^T\mathbf{X}(t)$; thus $\lambda = \lambda_0$ and $\beta = \beta_0$ from Condition 2. Then $\mathbf{b}_1^T\mathbf{Z}(t)$ has the same distribution as $\mathbf{b}_2^T\mathbf{Z}(t)$. From Condition 2, $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$.

Proposition A.2. Suppose that $l_\Lambda(\Lambda_0, \boldsymbol{\theta}_0)[q] + l_{\boldsymbol{\theta}_0}^T\mathbf{v} = 0$ almost surely. Then $\mathbf{v} = \mathbf{0}$ and $q = 0$.

Proof. In the score function $l_\Lambda(\Lambda_0, \boldsymbol{\theta}_0)[q] + l_{\boldsymbol{\theta}_0}^T\mathbf{v}$, we consider the case where the counting process has jumps at time $t_1, \ldots, t_m$ and

$C \geq \tau$. We integrate t_i from 0 to t_i for $1 \leq i \leq l$ and from 0 to τ for $(l+1) \leq i \leq m$ and multiply by $1/(m-l)!$. After summing over all nonnegative integers for $m - l = 0, 1, 2, \ldots$, we have

$$\int_{\mathbf{b}} \left\{ \sum_{i=1}^{l} F_2(\mathbf{b}, t_i) + \frac{\psi'(\mathbf{b}; \boldsymbol{\gamma}_0)^T \mathbf{v}_{\boldsymbol{\gamma}}}{\psi(\mathbf{b}; \boldsymbol{\gamma}_0)} \right\} \times \prod_{i=1}^{l} \{H_2(\mathbf{b}, t_i)\} \psi(\mathbf{b}; \boldsymbol{\gamma}_0)\, d\mathbf{b} = 0. \quad \text{(A.5)}$$

Here and in the sequel,

$$F_1(\mathbf{b}, t_i) = (q(t_i) + \mathbf{X}(t_i)^T \mathbf{v}_{\boldsymbol{\beta}}) + \int_0^{t_i} (q(s) + \mathbf{X}(s)^T \mathbf{v}_{\boldsymbol{\beta}}) e^{\boldsymbol{\beta}_0^T \mathbf{X}(s) + \mathbf{b}^T \mathbf{Z}(s)}\, d\Lambda_0 \times \frac{G''(\int_0^{t_i} e^{\boldsymbol{\beta}_0^T \mathbf{X}(s) + \mathbf{b}^T \mathbf{Z}(s)}\, d\Lambda_0)}{G'(\int_0^{t_i} e^{\boldsymbol{\beta}_0^T \mathbf{X}(s) + \mathbf{b}^T \mathbf{Z}(s)}\, d\Lambda_0)},$$

$$F_2(\mathbf{b}, t_i) = \int_0^{t_i} (q(s) + \mathbf{X}(s)^T \mathbf{v}_{\boldsymbol{\beta}}) e^{\boldsymbol{\beta}_0^T \mathbf{X}(s) + \mathbf{b}^T \mathbf{Z}(s)}\, d\Lambda_0 \times G'\left(\int_0^{t_i} e^{\boldsymbol{\beta}_0^T \mathbf{X}(s) + \mathbf{b}^T \mathbf{Z}(s)}\, d\Lambda_0 \right),$$

$$H_1(\mathbf{b}, t_i) = \lambda(t_i) e^{\boldsymbol{\beta}_0^T \mathbf{X}(t_i) + \mathbf{b}^T \mathbf{Z}(t_i)} G'\left(\int_0^{t_i} e^{\boldsymbol{\beta}_0^T \mathbf{X}(s) + \mathbf{b}^T \mathbf{Z}(s)}\, d\Lambda_0 \right),$$

and $H_2(\mathbf{b}, t_i) = G(\int_0^{t_i} e^{\boldsymbol{\beta}_0^T \mathbf{X}(s) + \mathbf{b}^T \mathbf{Z}(s)}\, d\Lambda_0)$.

By applying the arguments used in the proof of Proposition A.1 to (A.5), we obtain

$$\int_{\mathbf{b}} \left\{ \sum_{j=1}^{l} i s_j F_2(\mathbf{b}, t_j) H_2(\mathbf{b}, t_j) + \frac{\psi'(\mathbf{b}; \boldsymbol{\gamma}_0)^T \mathbf{v}_{\boldsymbol{\gamma}}}{\psi(\mathbf{b}; \boldsymbol{\gamma}_0)} \right\} \times \exp\left\{ \sum_{j=1}^{l} i s_j H_2(\mathbf{b}, t_j) \right\} \psi(\mathbf{b}; \boldsymbol{\gamma}_0)\, d\mathbf{b} = 0$$

for any $s_1, \ldots, s_l$. We make the variable transformation $\{y_1, \ldots, y_l\} = \{H_2(\mathbf{b}, t_1), \ldots, H_2(\mathbf{b}, t_l)\}$ and take the Fourier transformation. Thus,

$$-\sum_{j=1}^{l} \frac{\partial}{\partial b_j} F_2(\mathbf{b}, t_j) H_2(\mathbf{b}, t_j) \psi(\mathbf{b}; \boldsymbol{\gamma}_0) + \psi'(\mathbf{b}; \boldsymbol{\gamma}_0)^T \mathbf{v}_{\boldsymbol{\gamma}} = 0$$

almost everywhere. Let $t_1, \ldots, t_l$ go to 0. We conclude that $\psi'(\mathbf{b}; \boldsymbol{\gamma}_0)^T \mathbf{v}_{\boldsymbol{\gamma}} = 0$ almost everywhere. Thus Condition 2(d) implies that $\mathbf{v}_{\boldsymbol{\gamma}} = \mathbf{0}$. By the fact that $\mathbf{v}_{\boldsymbol{\gamma}} = \mathbf{0}$, (A.5) with $l = 1$ is a homogeneous equation for $(\mathbf{X}(t)^T \mathbf{v}_{\boldsymbol{\beta}} + q(t))$. It is easy to see that the equation has only a trivial solution; thus $\mathbf{X}(t)^T \mathbf{v}_{\boldsymbol{\beta}} + q(t) = 0$. It follows from Condition 2(a) that $\mathbf{v}_{\boldsymbol{\beta}} = \mathbf{0}$ and $q = 0$.

A.5 Proof of Theorem 4

If we can verify the two propositions in Section A.4, then the proof of Theorem 4 is the same as the proofs of Theorems 1–3. To verify the first identifiability condition as stated in Proposition A.1, we implement the same technique in its proof and obtain that the processes

$$\left\{ G_\eta\left(\int_0^t e^{\boldsymbol{\beta}^T \mathbf{X}(s) + \mathbf{b}_1^T \mathbf{Z}(s)}\, d\Lambda \right) : t \in [0, \tau) \right\} \quad \text{and}$$

$$\left\{ G_{\eta_0}\left(\int_0^t e^{\boldsymbol{\beta}_0^T \mathbf{X}(s) + \mathbf{b}_2^T \mathbf{Z}(s)}\, d\Lambda_0 \right) : t \in [0, \tau) \right\}$$

have the same distribution, where $\mathbf{b}_1 \sim \psi(\mathbf{b}; \boldsymbol{\gamma})$ and $\mathbf{b}_2 \sim \psi(\mathbf{b}; \boldsymbol{\gamma}_0)$. Thus their derivatives at $t = 0$ have the same distribution. Because $G'_\eta(0) = 1$, we see that $\log \lambda(0) + \boldsymbol{\beta}^T \mathbf{X}(0) + \mathbf{b}_1^T \mathbf{Z}(0)$ and $\log \lambda_0(0) + \boldsymbol{\beta}_0^T \mathbf{X}(0) + \mathbf{b}_2^T \mathbf{Z}(0)$ have the same distribution. Taking the expectation and using Condition 2′(b), we have $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\lambda(0) = \lambda_0(0)$. Thus $\mathbf{b}_1^T \mathbf{Z}(0)$ and $\mathbf{b}_2^T \mathbf{Z}(0)$ have the same distribution. From Condition 2′(c), $\psi(\mathbf{b}; \boldsymbol{\gamma}) = \psi(\mathbf{b}; \boldsymbol{\gamma}_0)$, so $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$. Thus $G_\eta(\int_0^t e^{\boldsymbol{\beta}_0^T \mathbf{X}(s) + \mathbf{b}^T \mathbf{Z}(s)}\, d\Lambda) = G_{\eta_0}(\int_0^t e^{\boldsymbol{\beta}_0^T \mathbf{X}(s) + \mathbf{b}^T \mathbf{Z}(s)}\, d\Lambda_0)$ almost surely for $\mathbf{b} \sim \psi(\mathbf{b}; \boldsymbol{\gamma}_0)$. We differentiate with respect to t and obtain

$$\log G'_\eta\left(\int_0^t e^{\boldsymbol{\beta}_0^T \mathbf{X}(s) + \mathbf{b}^T \mathbf{Z}(s)}\, d\Lambda \right) + \log \lambda(t) = \log G'_{\eta_0}\left(\int_0^t e^{\boldsymbol{\beta}_0^T \mathbf{X}(s) + \mathbf{b}^T \mathbf{Z}(s)}\, d\Lambda_0 \right) + \log \lambda_0(t).$$

Further differentiation at $t = 0$ yields

$$\{G''_\eta(0) - G''_{\eta_0}(0)\} e^{\boldsymbol{\beta}_0^T \mathbf{X}(0) + \mathbf{b}^T \mathbf{Z}(0)} = \delta,$$

where δ is the derivative of $-\log \lambda(t)/\lambda_0(t)$ at $t = 0$. If $G''_\eta(0) - G''_{\eta_0}(0) \neq 0$, then $\boldsymbol{\beta}_0^T \mathbf{X}(0) + \mathbf{b}^T \mathbf{Z}(0)$ is a constant, and so is its expectation with respect to $\mathbf{b}$. Thus $\boldsymbol{\beta}_0^T \mathbf{X}(0)$ is a constant, implying that $\mathbf{b}^T \mathbf{Z}(0) = 0$. This result contradicts Condition 2′(d). Thus $G''_\eta(0) = G''_{\eta_0}(0)$. Because $G''_\eta(0)$ is strictly monotone in η due to Condition 6′, we have $\eta = \eta_0$. As a result, $\Lambda = \Lambda_0$.

To verify the second identifiability condition as stated in Proposition 2, we note that the arguments for proving $\mathbf{v}_{\boldsymbol{\gamma}} = \mathbf{0}$ are valid if we redefine $F_1(\mathbf{b}, t_i)$ as the original expression plus

$$u \dot{G}_{\eta_0}\left(\int_0^{t_i} e^{\boldsymbol{\beta}_0^T \mathbf{X}(s) + \mathbf{b}^T \mathbf{Z}(s)}\, d\Lambda_0 \right)$$

and $F_2(\mathbf{b}, t_i)$ as the original expression minus

$$u \dot{G}_{\eta_0}\left(\int_0^{\tau} e^{\boldsymbol{\beta}_0^T \mathbf{X}(s) + \mathbf{b}^T \mathbf{Z}(s)}\, d\Lambda_0 \right),$$

where u is the direction of the score for η_0. Thus, (A.5), together with the fact that $\mathbf{v}_{\boldsymbol{\gamma}} = \mathbf{0}$, implies that $F_1(\mathbf{b}; t) = 0$ almost surely. Let $t = 0$. By Condition 2′(b), $\mathbf{v}_{\boldsymbol{\beta}} = \mathbf{0}$. Then

$$0 = q(t) + \int_0^t q(s) e^{\boldsymbol{\beta}_0^T \mathbf{X}(s) + \mathbf{b}^T \mathbf{Z}(s)}\, d\Lambda_0 \times \frac{G''_{\eta_0}(Q(t, \mathbf{O}, \mathbf{b}; \Lambda_0, \boldsymbol{\beta}_0))}{G'_{\eta_0}(Q(t, \mathbf{O}, \mathbf{b}; \Lambda_0, \boldsymbol{\beta}_0))} + u \dot{G}_{\eta_0}(Q(t, \mathbf{O}, \mathbf{b}; \Lambda_0, \boldsymbol{\beta}_0)).$$

We multiply both sides by $H_1(\mathbf{b}; t)$ and integrate from 0 to t to obtain

$$q(t) = -\left\{ \frac{u \dot{G}'_{\eta_0}(Q(t, \mathbf{O}, \mathbf{b}; \Lambda_0, \boldsymbol{\beta}_0))}{G'_{\eta_0}(Q(t, \mathbf{O}, \mathbf{b}; \Lambda_0, \boldsymbol{\beta}_0))} - \frac{u \dot{G}_{\eta_0}(Q(t, \mathbf{O}, \mathbf{b}; \Lambda_0, \boldsymbol{\beta}_0)) G''_{\eta_0}(Q(t, \mathbf{O}, \mathbf{b}; \Lambda_0, \boldsymbol{\beta}_0))}{G'_{\eta_0}(Q(t, \mathbf{O}, \mathbf{b}; \Lambda_0, \boldsymbol{\beta}_0))} \right\}.$$

We further differentiate with respect to t and let $t = 0$, obtaining

$$q'(0) = -u \lambda_0(0) \dot{G}''_{\eta_0}(0) e^{\boldsymbol{\beta}_0^T \mathbf{X}(0) + \mathbf{b}^T \mathbf{Z}(0)}.$$

The arguments used in verifying the first identifiability condition then yield that $u = 0$, which implies that $q(t) = 0$.

APPENDIX B: NUMERICAL ALGORITHM

Here we provide an EM algorithm for calculating the NPMLEs and their variances. In the EM algorithm, the random effects $\mathbf{b}_i$, $i = 1, \ldots, n$, are treated as missing data. Let $\widehat{E}[\cdot]$ denote the conditional

expectation given the observable data and the current parameter estimates. In the E-step, we calculate the expectation $\widehat{E}[H(\mathbf{b}_i)]$ for some function $H(\mathbf{b}_i)$ as

$$\begin{aligned}\widehat{E}[H(\mathbf{b}_i)] = \Bigg[\int_{\mathbf{b}_i} H(\mathbf{b}_i) \prod_{t\le\tau}\Bigg\{ & e^{\mathbf{b}_i^T \mathbf{Z}_i(t)} \\ & \times G'\Bigg(\int_0^t I(C_i \ge s) e^{\boldsymbol{\beta}^T \mathbf{X}_i(s) + \mathbf{b}_i^T \mathbf{Z}_i(s)}\, d\Lambda\Bigg)\Bigg\}^{\Delta N_i(t)} \\ & \times \exp\Bigg\{-G\Bigg(\int_0^\tau I(C_i \ge t) e^{\boldsymbol{\beta}^T \mathbf{X}_i(t) + \mathbf{b}_i^T \mathbf{Z}_i(t)}\, d\Lambda\Bigg)\Bigg\} \\ & \times \psi(\mathbf{b}_i; \boldsymbol{\gamma})\, d\mathbf{b}_i\Bigg] \\ & \times \Bigg[\int_{\mathbf{b}_i} \prod_{t\le\tau}\Bigg\{ e^{\mathbf{b}_i^T \mathbf{Z}_i(t)} \\ & \times G'\Bigg(\int_0^t I(C_i \ge s) e^{\boldsymbol{\beta}^T \mathbf{X}_i(s) + \mathbf{b}_i^T \mathbf{Z}_i(s)}\, d\Lambda\Bigg)\Bigg\}^{\Delta N_i(t)} \\ & \times \exp\Bigg\{-G\Bigg(\int_0^\tau I(C_i \ge t) e^{\boldsymbol{\beta}^T \mathbf{X}_i(t) + \mathbf{b}_i^T \mathbf{Z}_i(t)}\, d\Lambda\Bigg)\Bigg\} \\ & \times \psi(\mathbf{b}_i; \boldsymbol{\gamma})\, d\mathbf{b}_i\Bigg]^{-1}. \end{aligned} \tag{B.1}$$

The integrations in this formula are evaluated by numerical approximations (e.g., Evans and Swartz 2000, chap. 4). In the M-step, we maximize the objective function

$$\begin{aligned} M(\Lambda, \boldsymbol{\theta}) = \sum_{i=1}^n \Bigg(& \int_0^\tau [\log \Lambda\{t\} + \boldsymbol{\beta}^T \mathbf{X}_i(t)]\, dN_i(t) \\ & + \int_0^\tau \widehat{E}\Bigg[\mathbf{b}_i^T \mathbf{Z}_i(t) \\ & + \log G'\Bigg(\int_0^t I(C_i \ge s) e^{\boldsymbol{\beta}^T \mathbf{X}_i(s) + \mathbf{b}_i^T \mathbf{Z}_i(s)}\, d\Lambda\Bigg)\Bigg] dN_i(t) \\ & - \widehat{E}\Bigg[G\Bigg(\int_0^\tau I(C_i \ge t) e^{\boldsymbol{\beta}^T \mathbf{X}_i(t) + \mathbf{b}_i^T \mathbf{Z}_i(t)}\, d\Lambda\Bigg)\Bigg] \\ & + \widehat{E}[\log \psi(\mathbf{b}_i; \boldsymbol{\gamma})]\Bigg). \end{aligned}$$

Clearly, $\boldsymbol{\gamma}$ can be updated by maximizing $\sum_{i=1}^n \widehat{E}[\log \psi(\mathbf{b}_i; \boldsymbol{\gamma})]$. It remains to update the values for Λ and $\boldsymbol{\beta}$ in the M-step.

Define $F(t) = \Lambda(t)/\Lambda(\tau)$. If we expand $\mathbf{X}_i(t)$ to $[1, \mathbf{X}_i(t)]$, still denoted by $\mathbf{X}_i(t)$, and expand $\boldsymbol{\beta}$ to $[\log \Lambda(\tau), \boldsymbol{\beta}]$, still denoted by $\boldsymbol{\beta}$, then the objective function in the M-step is equivalent to

$$\begin{aligned} \widetilde{M}(F, \boldsymbol{\beta}) = \sum_{i=1}^n \Bigg(& \int_0^\tau [\log F\{t\} + \boldsymbol{\beta}^T \mathbf{X}_i(t)]\, dN_i(t) \\ & - \widehat{E}\Bigg[G\Bigg(\int_0^\tau I(C_i \ge t) e^{\boldsymbol{\beta}^T \mathbf{X}_i(t) + \mathbf{b}_i^T \mathbf{Z}_i(t)}\, dF\Bigg)\Bigg] \\ & + \int_0^\tau \widehat{E}\Bigg[\mathbf{b}_i^T \mathbf{Z}_i(t) + \log G'\Bigg(\int_0^t I(C_i \ge s) \\ & \times e^{\boldsymbol{\beta}^T \mathbf{X}_i(s) + \mathbf{b}_i^T \mathbf{Z}_i(s)}\, dF\Bigg)\Bigg] dN_i(t)\Bigg), \end{aligned}$$

with constraint $\sum_{i=1}^n \int_0^\tau F\{t\}\, dN_i(t) = 1$. The constraint means that the sum of the jump sizes of the normalized Λ is 1. We order the observed event times as $\omega_1 < \omega_2 < \cdots < \omega_m$. In addition, we let $F\{\omega_j\} = \widetilde{f}_j$ and $\widetilde{F}_j = F(\omega_j)$. Then the objective function becomes

$$\begin{aligned} \widetilde{M}(F, \boldsymbol{\beta}) = & \sum_{j=1}^m \log \widetilde{f}_j + \sum_{i=1}^n \Bigg(\sum_{k=1}^{n_i - 1} \boldsymbol{\beta}^T \mathbf{X}_i(T_{ik}) \\ & - \widehat{E}\Bigg[G\Bigg(\sum_{\omega_j \le C_i} e^{\boldsymbol{\beta}^T \mathbf{X}_i(\omega_j) + \mathbf{b}_i^T \mathbf{Z}_i(\omega_j)} \widetilde{f}_j\Bigg)\Bigg] \\ & + \sum_{k=1}^{n_i - 1} \widehat{E}\Bigg[\mathbf{b}_i^T \mathbf{Z}_i(T_{ik}) \\ & + \log G'\Bigg(\sum_{\omega_j \le T_{ik}} e^{\boldsymbol{\beta}^T \mathbf{X}_i(\omega_j) + \mathbf{b}_i^T \mathbf{Z}_i(\omega_j)} \widetilde{f}_j\Bigg)\Bigg]\Bigg). \end{aligned}$$

By introducing the Lagrange multiplier μ, we then solve the equations

$$\begin{aligned} 0 = \sum_{i=1}^n \Bigg(& \sum_{k=1}^{n_i - 1} \mathbf{X}_i(T_{ik}) - \widehat{E}\Bigg[G'\Bigg(\sum_{\omega_j \le C_i} e^{\boldsymbol{\beta}^T \mathbf{X}_i(\omega_j) + \mathbf{b}_i^T \mathbf{Z}_i(\omega_j)} \widetilde{f}_j\Bigg) \\ & \times \Bigg\{\sum_{\omega_j \le C_i} \mathbf{X}_i(\omega_j) e^{\boldsymbol{\beta}^T \mathbf{X}_i(\omega_j) + \mathbf{b}_i^T \mathbf{Z}_i(\omega_j)} \widetilde{f}_j\Bigg\}\Bigg] \\ & + \sum_{k=1}^{n_i - 1} \widehat{E}\Bigg[\frac{G''(\sum_{\omega_j \le T_{ik}} e^{\boldsymbol{\beta}^T \mathbf{X}_i(\omega_j) + \mathbf{b}_i^T \mathbf{Z}_i(\omega_j)} \widetilde{f}_j)}{G'(\sum_{\omega_j \le T_{ik}} e^{\boldsymbol{\beta}^T \mathbf{X}_i(\omega_j) + \mathbf{b}_i^T \mathbf{Z}_i(\omega_j)} \widetilde{f}_j)} \\ & \times \Bigg\{\sum_{\omega_j \le T_{ik}} \mathbf{X}_i(\omega_j) e^{\boldsymbol{\beta}^T \mathbf{X}_i(\omega_j) + \mathbf{b}_i^T \mathbf{Z}_i(\omega_j)} \widetilde{f}_j\Bigg\}\Bigg]\Bigg) \end{aligned} \tag{B.2}$$

and

$$\begin{aligned} \mu = \frac{1}{\widetilde{f}_j} + \sum_{i=1}^n \Bigg(& -\widehat{E}\Bigg[G'\Bigg(\sum_{\omega_s \le C_i} e^{\boldsymbol{\beta}^T \mathbf{X}_i(\omega_s) + \mathbf{b}_i^T \mathbf{Z}_i(\omega_s)} \widetilde{f}_s\Bigg) \\ & \times \{I(\omega_j \le C_i) e^{\boldsymbol{\beta}^T \mathbf{X}_i(\omega_j) + \mathbf{b}_i^T \mathbf{Z}_i(\omega_j)}\}\Bigg] \\ & + \sum_{k=1}^{n_i - 1} \widehat{E}\Bigg[\frac{G''(\sum_{\omega_s \le T_{ik}} e^{\boldsymbol{\beta}^T \mathbf{X}_i(\omega_s) + \mathbf{b}_i^T \mathbf{Z}_i(\omega_s)} \widetilde{f}_s)}{G'(\sum_{\omega_s \le T_{ik}} e^{\boldsymbol{\beta}^T \mathbf{X}_i(\omega_s) + \mathbf{b}_i^T \mathbf{Z}_i(\omega_s)} \widetilde{f}_s)} \\ & \times \{I(\omega_j \le T_{ik}) e^{\boldsymbol{\beta}^T \mathbf{X}_i(\omega_j) + \mathbf{b}_i^T \mathbf{Z}_i(\omega_j)}\}\Bigg]\Bigg). \end{aligned} \tag{B.3}$$

When both $\mathbf{X}$ and $\mathbf{Z}$ are time-independent, (B.3) becomes

$$\begin{aligned} \mu = & \frac{1}{\widetilde{f}_j} - \sum_{i=1}^n \widehat{E}\big[G'\big(e^{\boldsymbol{\beta}^T \mathbf{X}_i + \mathbf{b}_i^T \mathbf{Z}_i} F(C_i)\big) e^{\boldsymbol{\beta}^T \mathbf{X}_i + \mathbf{b}_i^T \mathbf{Z}_i}\big] I(C_i \ge \omega_j) \\ & + \sum_{i=1}^n \sum_{k=1}^{n_i - 1} \widehat{E}\Bigg[\frac{G''(e^{\boldsymbol{\beta}^T \mathbf{X}_i + \mathbf{b}_i^T \mathbf{Z}_i} F(T_{ik}))}{G'(e^{\boldsymbol{\beta}^T \mathbf{X}_i + \mathbf{b}_i^T \mathbf{Z}_i} F(T_{ik}))} e^{\boldsymbol{\beta}^T \mathbf{X}_i + \mathbf{b}_i^T \mathbf{Z}_i}\Bigg] \\ & \times I(T_{ik} \ge \omega_j). \end{aligned} \tag{B.4}$$

Thus

$$\begin{aligned} \frac{1}{\widetilde{f}_j} = & \frac{1}{\widetilde{f}_{j+1}} + \sum_{i=1}^n \widehat{E}\big[G'\big(e^{\boldsymbol{\beta}^T \mathbf{X}_i + \mathbf{b}_i^T \mathbf{Z}_i} \widetilde{F}_j\big) e^{\boldsymbol{\beta}^T \mathbf{X}_i + \mathbf{b}_i^T \mathbf{Z}_i}\big] \\ & \times I(\omega_{j+1} > C_i \ge \omega_j) \\ & - \sum_{i=1}^n \sum_{k=1}^{n_i - 1} \widehat{E}\Bigg[\frac{G''(e^{\boldsymbol{\beta}^T \mathbf{X}_i + \mathbf{b}_i^T \mathbf{Z}_i} \widetilde{F}_j)}{G'(e^{\boldsymbol{\beta}^T \mathbf{X}_i + \mathbf{b}_i^T \mathbf{Z}_i} \widetilde{F}_j)} e^{\boldsymbol{\beta}^T \mathbf{X}_i + \mathbf{b}_i^T \mathbf{Z}_i}\Bigg] \\ & \times I(T_{ik} = \omega_j). \end{aligned}$$

Because $\widetilde{F}_j = 1 - (\widetilde{f}_{j+1} + \cdots + \widetilde{f}_m)$, (B.4) provides a recursive formula for calculating $\widetilde{f}_{m-1}, \ldots, \widetilde{f}_1$ from $\widetilde{f}_m$. If we denote $\alpha = \widetilde{f}_m$ and treat

$\widetilde{f}_{m-1}, \ldots, \widetilde{f}_1$ as functions of α and $\boldsymbol{\beta}$, then (B.2) and (B.3) become

$$\begin{aligned}
&(n_i - 1)\sum_{i=1}^{n} \mathbf{X}_i \\
&\quad - \sum_{i=1}^{n} \widehat{E}\big[G'\big(e^{\boldsymbol{\beta}^T\mathbf{X}_i + \mathbf{b}_i^T\mathbf{Z}_i} F(C_i)\big) e^{\boldsymbol{\beta}^T\mathbf{X}_i + \mathbf{b}_i^T\mathbf{Z}_i}\big] F(C_i)\mathbf{X}_i \\
&\quad + \sum_{i=1}^{n}\sum_{k=1}^{n_i - 1} \widehat{E}\left[\frac{G''(e^{\boldsymbol{\beta}^T\mathbf{X}_i + \mathbf{b}_i^T\mathbf{Z}_i} F(T_{ik}))}{G'(e^{\boldsymbol{\beta}^T\mathbf{X}_i + \mathbf{b}_i^T\mathbf{Z}_i} F(T_{ik}))} e^{\boldsymbol{\beta}^T\mathbf{X}_i + \mathbf{b}_i^T\mathbf{Z}_i}\right] \\
&\quad \times F(T_{ik})\mathbf{X}_i = \mathbf{0}
\end{aligned} \tag{B.5}$$

and

$$\sum_{j=1}^{m} \widetilde{f}_j = 1. \tag{B.6}$$

We can use the Newton–Raphson method to solve these equations. In the Newton–Raphson iterations, the derivatives of $\widetilde{f}_j$ with respect to α and $\boldsymbol{\beta}$ can be calculated using the recursive formula given in (B.4) with the initial values $\partial \widetilde{f}_m / \partial \alpha = 1$ and $\partial \widetilde{f}_m / \partial \boldsymbol{\beta} = \mathbf{0}$. Thus in the M-step we can reduce solving a large equation system to solving only a few equations.

Denote the ith term in the complete-data log-likelihood function as $l_i(\mathbf{b}_i; \Lambda, \boldsymbol{\theta})$. Then the observed information matrix can be evaluated by

$$\begin{aligned}
&-\sum_{i=1}^{n} \widehat{E}[\nabla^2 l_i(\mathbf{b}_i; \Lambda, \boldsymbol{\theta})] \\
&\qquad - \sum_{i=1}^{n} \big(\widehat{E}[\nabla l_i(\mathbf{b}_i; \Lambda, \boldsymbol{\theta})^{\otimes 2}] - \widehat{E}[\nabla l_i(\mathbf{b}_i; \Lambda, \boldsymbol{\theta})]^{\otimes 2}\big),
\end{aligned}$$

where $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$ and ∇ and ∇^2 denote the first and the second derivatives with respect to $\boldsymbol{\theta}$ and the jump sizes of Λ.

[Received March 2006. Revised July 2006.]

REFERENCES

Akaike, H. (1985), "Prediction and Entropy," in *A Celebration of Statistics*, eds. A. C. Atkinson and S. E. Fienberg, New York: Springer-Verlag, pp. 1–24.

Andersen, P. K., Borgan, O., Grill, R. D., and Keiding, N. (1993), *Statistical Models Basel on Counting Processes*, New York: Springer-Verlag.

Andersen, P. K., and Gill, R. D. (1982), "Cox's Regression Model for Counting Processes: A Large-Sample Study," *The Annals of Statistics*, 10, 1100–1200.

Bennett, S. (1983), "Analysis of Survival Data by the Proportional Odds Model," *Statistics in Medicine*, 2, 273–277.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: Johns Hopkins University Press.

Burdick, R. K., and Graybill, F. A. (1992), *Confidence Intervals on Variance Components*, New York: Marcel Dekker.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Evans, M., and Swartz, T. (2000), *Approximating Integrals via Monte Carlo and Deterministic Methods*, Oxford, U.K.: Oxford University Press.

Kosorok, M. R., Lee, B. L., and Fine, J. P. (2004), "Robust Inference for Univariate Proportional Hazards Frailty Regression Models," *The Annals of Statistics*, 32, 1448–1491.

Lawless, J. F., and Nadeau, C. (1995), "Some Simple Robust Methods for the Analysis of Recurrent Events," *Technometrics*, 37, 158–168.

Lawless, J. F., Nadeau, C., and Cook, R. J. (1997), "Analysis of Mean and Rate Functions for Recurrent Events," in *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, eds. D. Y. Lin and T. R. Fleming, New York: Springer-Verlag, pp. 37–49.

Lin, D. Y., Wei, L. J., Yang, I., and Ying, Z. (2000), "Semiparametric Regression for the Mean and Rate Functions of Recurrent Events," *Journal of the Royal Statistical Society*, Ser. B, 62, 711–730.

Lin, D. Y., Wei, L. J., and Ying, Z. (2001), "Semiparametric Transformation Models for Point Processes," *Journal of the American Statistical Association*, 96, 620–628.

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 44, 226–233.

Miloslavsky, M., Keles, S., and van der Laan, M. J. (2004), "Recurrent Events Analysis in the Presence of Time-Dependent Covariates and Dependent Censoring," *Journal of the Royal Statistical Society*, Ser. B, 66, 239–257.

Murphy, S. A. (1994), "Consistency in a Proportional Hazards Model Incorporating a Random Effect," *The Annals of Statistics*, 22, 712–731.

——— (1995), "Asymptotic Theory for the Frailty Model," *The Annals of Statistics*, 23, 182–198.

Murphy, S. A., and van der Vaart, A. W. (2000), "On the Profile Likelihood," *Journal of the American Statistical Association*, 95, 449–465.

Nielsen, G. G., Gill, R. D., Andersen, P. K., and Sorensen, T. I. A. (1992), "A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models," *Scandinavian Journal of Statistics*, 19, 25–43.

Oakes, D. (1992), "Frailty Models for Multiple Event Times," in *Survival Analysis: State of the Art*, eds. J. P. Klein and P. K. Goel, Amsterdam: Kluwer Academic, pp. 371–379.

Pepe, M. S., and Cai, J. (1993), "Some Graphical Displays and Marginal Regression Analyses for Recurrent Failure Times and Time-Dependent Covariates," *Journal of the American Statistical Association*, 88, 811–820.

Pettitt, A. N. (1984), "Proportional Odds Models for Survival Data and Estimates Using Ranks," *Applied Statistics*, 33, 169–175.

Prentice, R. L., Williams, B. J., and Peterson, A. V. (1981), "On the Regression Analysis of Multivariate Failure Time Data," *Biometrika*, 68, 373–379.

Schaubel, D. E., and Cai, J. W. (2004), "Regression Methods for the Gap Time Hazard Functions of Sequentially Ordered Multivariate Failure Time Data," *Biometrika*, 91, 291–303.

Therneau, T. M., and Grambsch, P. M. (2000), *Modeling Survival Data: Extending the Cox Model*, New York: Springer-Verlag.

Therneau, T. M., and Hamilton, S. A. (1997), "rhDNase as an Example of Recurrent Event Analysis," *Statistics in Medicine*, 16, 2029–2047.

van der Laan, M. J., Dudoit, S., and Keles, S. (2004), "Asymptotic Optimality of Likelihood-Based Cross-Validation," *Statistical Applications in Genetics and Molecular Biology*, 3, No. 1, article 4.

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer-Verlag.

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989), "Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions," *Journal of the American Statistical Association*, 84, 1065–1073.