Biometrika Advance Access published April 27, 2009

Biometrika (2009), *pp*. 1–15 doi: 10.1093/biomet/asp008
© 2009 Biometrika Trust
Printed in Great Britain

Gamma frailty transformation models for multivariate survival times

BY DONGLIN ZENG

Department of Biostatistics, University of North Carolina, 3105-D McGavran-Greenberg Hall, Campus Box 7420, Chapel Hill, North Carolina, 27516, U.S.A.

dzeng@bios.unc.edu

QINGXIA CHEN

Department of Biostatistics, Vanderbilt University, 1161 21st Avenue South, S-2323 Medical Center North, Nashville, Tennessee, 37232, U.S.A.

cindy.chen@vanderbilt.edu

AND JOSEPH G. IBRAHIM

Department of Biostatistics, University of North Carolina, 3109 McGavran-Greenberg Hall, Campus Box 7420, Chapel Hill, North Carolina, 27516, U.S.A.

ibrahim@bios.unc.edu

SUMMARY

We propose a class of transformation models for multivariate failure times. The class of transformation models generalize the usual gamma frailty model and yields a marginally linear transformation model for each failure time. Nonparametric maximum likelihood estimation is used for inference. The maximum likelihood estimators for the regression coefficients are shown to be consistent and asymptotically normal, and their asymptotic variances attain the semiparametric efficiency bound. Simulation studies show that the proposed estimation procedure provides asymptotically efficient estimates and yields good inferential properties for small sample sizes. The method is illustrated using data from a cardiovascular study.

Some key words: Gamma frailty model; Linear transformation model; Proportional hazards model; Semiparametric efficiency.

1. INTRODUCTION

Multivariate failure time data arise when each study subject can potentially experience several events (Kalbfleisch & Prentice, 2002, Chs. 8–10). For multivariate failure times, it is often interesting to determine risk factors that are predictive for all the failures or predictive for some failures but not others. For example, in a colon cancer study (Lin, 1994), patients with resected colon cancer could experience cancer recurrence and then die; in this study, investigators wished to access the efficacy of adjuvant therapy for both types of failures. Since multivariate failure times are from the same subject, they are potentially dependent on one another, and ignoring this may lead to biased inference. To account for the dependence of correlated failure times in a statistical model, it is natural and convenient to represent such dependence through a frailty term or a random effect (Clayton & Cuzick, 1985; Oakes, 1989, 1991; Hougaard, 2000). In particular, the proportional hazards model (Cox, 1972) with gamma frailty was used to incorporate

covariates by Nielsen et al. (1992) and Klein (1992), and for a subject with covariate X, such a model takes the form $\Lambda_k(t \mid X) = \omega \Lambda_k(t) \exp(\beta_k^{\mathrm{T}} X)$, where $\Lambda_k(t \mid X)$ is the cumulative hazard function for failure time of type k, $\Lambda_k(t)$ is an unknown baseline function and ω is the gamma frailty. The asymptotic properties of the nonparametric maximum likelihood estimates from this model can be studied using the same theory as in Murphy (1994, 1995) and Parner (1998).

The proportional hazards model assumes that the hazard ratios across covariate levels are constant over time. Such an assumption is often violated in scientific studies. Instead, other semiparametric models may provide more accurate or more concise summarization of the data. One alternative is the proportional odds model (Bennett, 1983; Murphy et al., 1997), which assumes that the hazard ratio between two sets of covariate values converges to unity as time increases rather than being constant. In fact, both the proportional hazards and proportional odds models belong to the class of linear transformation models, which relate an unknown transformation of the failure time linearly to the covariates (Kalbfleisch & Prentice, 2002, p. 241). For univariate survival data, Dabrowska & Doksum (1988), Cheng et al. (1995) and Chen et al. (2002) proposed general estimators for this class of models. More general transformation models have been recently proposed and studied by Zeng & Lin (2006).

The literature on the generalization of transformation models to multivariate failure times by incorporating random effects is very limited. Only Zeng & Lin (2007) discussed a class of such transformation models. In their paper, the cumulative hazard function for the failure time of type k is assumed to have the form $G_k\{\exp(\beta_k^{\mathrm{T}} X + b^{\mathrm{T}} Z)\Lambda_k(t)\}$, where X and Z are subject-specific covariates, G_k is the transformation used and b is the random effect. Such a class of transformation models include the proportional hazards and the proportional odds models as special cases, and incorporates the dependence across failure times through the random effect b. However, as discussed by Zeng & Lin (2007), this class does not include the usual gamma frailty model, and inference requires that the latent random effects be nonzero. Additionally, for many simple transformations, β_k does not have a clear interpretation regarding the relationship between X and the failure times.

In this paper, we propose a different general class of transformation models with gamma frailty. Our class includes the gamma frailty model as a special case and allows the random effects to be zero. The proposed models are not covered by the general class of models in Zeng & Lin (2007), and imply another set of marginal transformation models for each failure time so their regression coefficients have a direct interpretation similar to the usual linear transformation models.

2. MODELS AND INFERENCE

Let T_{ik} denote the failure time of event type k ($k = 1, \ldots, K$). We assume that given ω_i,

$$\Lambda_{ik}(t \mid \omega_i, X_i) = G_k\{\Lambda_k(t)\exp(\beta_k^{\mathrm{T}} X_i)\}\omega_i,$$

where ω_i follows the gamma distribution with mean unit and variance θ. The model is a natural generalization of the usual frailty model, which yields it as a special case, but we allow more flexible choices for the transformation model. For example, when $G_k(x) = x$, we obtain the proportional hazards model, and when $G_k(x) = \log(1 + x)$, we obtain the proportional odds model. Furthermore, under the above model, we can easily obtain the marginal cumulative hazard function for the failure time of event type k as $\log[1 + \theta G_k\{\Lambda_k(t)\exp(\beta_k^{\mathrm{T}} X_i)\}]/\theta$. Equivalently, T_{ik} given X_i satisfies another linear transformation model $\log \Lambda_k(T_{ik}) = -\beta_k^{\mathrm{T}} X_i + \epsilon_{ik}$, where ϵ_{ik} follows the distribution $\log G_k^{-1}\{(U^{-\theta} - 1)/\theta\}$, and $U \sim$ Uniform(0,1). Hence, β_k has the same interpretation as the coefficients in the usual linear transformation models (Cheng et al., 1995; Zeng & Lin, 2006). Furthermore, we allow $\theta = 0$, i.e. no frailty exists among the failure times.

We further assume that the study ends at some finite τ. However, when the event times are cure-type events (Berkson & Gage, 1952), that is, the survival probability at infinity is nonzero, we also allow $\tau = \infty$. Under these assumptions, we can rewrite the above model as

$$\Lambda_{ik}(t \mid \omega_i, X_i) = G_k\{F_k(t)\exp(\alpha_k + \beta_k^{\mathrm{T}} X_i)\}\omega_i \tag{1}$$

by defining $F_k(t) = \Lambda_k(t)/\Lambda_k(\tau)$ and $\alpha_k = \log \Lambda_k(\tau)$. Clearly, $F_k(\tau) = 1$; i.e. $F_k(t)$ is a distribution function in $[0, \tau]$.

If we have a set of right-censored data, the observed data for a single cluster are $\{Y_k = \min(T_k, C_k), \Delta_k = I(T_k \leqslant C_k), X\}(k = 1, \ldots, K)$, where C_k is the censoring time for event type k. Therefore, under the assumption that the censoring time is independent of the failure time and the frailty given the covariates, the likelihood function is

$$L_n(\alpha, \beta, \theta, F) = \prod_{k=1}^{K} \left[G_k'\{F_k(Y_k)e^{\alpha_k+\beta_k^{\mathrm{T}}X}\}F_k(Y_k)e^{\alpha_k+\beta_k^{\mathrm{T}}X}\right]^{\Delta_k} \times \int \omega^{\sum_{k=1}^{K}\Delta_k} \exp\left(-\omega\left[\sum_{k=1}^{K} G_k\{F_k(Y_k)e^{\alpha_k+\beta_k^{\mathrm{T}}X}\}\right]\right) g_\theta(\omega)\, d\omega.$$

The estimation of the parameters α_k, β_k, F_k and θ is based on nonparametric maximum likelihood estimation. In this approach, we treat F_k as a discrete distribution function with positive jumps at the Y_k for which $\Delta_k = 1$. Thus, the estimates of all the parameters maximize the following loglikelihood function from n independent and identically distributed clusters:

$$l_n(\alpha, \beta, \theta, F) = \sum_{i=1}^{n}\sum_{k=1}^{K} \Delta_{ik}\left[\log G_k'\{F_k(Y_{ik})e^{\alpha_k+\beta_k^{\mathrm{T}}X_i}\} + \log F_k\{Y_{ik}\} + \alpha_k + \beta_k^{\mathrm{T}}X_i\right] + \sum_{i=1}^{n} \log \int \omega^{\sum_{k=1}^{K}\Delta_{ik}} \exp\left(-\omega\left[\sum_{k=1}^{K} G_k\{F_k(Y_{ik})e^{\alpha_k+\beta_k^{\mathrm{T}}X_i}\}\right]\right) g_\theta(\omega)\, d\omega,$$

where $l_n(\alpha, \beta, \theta, F) = \log L_n(\alpha, \beta, \theta, F)$, and $F_k\{t\}$ denotes the jump size of F_k at t. In fact, it is easy to show that the estimates for F_k must be a cumulative distribution function with positive jumps only at the Y_{ik}s for which $\Delta_{ik} = 1$. Therefore, the above maximization should be performed over the parameters α_k, β_k, θ and these positive jumps. Additionally, the summation of these positive jumps for F_k should be unity due to the fact that F_k is a distribution function in $[0, \tau]$. We denote the nonparametric maximum likelihood estimates for α_k, β_k, θ and F_k by $\hat{\alpha}_k$, $\hat{\beta}_k$, $\hat{\theta}$ and $\hat{F}_k$, respectively. Correspondingly, we let $\hat{\Lambda}_k$ denote the estimate of Λ_k.

Our subsequent theory will establish asymptotic normality of the nonparametric maximum likelihood estimators, and specifically that the asymptotic covariance for the nonparametric maximum likelihood estimators can be consistently estimated using the inverse of the observed Fisher information matrix. That is, we treat the jump sizes of $\hat{F}_k$ as usual parameters along with α_k, β_k and θ. We then calculate the observed information matrix for these parameters, denoted by $\hat{J}$. The asymptotic covariance for $\hat{F}_k$, $\hat{\alpha}_k$, $\hat{\beta}_k$ and $\hat{\theta}$ is thus estimated using the delta method. For example, to estimate the asymptotic covariance of $\sum_k(\int h_k d\hat{F}_k + t_{0k}\hat{\alpha}_k + t_{1k}^{\mathrm{T}}\hat{\beta}_k)$ for some deterministic functions h_k and constants t_{0k} and t_{1k} $(k = 1, \ldots, K)$, which can also be written as $\sum_k\{\sum_{i=1}^{n}\Delta_{ik}h_k(Y_{ik})\hat{F}_k\{Y_{ik}\} + t_{0k}\hat{\alpha}_k + t_{1k}^{\mathrm{T}}\hat{\beta}_k\}$, we can use $\tilde{h}^{\mathrm{T}}\hat{J}^{-1}\tilde{h}$, where $\tilde{h}$ is the vector consisting of all $h_k(Y_{ik})$ for which $\Delta_{ik} = 1$ and t_{0k} and t_{1k} for $k = 1, \ldots, K$.

When the true frailty variance is zero, the estimate for θ, $\hat{\theta}$, can be negative. In this case, we use $\max(0, \hat{\theta})$ as the estimate for θ. Then from the theory given later, such a modified estimate has a half-truncated normal distribution, so that inference for θ can be carried out accordingly.

3. COMPUTATIONAL ALGORITHM

We present a computationally convenient algorithm to compute the nonparametric maximum likelihood estimator, which avoids maximization over a large number of parameters. We apply the expectation-maximization, EM, algorithm by treating ω_i as missing data. In the E-step, we evaluate the conditional expectation of some function $Q(\omega_i)$ given the observed data. The conditional density of ω_i given the observed data is

$$\text{Gamma}\left(\theta^{-1}+\sum_{k=1}^{K}\Delta_k,\ \left[\theta^{-1}+\sum_{k=1}^{K}G_k\{F_k(Y_{ik})e^{\alpha_k+\beta_k^{\mathrm{T}}X_i}\}\right]^{-1}\right).$$

The conditional expectation of $Q(\omega_i)$ can be calculated analytically or by a Laplace approximation; we denote it by $\hat{E}\{Q(\omega_i)\}$. In the M-step, we need to maximize the following loglikelihood function:

$$\begin{aligned} l_n(\alpha,\beta,\theta,F) = {} & \sum_{i=1}^{n}\sum_{k=1}^{K}\Delta_{ik}\big[\log G_k'\{F_k(Y_{ik})e^{\alpha_k+\beta_k^{\mathrm{T}}X_i}\}+\log F_k(Y_{ik}) \\ & +\alpha_k+\beta_k^{\mathrm{T}}X_i+\hat{E}(\log\omega_i)\big]-\sum_{i=1}^{n}\hat{E}(\omega_i)\sum_{k=1}^{K}G_k\{F_k(Y_{ik})e^{\alpha_k+\beta_k^{\mathrm{T}}X_i}\} \\ & -n\log\theta^{1/\theta}\Gamma(1/\theta)+(1/\theta-1)\sum_{i=1}^{n}\hat{E}(\log\omega_i)-\theta^{-1}\sum_{i=1}^{n}\hat{E}(\omega_i). \end{aligned}$$

To this end, we order the Y_{ik} for which $\Delta_{ik}=1$ from smallest to largest, and denote them as $y_{1k}<\cdots<y_{n_k,k}$. The corresponding covariates are denoted by $x_{1k},\ldots,x_{n_k,k}$. Let f_{lk} be the jump size of $F_k(\cdot)$ at y_{lk} and let F_{lk} denote $F_k(y_{lk})$ $(l=1,\ldots,n_k)$. After differentiating the loglikelihood function with respect to f_{lk}, we obtain

$$\begin{aligned} \frac{1}{f_{lk}} = {} & -\sum_{Y_{ik}\geqslant y_{lk}}\Delta_{ik}\frac{G_k''\{F_k(Y_{ik})e^{\alpha_k+\beta_k^{\mathrm{T}}X_i}\}}{G_k'\{F_k(Y_{ik})e^{\alpha_k+\beta_k^{\mathrm{T}}X_i}\}}e^{\alpha_k+\beta_k^{\mathrm{T}}X_i} \\ & +\sum_{Y_{ik}\geqslant y_{lk}}\hat{E}(\omega_i)G_k'\{F_k(Y_{ik})e^{\alpha_k+\beta_k^{\mathrm{T}}X_i}\}e^{\alpha_k+\beta_k^{\mathrm{T}}X_i}+\lambda_k, \end{aligned}$$

where λ_k is the Lagrange multiplier for the constraint $\sum_{l=1}^{n_k}f_{lk}=1$. This yields

$$\begin{aligned} \frac{1}{f_{lk}} = {} & \frac{1}{f_{l+1,k}}-\sum_{y_{lk}\leqslant Y_{ik}<y_{l+1,k}}\Delta_{ik}\frac{G_k''(F_{lk}e^{\alpha_k+\beta_k^{\mathrm{T}}X_i})}{G_k'(F_{lk}e^{\alpha_k+\beta_k^{\mathrm{T}}X_i})}e^{\alpha_k+\beta_k^{\mathrm{T}}X_i} \\ & +\sum_{y_{lk}\leqslant Y_{ik}<y_{l+1,k}}\hat{E}(\omega_i)G_k'(F_{lk}e^{\alpha_k+\beta_k^{\mathrm{T}}X_i})e^{\alpha_k+\beta_k^{\mathrm{T}}X_i}. \end{aligned} \tag{2}$$

Since $F_{lk}=1-f_{l+1,k}-f_{l+2,k}-\cdots-f_{n_k,k}$, this gives us a recursive formula for calculating f_{lk} from $f_{l+1,k},\ldots,f_{n_k,k}$. Therefore, we can treat α_k, β_k, $\xi_k=f_{n_k,k}$ $(k=1,\ldots,K)$, and θ as the parameters to be updated in the M-step, since any other f_{lk} can be indirectly expressed as a function of these parameters using formula (2). Hence, the maximization in the M-step can be carried out over a small set of parameters including α_k, β_k, ξ_k $(k=1,\ldots,K)$, and θ. In practice, a one-step Newton–Raphson algorithm can be used to update these parameters. In particular, the

equations to be solved are

$$\sum_{i=1}^{n} \Delta_{ik} \left[\frac{G_k''\{F_k(Y_{ik})e^{\alpha_k+\beta_k^{\mathrm{T}}X_i}\}}{G_k'\{F_k(Y_{ik})e^{\alpha_k+\beta_k^{\mathrm{T}}X_i}\}} F_k(Y_{ik})e^{\alpha_k+\beta_k^{\mathrm{T}}X_i} + 1 \right] (1, X_i^{\mathrm{T}})^{\mathrm{T}}$$
$$- \sum_{i=1}^{n} \hat{E}(\omega_i) G_k'\big(F_k(Y_{ik})e^{\alpha_k+\beta_k^{\mathrm{T}}X_i}\big) F_k\big(Y_{ik}e^{\alpha_k+\beta_k^{\mathrm{T}}X_i}\big)(1, X_i^{\mathrm{T}})^{\mathrm{T}} = 0, \tag{3}$$

$$\sum_{l=1}^{n_k} f_{lk} = 1 \quad (k = 1, \ldots, K), \tag{4}$$

and

$$\frac{n}{\theta^2}\log\theta - \frac{n}{\theta^2} + n\frac{\Gamma'(1/\theta)}{\theta^2\Gamma(1/\theta)} - \frac{1}{\theta^2}\sum_{i=1}^{n}\hat{E}(\log\omega_i) + \frac{1}{\theta^2}\sum_{i=1}^{n}\hat{E}(\omega_i) = 0, \tag{5}$$

where f_{lk} is a function of α_k, β_k and ξ_k. We iterate between the E- and M-steps until convergence. The resulting estimates are then the nonparametric maximum likelihood estimators.

One limitation in the above algorithm is that the estimate of θ must be positive. However, when the frailty variance is zero, the maximum likelihood estimate of θ can be zero or even negative. The EM algorithm is not applicable in this case since the frailty ω_i has an improper density when $\theta < 0$. To overcome this dilemma, we add a second set of equations by fixing $\theta = 0$ and then compute the nonparametric maximum likelihood estimators for the other parameters. The same EM algorithm can be used in this case except that θ is set to be zero. We then compare the observed likelihood functions for the nonparametric maximum likelihood estimators obtained in the first set of equations with $\theta \neq 0$ and the observed likelihood function for the nonparametric maximum likelihood estimators obtained in the second set of equations with $\theta = 0$. The equations with the larger likelihood function are treated as the final estimates.

Finally, the inverse of the observed information matrix can be used to estimate the asymptotic covariance matrix of the parameter estimates. The observed information matrix can be calculated using Louis' formula (Louis, 1982) based on equations (3)–(5).

4. Asymptotic results

We establish asymptotic results for the proposed estimators. For a noncured failure time in which all subjects eventually fail, we assume τ to be finite; while for a cured failure time in which some subjects never failure, we allow τ to equal infinity. The latter generalizes the development in Zeng et al. (2006) to the multivariate case, and it does not assume that we observe events at $\tau = \infty$, i.e. all the cured subjects are right censored.

Let θ_0 and $(\alpha_{0k}, \beta_{0k}, F_{0k})$ denote the true values for the parameters θ and (α_k, β_k, F_k), respectively. We need the following assumptions.

Assumption 1. The true parameter β_{0k} belongs to a known bounded region $\mathcal{B}_k$ and $0 \leqslant \theta_0 < \theta_M$ for some constant θ_M, $f_{0k}(t) > 0$ for $t \in [0, \tau]$ and $k = 1, \ldots, K$. The vectors $(1, X)$ are linearly independent with positive probability and X is bounded with probability one.

Assumption 2. The transformation G_k is a strictly increasing function with $G_k(0) = 0$ and $G_k'(0) > 0$, and is three-times continuously differentiable in $[0, \tau]$. Moreover, $G_k''(x) \leqslant 0$, $\limsup_{x\to\infty}\{G_k'(x)/G_k'(Mx) + G_k(Mx)/G_k(x)\} < \infty$, where $M = \max_{k=1}^{K} \sup_{\beta\in\mathcal{B}_k, X} \beta_k^{\mathrm{T}}X$.

Assumption 3. The censoring times $(C_1, \ldots, C_K)$ are independent of $(T_1, \ldots, T_K)$ conditional on X. Moreover, $\inf_x P(C_1 = \tau, \ldots, C_K = \tau \mid X = x) > 0$.

The assumption about the known boundness of β_{0k} in Assumption 1 is standard. In the assumption, we allow the possibility that $\theta_0 = 0$, corresponding to no correlation among all types of events. Moreover, we do not assume a known bound for α_k and $\Lambda_k(\tau)$. Therefore, we do not impose a bound for $\hat{\alpha}_k$. This yields a very challenging statistical issue for establishing consistency of the parameter estimates. Assumption 2 implies that there exists some constant $c_0 > 0$ such that $G_k(x) \leqslant c_0 x$, $G_k'(x) \leqslant c_0 G_k'(Mx)$ and $G_k(Mx) \leqslant c_0 G_k(x)$. Many classes of transformations satisfy these properties. For example, $G_k(x) = r^{-1}\log(1 + xr)$ with $r \geqslant 0$ and $G_k(x) = \{(1 + x)^\rho - 1\}/\rho$ with $\rho \in [0, 1]$. Assumption 3 assumes that all subjects surviving at τ are right-censored.

Under these assumptions, we first establish identifiability of the model parameters.

THEOREM 1. *Under Assumptions* 1–3*, all the parameters, including* (α_k, β_k, F_k) $(k = 1, \ldots, K)$ *and* θ*, are identifiable.*

The proof of Theorem 1, which is given in the Appendix, utilizes the expression of the likelihood function but considers the cases $\theta_0 = 0$ and $\theta_0 > 0$ separately. Our next result shows that any nontrivial one-dimensional submodel possesses a nonsingular Fisher information matrix. Such a result is necessary to establish the subsequent asymptotic properties.

THEOREM 2. *Under Assumptions* 1–3*, for any one-dimensional submodel given as* $\{\alpha_{0k} + \epsilon a_k, \beta_{0k} + \epsilon b_k, dF_{0k} + \epsilon \int h_{0k} dF_{0k}, \theta_0 + \epsilon w, (k = 1, \ldots, K)\}$*, the Fisher information along this submodel is nonsingular, where* h_{0k} *is a function in* $BV[0, \tau]$*, the space of all functions with bounded total variation, satisfying* $\int_0^\tau h_k(s) dF_{0k}(s) = 0$.

The proof is given in the Appendix and is based on an explicit expression for the score function along this submodel and considers $\theta_0 = 0$ and $\theta_0 > 0$ separately. Using both Theorem 1 and Theorem 2, we are able to obtain the following consistency results.

THEOREM 3. *Under Assumptions* 1–3*, it follows that*

$$\sum_{k=1}^{K}(|\,\hat{\alpha}_k - \alpha_{0k}\,| + |\,\hat{\beta}_k - \beta_{0k}\,|) + |\,\hat{\theta} - \theta_0\,| + \sum_{k=1}^{K} \sup_{t\in[0,\tau]} |\,\hat{F}_k - F_{0k}\,| \to 0,$$

almost surely. The proof of Theorem 3, given in the Appendix, relies on first obtaining the compactness of $\hat{\alpha}_k$, or equivalently, the uniform boundedness of $\hat{\Lambda}_k$. However, in contrast to Murphy (1994) and Parner (1998) who worked on $\hat{\Lambda}_k$ directly, our proof works on $G_k(\Lambda_k)$ instead due to the nature of the transformation models. Moreover, the proof of consistency needs to consider $\theta_0 > 0$ and $\theta_0 = 0$ separately.

Our last theorem gives the asymptotic properties of the nonparametric maximum likelihood estimators.

THEOREM 4. *Under Assumptions* 1–3 *and treating* $\hat{\Lambda}_k$ *as a function in* $BV[0, \tau]$*,* $n^{1/2}(\hat{\beta}_k - \beta_{k0}, \hat{\theta} - \theta_0, \hat{\Lambda}_k - \Lambda_{0k})_{k=1,\ldots,K}$*, converges in distribution to a zero-mean Gaussian process in the product of real spaces and* $BV[0, \tau]^{\otimes K}$*. Moreover, the asymptotic covariances of* $\hat{\beta}_k$ $(k = 1, \ldots, K)$ *and* $\hat{\alpha}$ *attain their semiparametric efficiency bound.*

The proof follows from Theorem 2 in the Appendix of Zeng & Lin (2007), so we only outline it below. First, since θ_0 may be zero, we consider an extended domain $\theta_M > \theta > -\epsilon_0$ where $\epsilon_0 = \min[1/K, \{\sum_{k=1}^K G_k(e^{\alpha_{0k} + \beta_{0k}^{\mathrm{T}} X})\}^{-1}]$. Under the extended likelihood function, the true parameter

value for θ is in the interior of the domain and by consistency, when n is large enough, $\hat{\theta}$ belongs to the interior of the domain. We can easily check that condition (C1) in Zeng & Lin (2007) follows from our Assumption 1 and the above extension. Condition (C2) in Zeng & Lin (2007) is implied by our Assumption 1. Conditions (C4), (C6) and (C8) in Zeng & Lin (2007) can be verified by direct calculations. Our previous results on parameter identifiability and nonsingularity of submodel information yield the two identifiability conditions (C5) and (C7) given in Zeng & Lin (2007). Condition (C3) in Zeng & Lin (2007) is not necessary for the asymptotic distribution once consistency holds.

Remark 1. As in Theorem 2 of Zeng & Lin (2007), the asymptotic covariance matrix attains the efficiency bound and can be estimated using the observed information matrix as described in the previous section. Furthermore, since $\hat{\theta}$ can be negative, we propose to estimate θ by $\breve{\theta} = \hat{\theta} I(\hat{\theta} \geqslant 0)$, where $I(x)$ is an indicator function of condition x. Thus, from Theorem 4, we conclude that $\breve{\theta}$ possesses the same asymptotic distribution as $\hat{\theta}$ if $\theta_0 > 0$; however, if $\theta_0 = 0$, then $n^{1/2}\breve{\theta} \to \sigma Z I(Z \geqslant 0)$, in distribution, where Z denotes the standard normal distribution, and σ^2 is the asymptotic variance of $\hat{\theta}$. This result can be used to test whether $\theta = 0$.

Remark 2. For a cured failure time, when $\tau = \infty$, we can estimate the cure rate for each event type. From the assumed model, it is easy to calculate

$$\Pr(T_k \geqslant \tau \mid X_i) = \int \exp\big[-\omega G_k\big\{F_k(\tau)e^{\alpha_k+\beta_k^{\mathrm{T}}X}\big\}\big] g_\theta(\omega) d\omega = \big\{1 + \theta G_k\big(e^{\alpha_k+\beta_k^{\mathrm{T}}X}\big)\big\}^{-1/\theta}.$$

Thus, an estimator of the cure rate for event type k given X is given by $\hat{\pi}_k = \{1 + \hat{\theta} G_k(e^{\hat{\alpha}_k+\hat{\beta}_k^{\mathrm{T}}X})\}^{-1/\hat{\theta}}$. From the previous derivation, $\hat{\pi}_k$ also has an asymptotically normal distribution and its variance can be estimated using the delta method.

5. Numerical studies

We conducted simulation studies to examine the small sample performance of the proposed methodology using 1000 replications. We considered two event types, $k = 1, 2$, with the following transformations:

$$G_k(x) = \begin{cases} \{(1+x)^\rho - 1\}/\rho & (\rho \geqslant 0), \\ \log(1 + rx)/r & (r \geqslant 0). \end{cases} \tag{6}$$

In this family, $\rho = 1$ or $r = 0$ (i.e. $G(x) = x$) yields the proportional hazards model, whereas $\rho = 0$ or $r = 1$, i.e. $G(x) = \log(1 + x)$ yields the proportional odds model. Other than these two special cases, we also considered two other transformations in the simulation study: $\rho = 0{\cdot}5$, which gives $G(x) = 2\{(1 + x)^{1/2} - 1\}$; $\rho = 0$ and $r = 0{\cdot}5$, which gives $G(x) = 2\log(1 + x/2)$. The transformation model had a cumulative hazard function of the form

$$\Lambda_{ik}(t \mid \omega_i, X_{i1}, X_{i2}) = G_k\{F_k(t)\exp(\alpha_k + \beta_{k1}X_{i1} + \beta_{k2}X_{i2})\}\omega_i \quad (i = 1, \ldots, n; \quad k = 1, 2),$$

where X_{i1} was simulated from a uniform distribution on $(0, 1)$, X_{i2} was simulated from a Bernoulli distribution with success probability $p = 0{\cdot}4$, ω_i was simulated from a Gamma $(1/\theta, \theta)$ distribution, and $F_k(t) = (1 - e^{-t})/(1 - e^{-3})I(0 \leqslant t \leqslant 3) + I(t > 3)$. The censoring times of the two event types were both generated from a mixture distribution with probability $0{\cdot}5$ from a uniform distribution on $(3/2, 3)$ and probability $0{\cdot}5$ from a point mass at $\tau = 3$.

We chose θ to be 1 and $0{\cdot}5$ in the simulation study. When $\theta = 1$, the results are summarized in Table 1 for each of the four transformation models and sample size n. The confidence intervals were constructed based on the asymptotic normal approximation for $\hat{\beta}$, a lognormal approximation

Table 1. *Simulation results from* 1000 *replications when* $\theta = 1$

n	Par	True	Est	SE	ASE	CP	Par	True	Est	SE	ASE	CP
			$G_1(x) = x$						$G_2(x) = x$			
200	β_{11}	1·0	1·00	0·50	0·48	0·94	β_{21}	−1·5	−1·53	0·48	0·49	0·95
	β_{12}	−0·5	−0·50	0·30	0·29	0·94	β_{22}	0·5	0·51	0·27	0·27	0·96
	$\Lambda_1(\tau/4)$	0·555	0·56	0·06	0·06	0·94	$\Lambda_2(\tau/4)$	0·555	0·55	0·06	0·06	0·96
	θ	1·0	0·99	0·29	0·29	0·95						
400	β_{11}	1·0	0·99	0·35	0·34	0·94	β_{21}	−1·5	−1·50	0·33	0·34	0·96
	β_{12}	−0·5	−0·50	0·20	0·20	0·96	β_{22}	0·5	0·49	0·19	0·19	0·96
	$\Lambda_1(\tau/4)$	0·555	0·56	0·05	0·05	0·94	$\Lambda_2(\tau/4)$	0·555	0·56	0·05	0·05	0·94
	θ	1·0	0·99	0·20	0·20	0·96						
			$G_1(x) = x$						$G_2(x) = \log(1+x)$			
200	β_{11}	1·0	1·01	0·50	0·48	0·94	β_{21}	−1·5	−1·54	0·62	0·63	0·95
	β_{12}	−0·5	−0·50	0·30	0·29	0·94	β_{22}	0·5	0·51	0·35	0·36	0·96
	$\Lambda_1(\tau/4)$	0·555	0·56	0·06	0·07	0·94	$\Lambda_2(\tau/4)$	0·555	0·55	0·07	0·08	0·95
	θ	1·0	0·99	0·32	0·31	0·95						
400	β_{11}	1·0	0·99	0·35	0·34	0·95	$0\beta_{21}$	−1·5	−1·50	0·43	0·44	0·96
	β_{12}	−0·5	−0·50	0·20	0·20	0·96	β_{22}	0·5	0·49	0·25	0·25	0·95
	$\Lambda_1(\tau/4)$	0·555	0·56	0·05	0·05	0·94	$\Lambda_2(\tau/4)$	0·555	0·56	0·05	0·05	0·94
	θ	1·0	0·99	0·22	0·22	0·97						
			$G_1(x) = 2\{(1+x)^{1/2} - 1\}$						$G_2(x) = 2\log(1+x/2)$			
200	β_{11}	1·0	1·00	0·58	0·55	0·94	β_{21}	−1·5	−1·53	0·56	0·57	0·96
	β_{12}	−0·5	−0·50	0·34	0·32	0·94	β_{22}	0·5	0·51	0·31	0·32	0·97
	$\Lambda_1(\tau/4)$	0·555	0·56	0·07	0·07	0·95	$\Lambda_2(\tau/4)$	0·555	0·55	0·07	0·07	0·95
	θ	1·0	0·98	0·32	0·31	0·95						
400	β_{11}	1·0	1·00	0·40	0·39	0·95	β_{21}	−1·5	−1·50	0·39	0·40	0·95
	β_{12}	−0·5	−0·50	0·23	0·23	0·95	β_{22}	0·5	0·49	0·23	0·23	0·96
	$\Lambda_1(\tau/4)$	0·555	0·55	0·05	0·05	0·96	$\Lambda_2(\tau/4)$	0·555	0·56	0·05	0·05	0·94
	θ	1·0	0·99	0·22	0·22	0·96						
			$G_1(x) = \log(1+x)$						$G_2(x) = \log(1+x)$			
200	β_{11}	1·0	1·00	0·65	0·62	0·94	β_{21}	−1·5	−1·53	0·62	0·63	0·95
	β_{12}	−0·5	−0·49	0·38	0·37	0·95	β_{22}	0·5	0·51	0·35	0·36	0·96
	$\Lambda_1(\tau/4)$	0·555	0·56	0·08	0·08	0·95	$\Lambda_2(\tau/4)$	0·555	0·56	0·08	0·08	0·95
	θ	1·0	0·97	0·34	0·33	0·96						
400	β_{11}	1·0	0·99	0·45	0·44	0·95	β_{21}	−1·5	−1·49	0·43	0·44	0·96
	β_{12}	−0·5	−0·50	0·26	0·26	0·95	β_{22}	0·5	0·49	0·25	0·25	0·95
	$\Lambda_1(\tau/4)$	0·555	0·56	0·05	0·05	0·95	$\Lambda_2(\tau/4)$	0·555	0·56	0·05	0·05	0·94
	θ	1·0	0·98	0·23	0·23	0·97						

Par, the parameter to be estimated; True, the true value of the parameter; Est, the average estimate; SE, the sample standard deviation of the estimates; ASE, the average standard error; CP, the coverage probability of the nominal 95% confidence intervals.

for $\hat{\Lambda}$ and the Satterthwaite approximation for $\hat{\theta}$, the last two because $\hat{\Lambda}$ is positive and $\hat{\theta}$ is the estimated frailty variance. The results in Table 1 indicate that for both event types, the proposed method performs well with sample sizes of $n = 200$ and $n = 400$. In particular, the results show that the biases are small, the estimated standard errors agree well with the sample standard errors and the coverage probabilities range from 93% to 97%. The same results hold for the simulations with $\theta = 0{\cdot}5$. We conclude that when $\theta > 0$, the proposed estimation procedure provides asymptotically efficient estimates and good inferential properties for small sample sizes.

Table 2. *Analysis of cardiovascular data*

	Time to myocardial infarction			Time to stroke		
Covariates	Est	SE	*P*-value	Est	SE	*P*-value
Race (black)	0·651	0·280	0·020	−0·095	0·289	0·744
Age	0·053	0·020	0·008	0·093	0·018	0·000
Gender (male)	0·158	0·212	0·000	−0·115	0·198	0·561
Hypertension	0·330	0·137	0·016	0·519	0·111	0·000
BMI	−0·000	0·001	0·856	0·000	0·001	0·912
SBP	0·003	0·004	0·524	−0·001	0·001	0·509
Smoker	0·241	0·377	0·523	0·563	0·314	0·073
Diabetes	−0·000	0·001	0·866	−0·000	0·001	0·746
Frailty variance	0·987	0·390	0·011			

Est, estimate of the parameter; SE, standard error of the estimates.

Our proposed method also allows θ_0 to be zero, implying no frailty among events. The third simulation study considered this scenario under the same setup as the first two. The results are summarized as before. The only difference is that the confidence intervals were constructed using the half-normal distribution as described in Remark 1. The results, not shown, indicate that the confidence interval for θ is slightly conservative; however, the inferences for all other parameters are accurate. The Q–Q plots for $\hat{\theta}$, not shown, show that the estimated quantiles are almost linearly related to the quantiles of the half-normal distribution. The two distributions have a small discrepancy at zero with a sample size of 200, but this is much improved with a sample size of 400.

6. Application

We now use the proposed methods to analyze a single county subset of data from the atherosclerosis risk in communities study (ARIC Investigators, 1989). We defined T_1 as the time to myocardial infarction, and T_2 as the time to stroke. Other risk factors included race, baseline age, sex, BMI, systolic blood pressure, diabetes, hypertension and smoking status. The main goal is to identify risk factors for either or both of the time-to-event variables. The data we used contain 1212 patients who are more than 65 years of age, and the censoring rates for myocardial infarction and time to stroke were 89·2% and 86·8%, respectively.

Two transformations G_1 and G_2 need to be chosen using our approach. We consider all the transformations from the family (6), which include both the proportional hazards models and the proportional odds models. We calculate the likelihood function at a number of equally spaced grid points of (ρ, r) in increments of 0·1. The best model is that which yields the largest likelihood function. The results show that $G_1(x) = G_2(x) = \log(1 + x)$ is the best choice, indicating that the proportional odds model fits the data best among our choices of transformations. The estimates from this model are given in Table 2. The table shows that black patients have higher risk of developing myocardial infarction than white patients; older patients have higher risk of both time to myocardial infarction and time to stroke; male patients have higher risks of time to myocardial infarction; patients with hypertension appear to have higher risk of both events. Moreover, the frailty variance is significant with p-value 0·011, indicating a strong association between the two types of failure times.

One advantage of using frailty models for both types of failures is for prediction. To illustrate this, suppose that we want to predict the survival distribution of one patient in one group, given that the patient experiences a myocardial infarction event at year t_1. From the proposed model,

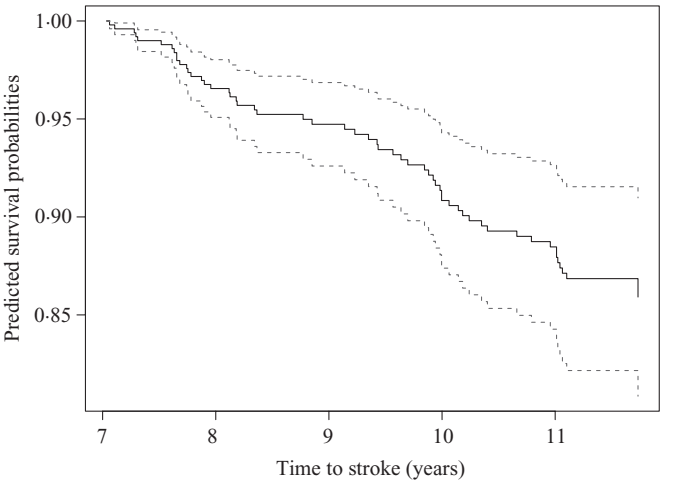

Fig. 1. Predicted survival probability of time to stroke for white male patients given that myocardial infarction occurred at seven years of follow-up; the solid line is the predicted survival curve and the dashed lines are the pointwise 95% confidence intervals.

we can easily obtain the survival probability of time to stroke of $P(T_2 > t \mid T_1 = t_1, T_2 > t_1, X)$, which is given by

$$\frac{\int_\omega \exp\left(-\omega\left[G_1\left\{F_1(t_1)\exp(\alpha_1+\beta_1^{\mathrm{T}}X)\right\}+G_2\left\{F_2(t)\exp(\alpha_2+\beta_2^{\mathrm{T}}X)\right\}\right]\right)\omega\,d\omega}{\int_\omega \exp\left(-\omega\left[G_1\left\{F_1(t_1)\exp(\alpha_1+\beta_1^{\mathrm{T}}X)\right\}+G_2\left\{F_2(t_1)\exp(\alpha_2+\beta_2^{\mathrm{T}}X)\right\}\right]\right)\omega\,d\omega}$$
$$=\left(\frac{1+\theta\left[G_1\left\{F_1(t_1)\exp(\alpha_1+\beta_1^{\mathrm{T}}X)\right\}+G_2\left\{F_2(t_1)\exp(\alpha_2+\beta_2^{\mathrm{T}}X)\right\}\right]}{1+\theta\left[G_1\left\{F_1(t_1)\exp(\alpha_1+\beta_1^{\mathrm{T}}X)\right\}+G_2\left\{F_2(t)\exp(\alpha_2+\beta_2^{\mathrm{T}}X)\right\}\right]}\right)^{1/\theta+1}. \quad (7)$$

Therefore, to predict the survival probability of time to stroke for a patient from one group, we can first estimate the above expression for each X_i in this group by substituting the maximum likelihood estimates into expression (7); we then take the average over all the patients in this group. Figure 1 gives the survival probability of time to stroke for white male patients, given that the patient had a myocardial infarction at seven years of follow-up.

7. REMARKS

The gamma frailty in our model is used for its computational tractability, but is likely to model early dependence (Hougaard, 2000). The proposed approach can be easily generalized to different frailty distributions, including the lognormal distribution and the positive stable distribution. It would also be interesting to consider a list of frailty distributions and consider how to check the frailty distribution empirically (Glidden, 2007).

Our models can be easily adapted to clustered failure times where only one failure time is of interest and subjects are sampled from clusters. In this case, we can use model (1) for each subject but assume the same G and β for each subject. The inference on nonparametric maximum likelihood estimators should be applicable in this case as well. Although we only consider time-independent covariates, time-dependent and external covariates can be easily incorporated into (1). Other possible generalizations include modelling multivariate or clustered counting processes.

Selecting an appropriate transformation is an important issue. The transformation $G(x)$ can be misspecified in practice due to limited knowledge or due to complex relationships between

the covariates and the time-to-event variable. In this paper, the function G was regarded as fixed. As a generalization, one may also specify a parametric family of functions and then estimate the relevant parameters. It is theoretically possible, although computationally demanding, to account for this extra variation. However, whether this kind of variation should be accounted for is debatable (Box & Cox, 1982). Nonparametric modelling and estimation of G is a challenging topic currently pursued by statisticians and econometricians in many contexts.

Appendix

Proof of Theorem 1. From the likelihood function $L_n(\alpha, \beta, \theta, F)$, we obtain that the joint survival distribution of $(T_1, \ldots, T_K)$ is given by $(1 + \theta[\sum_{k=1}^K G_k\{F_k(t_k)\exp(\alpha_k + \beta_k^{\mathrm{T}} X)\}])^{-1/\theta}$. To show identifiability, we first set

$$\left(1 + \theta\left[\sum_{k=1}^K G_k\left\{F_k(t_k)e^{\alpha_k+\beta_k^{\mathrm{T}}X}\right\}\right]\right)^{1/\theta} = \left(1 + \theta_0\left[\sum_{k=1}^K G_k\left\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\right\}\right]\right)^{1/\theta_0}, \tag{A1}$$

and then show $\theta = \theta_0$. We consider two situations.

We consider the first case when $\theta_0 > 0$. Suppose $\theta \neq \theta_0$. Without loss of generality, we assume $\theta > \theta_0$. For each $k = 1, \ldots, K$, let $t_l = 0$ for $l \neq k$. We obtain

$$\theta G_k\left\{F_k(t_k)e^{\alpha_k+\beta_k^{\mathrm{T}}X}\right\} = \left(1 + \theta_0\left[\sum_{k=1}^K G_k\left\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\right\}\right]\right)^{\theta/\theta_0} - 1.$$

Therefore, it follows that

$$\sum_{k=1}^K \left\{\left(1 + \theta_0\left[\sum_{k=1}^K G_k\left\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\right\}\right]\right)^{\theta/\theta_0} - 1\right\}$$
$$= \left(1 + \theta_0\left[\sum_{k=1}^K G_k\left\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\right\}\right]\right)^{\theta/\theta_0} - 1.$$

As the function $f(x) = (1 + x + y)^{\theta/\theta_0} - (1 + x)^{\theta/\theta_0}$ is strictly increasing for $x > 0$ given $y > 0$, it follows that $(1 + x + y)^{\theta/\theta_0} + 1 > (1 + x)^{\theta/\theta_0} + (1 + y)^{\theta/\theta_0}$. We therefore obtain a contradiction. Hence, $\theta = \theta_0$.

We then consider the second case when $\theta_0 = 0$. In this case, (A1) is equivalent to

$$\left(1 + \theta\left[\sum_{k=1}^K G_k\left\{F_k(t_k)e^{\alpha_k+\beta_k^{\mathrm{T}}X}\right\}\right]\right)^{1/\theta} = \exp\left[\sum_{k=1}^K G_k\left\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\right\}\right].$$

For each $k = 1, \ldots, K$, we let $t_l = 0$ for $l \neq k$ and obtain $\theta G_k\{F_k(t_k)e^{\alpha_k+\beta_k^{\mathrm{T}}X}\} = \exp[\theta G_k\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\}] - 1$. Therefore, we get

$$\exp\left[\theta\sum_{k=1}^K G_k\left\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\right\}\right] - 1 = \sum_{k=1}^K \left(\exp\left[\theta G_k\left(F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\right)\right] - 1\right).$$

As $e^{x+y} - e^x \geqslant e^y - 1$ for $x \geqslant 0$, with equality if and only if $x = 0$, we obtain $\theta = 0$.

In both cases, $\theta = \theta_0$ gives $F_k(t)\exp(\alpha_k + \beta_k^{\mathrm{T}} X) = F_{0k}(t)\exp(\alpha_{0k} + \beta_{0k}^{\mathrm{T}} X)$. By Assumption (A1) and the fact that $F_k(\tau) = 1$, we obtain $\alpha_k = \alpha_{0k}$, $\beta_k = \beta_{0k}$, $F_k = F_{0k}$. This establishes parameter identifiability.

Proof of Theorem 2. If the Fisher information along this submodel is singular, then the score function along this submodel is zero with probability one. To show that $w = 0$, in the score equations, we let $\Delta_k = 1$

and integrate Y_k from t_k to τ, then subtract that from the score equation for which $\Delta_k = 0$ and $Y_k = \tau$. We discuss the cases $\theta_0 > 0$ and $\theta_0 = 0$ separately.

When $\theta_0 > 0$, the score equation is

$$\begin{aligned}
&\sum_{k=1}^{K} G_k'\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\}e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\left\{\int_0^{t_k} h_k(s)dF_{0k}(s) + F_{0k}(t_k)\big(a_k + b_k^{\mathrm{T}}X\big)\right\}\\
&\quad + w\sum_{k=1}^{K} G_k\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\}\\
&\quad - \frac{w}{\theta_0^2}\left[1+\theta_0\sum_{k=1}^{K} G_k\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\}\right]\log\left[1+\theta_0\sum_{k=1}^{K} G_k\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\}\right] = 0.
\end{aligned}$$

For each $k = 1, \ldots, K$, let $t_l = 0$ for $l \neq k$, so we obtain

$$\begin{aligned}
&G_k'\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\}e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\left\{\int_0^{t_k} h_k(s)dF_{0k}(s) + F_{0k}(t_k)\big(a_k + b_k^{\mathrm{T}}X\big)\right\}\\
&\quad + wG_k\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\}\\
&\quad - \frac{w}{\theta_0^2}\big[1+\theta_0 G_k\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\}\big]\log\big[1+\theta_0 G_k\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\}\big] = 0.
\end{aligned}$$

Therefore, it follows that

$$\begin{aligned}
&w\sum_{k=1}^{K}\big[1+\theta_0 G_k\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\}\big]\log\big[1+\theta_0 G_k\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\}\big]\\
&\quad = w\left[1+\theta_0\sum_{k=1}^{K} G_k\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\}\right]\log\left[1+\theta_0\sum_{k=1}^{K} G_k\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\}\right].
\end{aligned}$$

As $f(x) = (1+x+y)\log(1+x+y) - (1+x)\log(1+x)$ is strictly increasing for $x > 0$ and $y > 0$, we have $(1+x+y)\log(1+x+y) > (1+x)\log(1+x) + (1+y)\log(1+y)$. We therefore conclude that $w = 0$.

When $\theta_0 = 0$, the score equation is

$$\begin{aligned}
&\sum_{k=1}^{K} G_k'\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\}e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\left\{\int_0^{t_k} h_k(s)dF_{0k}(s) + F_{0k}(t_k)\big(a_k + b_k^{\mathrm{T}}X\big)\right\}\\
&\quad - \frac{w}{2}\left[\sum_{k=1}^{K} G_k\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\}\right]^2 = 0.
\end{aligned}$$

Following the same arguments as before, we have

$$\frac{w}{2}\left[\sum_{k=1}^{K} G_k\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\}\right]^2 = \frac{w}{2}\sum_{k=1}^{K}\left[G_k\{F_{0k}(t_k)e^{\alpha_{0k}+\beta_{0k}^{\mathrm{T}}X}\}\right]^2,$$

which holds if and only if $w = 0$.

In both cases, $w = 0$ implies $\int_0^{t_k} h_k(s)\,dF_{0k}(s) + F_{0k}(t_k)(a_k + b_k^{\mathrm{T}}X) = 0$. Let $t_k = \tau$. Since $\int_0^{\tau} h_k(s)\,dF_{0k}(s) = 0$ and $F_{0k}(\tau) = 1$, we have $a_k = 0$ and $b_k = 0$. This further gives $h_k(s)\,dF_{0k}(s) = 0$. We have thus proved nonsingularity of the Fisher information matrix along any nontrivial submodel.

Proof of Theorem 3. We recall that $\Lambda_k(t) = F_k(t)e^{\alpha_k}$. Thus, nonparametric maximum likelihood estimation for $(\alpha, \beta, \theta, F)$ is equivalent to maximizing

$$l_n(\theta, \beta_1, \ldots, \beta_K, \Lambda_1, \ldots, \Lambda_K) = \sum_{i=1}^{n}\sum_{k=1}^{K} \Delta_{ik} \log\left[G_k'\left\{\Lambda_k(Y_{ik})e^{\beta_k^{\mathrm{T}}X_i}\right\}\Lambda_k\{Y_{ik}\}e^{\beta_k^{\mathrm{T}}X_i}\right]$$
$$+ \log\frac{\Gamma\left(\sum_{k=1}^{K}\Delta_{ik} + 1/\theta\right)}{\Gamma(1/\theta)(1/\theta)^{\sum_{k=1}^{K}\Delta_{ik}}} - \left(\sum_{k=1}^{K}\Delta_{ik} + \frac{1}{\theta}\right)\log\left(1 + \theta\left[\sum_{k=1}^{K} G_k\left\{\Lambda_k(Y_{ik})e^{\alpha_k+\beta_k^{\mathrm{T}}X_i}\right\}\right]\right).$$

The corresponding nonparametric maximum likelihood estimators are denoted by $(\hat\theta, \hat\beta_k, \hat\Lambda_k, k = 1, \ldots, K)$. Therefore, proving Theorem 3 is equivalent to establishing consistency for $(\hat\theta, \hat\beta, \hat\Lambda)$.

We define $\Psi_k(t) = G_k\{\Lambda_k(t)\}$. By the mean-value theorem,

$$\Psi_k\{t\} = G_k\{\Lambda_k(t)\} - G_k\{\Lambda_k(t-)\} = \Lambda_k\{t\}G_k'\{\nu\Lambda_k(t) + (1-\nu)\Lambda_k(t-)\}$$

for some $\nu \in [0, 1]$. Since G_k' is nonincreasing by Assumption (A2), we obtain $\Lambda_k\{t\} \leqslant \Psi_k\{t\}/G_k'\{\Lambda_k(t)\}$. Thus, the loglikelihood function l_n is bounded above by

$$O_p(n) + \sum_{i=1}^{n}\sum_{k=1}^{K}\Delta_{ik}\log\Psi_k\{Y_{ik}\} + \sum_{i=1}^{n}\sum_{k=1}^{K}\Delta_{ik}\log\frac{G_k'\left\{\Lambda_k(Y_{ik})e^{\beta_k^{\mathrm{T}}X_i}\right\}}{G_k'\{\Lambda_k(Y_{ik})\}}$$
$$- \sum_{i=1}^{n}\left(\sum_{i=1}^{K}\Delta_{ik} + \frac{1}{\theta}\right)\log\left(1 + \theta\left[\sum_{k=1}^{K} G_k\left\{\Lambda_k(Y_{ik})e^{\beta_k^{\mathrm{T}}X_i}\right\}\right]\right).$$

From Assumption (A2),

$$\frac{G_k'\left\{\Lambda_k(Y_{ik})e^{\beta_k^{\mathrm{T}}X_i}\right\}}{G_k'\{\Lambda_k(Y_{ik})\}} \leqslant c_0, \quad G_k\left\{\Lambda_k(Y_{ik})e^{\beta_k^{\mathrm{T}}X_i}\right\} \geqslant \frac{1}{c_0}G_k\{\Lambda_k(Y_{ik})\}.$$

We have

$$l_n(\theta, \beta_1, \ldots, \beta_K, \Lambda_1, \ldots, \Lambda_K)$$
$$\leqslant O_p(n) + \sum_{i=1}^{n}\sum_{k=1}^{K}\Delta_{ik}\log\frac{\Psi\{Y_{ik}\}}{1 + \theta\sum_{k=1}^{K}\Psi(Y_{ik})} - \frac{1}{\theta}\sum_{i=1}^{n}\log\left\{1 + c_0^{-1}\theta\sum_{k=1}^{K}\Psi(Y_{ik})\right\}. \quad \text{(A2)}$$

If Y_{ik} is the smallest observation so that $\hat\Lambda_k\{Y_{ik}\} = \infty$, then clearly $\hat\Psi_k\{Y_{ik}\} = \infty$. However, from (A2), the loglikelihood function will be $-\infty$ whenever $\theta \geqslant 0$. Hence, we conclude that $\hat\Lambda_k\{Y_{ik}\} < \infty$ for all Y_{ik}. The same holds for $\hat\Psi_k$.

To prove consistency, we first show that $\limsup_n \hat\Lambda_k(\tau) < \infty$ with probability one. Otherwise, suppose that for some subsequence and some k, $\hat\Lambda_k(\tau) \to \infty$. Since $1 + \hat\theta[\sum_{k=1}^{K} G_k\{\hat\Lambda_k(\tau)\exp(\hat\beta_k^{\mathrm{T}}X_i)\}] > 0$, we immediately conclude that $\liminf_n \hat\theta \geqslant 0$. By choosing a subsequence, we may assume $\hat\theta \to \theta^*$ and $\hat\beta_k \to \beta_k^*$ for $k = 1, \ldots, K$. Note $\theta_M \geqslant \theta^* \geqslant 0$. There are two cases.

Let $\theta^* = 0$. After differentiating with respect to $\Lambda_k\{Y_{ik}\}$, we obtain $\hat\Lambda_k(t) = \int_0^t S_{nk} \times (t; \hat\Lambda_1, \ldots, \hat\Lambda_K, \hat\beta, \hat\theta)^{-1} dN_k(t)$, where $N_k(t) = n^{-1}\sum_{i=1}^{n}\Delta_{ik}I(Y_{ik} \leqslant t)$ and

$$S_{nk}(t; \Lambda_1, \ldots, \Lambda_K, \beta, \theta) = -n^{-1}\sum_{i=1}^{n}\Delta_{ik}\frac{G_k''\left\{\Lambda_k(Y_{ik})e^{\beta^{\mathrm{T}}X_{ik}}\right\}}{G_k'\left\{\Lambda_k(Y_{ik})e^{\beta^{\mathrm{T}}X_{ik}}\right\}}I(Y_{ik} \geqslant t)$$
$$+ n^{-1}\sum_{i=1}^{n}\frac{\theta\sum_{k=1}^{K}\Delta_{ik} + 1}{\theta\sum_{k=1}^{K}G_k\left\{\Lambda_k(Y_{ik})e^{\beta_k^{\mathrm{T}}X_i}\right\} + 1}I(Y_{ik} \geqslant t).$$

From assumption (A2), when θ is small and satisfies $\theta \sum_{k=1}^{K} G_k\{\Lambda_k(Y_{ik})e^{\beta_k^{\mathrm{T}} X_i}\} + 1 > 0$, we have

$$\begin{aligned} S_{nk}(t;\Lambda_1,\ldots,\Lambda_K,\beta,\theta) &\geqslant \frac{1}{2n}\sum_{i=1}^{n}\frac{1}{\mid\theta\mid\sum_{k=1}^{K} G_k\{\Lambda_k(Y_{ik})e^{\beta_k^{\mathrm{T}} X_i}\}+1} I(Y_{ik}\geqslant t), \\ &\geqslant \frac{1}{2n}\sum_{i=1}^{n}\frac{1}{\mid\theta\mid\sum_{k=1}^{K} G_k\{\Lambda_k(\tau)e^{\beta_k^{\mathrm{T}} X_i}\}+1} I(Y_{ik}\geqslant t), \\ &\geqslant \frac{1}{2n}\sum_{i=1}^{n}\frac{1}{c_1\{1+\mid\theta\mid\sum_{k=1}^{K}\Lambda_k(\tau)M\}} I(Y_{ik}\geqslant t), \end{aligned}$$

where $c_1 = \max(c_0, 1)$ and M is the supremum of $e^{\beta_k^{\mathrm{T}} X}$. Therefore, $\hat{\Lambda}_k(\tau) \leqslant 2c_1\{1 + |\hat{\theta}|\sum_{k=1}^{K}\hat{\Lambda}_k(\tau)M\}\int_0^{\tau}\{n^{-1}\sum_{i=1}^{n} I(Y_{ik}\geqslant t)\}^{-1} dN_k(t)$. Since $\hat{\theta}$ converges to zero, we obtain a contradiction to $\hat{\Lambda}_k(\tau)\to\infty$ for some k.

Suppose $\theta^* > 0$. We can assume $\hat{\theta} > \theta^*/2$. Note that $\hat{\Lambda}_k(\tau)\to\infty$ is equivalent to $\hat{\Psi}_k(\tau)\to\infty$. From (A2), since $\theta\in(\theta^*/2,\theta_M)$, the observed loglikelihood function at the nonparametric maximum likelihood estimators is further bounded above by

$$\begin{aligned} &l_n(\hat{\theta},\hat{\beta}_1,\ldots,\hat{\beta}_K,\hat{\Lambda}_1,\ldots,\hat{\Lambda}_K) \\ &\leqslant O_p(n) + \sum_{i=1}^{n}\sum_{k=1}^{K}\Delta_{ik}\log\frac{\hat{\Psi}_k\{Y_{ik}\}}{1+\hat{\Psi}_k(Y_{ik})} - \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{1}{\theta_M}\log\left\{1+c_0^{-1}\theta_M\sum_{k=1}^{K}\hat{\Psi}_k(Y_{ik})\right\} \\ &\leqslant O_p(n) + \sum_{k=1}^{K}\left[\sum_{i=1}^{n}\Delta_{ik}\log\frac{\hat{\Psi}_k\{Y_{ik}\}}{1+\hat{\Psi}_k(Y_{ik})} - \sum_{i=1}^{n}\frac{1}{\theta_M}\log\{1+\hat{\Psi}_k(Y_{ik})\}\right]. \end{aligned}$$

To complete the proof, we define $\tilde{\Lambda}_k(t) = \int_0^t S_{nk}(t;\Lambda_{01},\ldots,\Lambda_{0K},\beta_0,\theta_0)^{-1} dN_k(t)$. By a direct check, we can show $\tilde{\Lambda}_k\to\Lambda_{0k}$ uniformly in $[0,\tau]$ and $n\tilde{\Lambda}_k\{Y_{ik}\} = O_p(1)$. Thus, from

$$0\leqslant n^{-1}\{l_n(\hat{\theta},\hat{\beta}_1,\ldots,\hat{\beta}_K,\hat{\Lambda}_1,\ldots,\hat{\Lambda}_K) - l_n(\theta_0,\beta_{01},\ldots,\beta_{0K},\Lambda_{01},\ldots,\Lambda_{0K})\},$$

we obtain

$$0\leqslant O_p(1) + \sum_{k=1}^{K}\left[n^{-1}\sum_{i=1}^{n}\Delta_{ik}\log\frac{n\hat{\Psi}_k\{Y_{ik}\}}{1+\hat{\Psi}_k(Y_{ik})} - n^{-1}\sum_{i=1}^{n}\frac{1}{\theta_M}\log\{1+\hat{\Psi}_k(Y_{ik})\}\right].$$

Each term for the kth type event has a similar expression as expression (25) in Parner (1998). Thus, following the same arguments as in Parner (1998), we can show that the right-hand side diverges to $-\infty$ if $\hat{\Lambda}_k(\tau)\to\infty$ for some k. A similar argument is given in the Appendix of Zeng & Lin (2007).

We have shown that $\hat{\Lambda}_k$ is uniformly bounded. By choosing a subsequence, we can assume $\hat{\Lambda}_k$ converges weakly to Λ_k^*. From the expression, $\hat{\Lambda}_k(t) = \int_0^t\{S_{nk}(s;\Lambda_{01},\ldots,\Lambda_{0K},\beta_0,r_0)\}\{S_{nk}(s;\hat{\Lambda}_1,\ldots,\hat{\Lambda}_K,\hat{\beta},\hat{r})\}^{-1} d\tilde{\Lambda}_k(s)$, and after applying the same arguments as in Zeng & Lin (2007), we conclude that $\hat{\Lambda}_k$ is dominated by $\tilde{\Lambda}_k$ and the derivative converges. On the other hand, by examining the observed loglikelihood function, we obtain

$$\begin{aligned} 0\leqslant\ & n^{-1}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\Delta_{ik}\left[\log\frac{G_k'\{\hat{\Lambda}_k(Y_{ik})e^{\hat{\beta}^{\mathrm{T}} X_{ik}}\}}{G_k'\{\tilde{\Lambda}_k(Y_{ik})e^{\beta_0^{\mathrm{T}} X_{ik}}\}} + \log\frac{\hat{\Lambda}_k\{Y_{ik}\}}{\tilde{\Lambda}_k\{Y_{ik}\}} + \hat{\beta}^{\mathrm{T}} X_{ik} - \beta_0^{\mathrm{T}} X_{ik}\right]\right) \\ &+ n^{-1}\sum_{i=1}^{n}\log\frac{\int\omega^{\sum_{k=1}^{K}\Delta_{ik}}\exp\big(-\omega\big[\sum_{k=1}^{K} G_k\{\hat{\Lambda}_k(Y_{ik})e^{\hat{\beta}^{\mathrm{T}} X_{ik}}\}\big]\big)g_{\hat{\theta}}(\omega)\,d\omega}{\int\omega^{\sum_{k=1}^{K}\Delta_{ik}}\exp\big(-\omega\big[\sum_{k=1}^{K} G_k\{\Lambda_{0k}(Y_{ik})e^{\beta_0^{\mathrm{T}} X_{ik}}\}\big]\big)g_{\theta_0}(\theta_0)\,d\omega}. \end{aligned}$$

Now we take limits on both sides and define $F_k^* = \Lambda_k^*/\Lambda_k^*(\tau)$ and $\alpha^* = \log\Lambda_k^*(\tau)$. Then we obtain that the Kullback–Leibler information for $(f_1^*,\ldots,f_K^*,\alpha_1^*,\ldots,\alpha_K^*,\beta_1^*,\ldots,\beta_K^*,\theta^*)$ is less than or equal to zero.

Therefore, from the identifiability result proved earlier, $F_k^* = F_{0k}$, $\alpha_k^* = \alpha_{0k}$, $\beta^* = \beta_0$ and $\theta^* = \theta_0$. This establishes consistency of $(\hat{F}_k, \hat{\alpha}_k, \hat{\beta}_k, k = 1, \ldots, K)$ and $\hat{\theta}$.

References

The ARIC Investigators. (1989). The Atherosclerosis Risk in Communities (ARIC) Study: Design and Objectives. *Am. J. Epidemiol.* **129**, 687–702.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statist. Med.* **2**, 273–7.

Berkson, J. & Gage, R. P. (1952). Survival curve for cancer patients following treatment. *J. Am. Statist. Assoc.* **47**, 501–15.

Box, G. E. P. & Cox, D. R. (1982). An analysis of transformations revisited, rebutted. *J. Am. Statist. Assoc.* **77**, 209–10.

Chen, K., Jin, Z. & Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika* **89**, 659–68.

Cheng, S. C., Wei, L. J. & Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835–45.

Clayton, D. G. & Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *J. R. Statist. Soc.* A **148**, 82–117.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B **34**, 187–220.

Dabrowska, D. M. & Doksum, K. A. (1988). Partial likelihood in transformation models with censored data. *Scand. J. Statist.* **18**, 1–23.

Glidden, D. V. (2007). Pairwise dependence diagnostics for clustered failure-time data. *Biometrika* **94**, 371–85.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data.* New York: Springer.

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. Hoboken, NJ: Wiley.

Klein, J. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **48**, 795–806.

Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: the marginal approach. *Statist. Med.* **13**, 2233–47.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc.* B **13**, 2233–47.

Murphy, S. A. (1994). Consistency in a proportional hazards model incorporating a random effect. *Ann. Statist.* **22**, 712–31.

Murphy, S. A. (1995). Asymptotic theory for the frailty model. *Ann. Statist.* **23**, 182–98.

Murphy, S. A., Rossini, A. J. & van der Vaart, A. W. (1997). Maximal likelihood estimation in the proportional odds model. *J. Am. Statist. Assoc.* **92**, 968–76.

Nielsen, G. G., Gill, R. D., Andersen, P. K. & Sorensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Statist.* **19**, 25–43.

Oakes, D. (1989). Bivariate survival models induced by frailties. *J. Am. Statist. Assoc.* **84**, 487–93.

Oakes, D. (1991). Frailty models for multiple event times. In *Surv. Analysis: State of the Art* (Ed. J. P. Klein and P. K. Goel) **84**, 371–9. Dordrecht: Kluwer.

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *Ann. Statist.* **26**, 183–214.

Zeng, D. & Lin, D. Y. (2006). Maximum likelihood estimation in semiparametric transformation models for counting processes. *Biometrika* **93**, 627–40.

Zeng, D. & Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data (with discussion). *J. R. Statist. Soc.* B **69**, 507–64.

Zeng, D., Yin, G. & Ibrahim, J. G. (2006). Semiparametric transformation models for survival data with a cure fraction. *J. Am. Statist. Assoc.* **101**, 670–84.

[*Received March* 2008. *Revised October* 2008]