Statistics and Its Interface Volume 1 (2008) 33–45

A penalized likelihood approach to joint modeling of longitudinal measurements and time-to-event data

Wen Ye, Xihong Lin and Jeremy M. G. Taylor

Recently joint models for longitudinal and time-to-event data have attracted a lot attention. A full joint likelihood approach using an EM algorithm or Bayesian methods of estimation not only eliminates the bias in naive and two-stage methods, but also improves efficiency. However, both the EM algorithm and a Bayesian method are computationally intensive, limiting the utilization of these joint models. We propose to use an estimation procedure based on a penalized joint likelihood generated by Laplace approximation of a joint likelihood and by using a partial likelihood instead of the full likelihood for the event time data. The results of a simulation study show that this penalized likelihood approach performs as well as the corresponding EM algorithm under a variety of scenarios, but only requires a fraction of the computational time. An additional advantage of this approach is that it does not require estimation of the baseline hazard function. The proposed procedure is applied to a data set for evaluating the effect of the longitudinal biomarker PSA on the recurrence of prostate cancer.

Keywords and phrases: Penalized likelihood, Survival, Longitudinal, Joint modeling, Laplace approximation.

---

1. INTRODUCTION

Clinical research studies, such as cancer and AIDS studies, often collect both longitudinal and time-to-event data. Typically, in such studies the longitudinal measurement is susceptible to measurement error and only measured intermittently. In addition, many subjects do not survive throughout the study and the risk of event is related to the longitudinal measurements. Due to these complications analyzing such data using separate longitudinal and hazard models often leads to bias in parameter estimates in both models (Faucett and Thomas, 1996; Raboud, Reid, Coates and Farewell, 1993; Little, 1995).

Recently joint models for longitudinal and time-to-event data have attracted a lot attention. By jointly maximizing a likelihood from both the covariate process and time-to-event data, one uses information from both sources to get parameter estimates for the two processes simultaneously. By doing so, the informative drop-out can be adjusted for by borrowing information from the survival model, and, conversely, unbiased information from the longitudinal covariate can be incorporated into the survival model, and therefore, one can correctly assess the dependence of the failure time on the longitudinal marker. As well as correcting biases, joint modeling can improve the efficiency of parameter estimates in either part of the model, because extra information is being used. Faucett and Thomas (1996) proposed a joint model in which they described the failure process with a proportional hazards Cox model and the longitudinal data with a linear growth curve model, and took a Bayesian approach to estimate the joint posterior distribution of model parameters. Wulfsohn and Tsiatis (1997) assumed a similar model but took a frequentist approach, with an EM algorithm to estimate parameters. Their work has been further extended by many authors (Xu and Zeger, 2001a; Xu and Zeger, 2001b; Wang and Taylor, 2001; Pauler and Finkelstein, 2002; Wu, 2002; Brown and Ibrahim, 2003b; Brown and Ibrahim, 2003a; Tsiatis and Davidian, 2004; Tseng, Hsieh and Wang, 2005; Ratcliffe, Guo and Ten Have, 2004), with a review presented in Yu, Law, Taylor and Sandler (2004).

However, both EM algorithm and Bayesian method are computationally intensive, limiting the utilization of these joint models. In addition, when a more complicated longitudinal model is used, an EM algorithm may not be feasible. For example, when the longitudinal model requires a large number of random effects to capture the structure of the observations, the EM algorithm for estimation in such models involves p dimensional integration in the E-step, where p is the number of random effects. When p is large, the task is very challenging and might be even impossible. Therefore, to make joint modeling more practical, our goal is to first build a more efficient computational algorithm for joint model estimation.

In this paper, we propose an estimation procedure that maximizes a penalized joint likelihood (PJL) generated by a Laplace approximation of the joint likelihood, which allows us to avoid high dimensional integration. In addition, we use a partial likelihood instead of the full likelihood for the time-to-event data. This further simplifies the estimation by avoiding estimation of the baseline hazard function. We test the efficiency and quality of this estimation procedure and compare it to the EM algorithm in a basic joint model setting.

The remainder of this paper is organized as follows. Section 2 describes the models and Section 3 describes the estimating procedure. Section 4 presents some simulation results. Section 5 applies this method to the Michigan Prostate Cancer Study (MPCS). Conclusions and discussions are offered in Section 6.

2. MODELS

For simplicity we assume a linear growth curve model with random intercept and slope for the longitudinal covariate measurements. At each time point the covariate is measured, we observe

$$x_i(t_{ij}) = b_{0i} + b_{1i}t_{ij} + \varepsilon_{ij} \tag{1}$$

We assume that the measurement error ε_{ij} is independently distributed $N(0, \sigma_\varepsilon^2)$ and the vector of subject-specific random effects $\mathbf{b}_i = (b_{0i}, b_{1i})$ is normally distributed with mean $\mu_i = (\mu_{0i}, \mu_{1i})$ and covariance matrix $D(\phi)$, where $\phi = (\sigma_1^2, \sigma_2^2, \sigma_{12})$. Further, we assume that the mean parameter $\mu_i = (\mu_{0i}, \mu_{1i})$ is a function of some time-independent covariates,

$$\begin{aligned} \mu_{0i} &= W_{10i}^T \boldsymbol{\beta}_0 \\ \mu_{1i} &= W_{11i}^T \boldsymbol{\beta}_1 \end{aligned}$$

where W_{10i}^T is a vector of $1 + p_{10}$ elements, including the number 1 and p_{10} time-independent covariates related to the intercept of subject i. W_{11i}^T is a vector of $1 + p_{11}$ elements, including the number 1 and p_{11} time-independent covariates related to the slope for subject i.

We model the hazard of failure through a proportional hazards model, where the hazard depends on the longitudinal covariate $x_i(\cdot)$ through its current value $x_i(t) = b_{0i} + b_{1i}t$ and its rate of change b_{1i},

$$\lambda_i\{t|\mathbf{b}_i, W_{2i}^T\} = \lambda_0(t) exp\left\{\alpha_1(b_{0i} + b_{1i}t) + \alpha_2 b_{1i} + W_{2i}^T\boldsymbol{\gamma}\right\} \tag{2}$$

where W_{2i}^T (p_2 elements) is the vector of covariates related to the hazard directly.

Let T_i and C_i be the event and censoring time for the ith subject $(i = 1, \ldots, m)$. We only observe $Y_i = \min(T_i, C_i)$ along with an indicator Δ_i that is equal to 1 if the subject fails and equal to 0 if the subject is censored. Independent censoring is assumed. Let $X_i = (x_{i1}, \ldots, x_{in_i})^T$ be the vector of the observed longitudinal measurements of subject i at time $(t_{i1}, \ldots, t_{in_i})$.

3. APPROXIMATE LIKELIHOOD AND ESTIMATION

3.1 Penalized joint likelihood

Let $S_i(t|\mathbf{b}_i, W_{2i}^T)$ be the survival function corresponding to the hazard function $\lambda_i(t|\mathbf{b}_i, W_{2i}^T)$ defined in (2), and let $f(x_{i1}, \ldots, x_{in_i}|\mathbf{b}_i)$ be the joint density function of the observed longitudinal measurements given $\mathbf{b}_i$ for subject i. The full joint likelihood function for the models we defined in section 2 is

$$\begin{aligned} &L(\lambda_0(t), \alpha_1, \alpha_2, \boldsymbol{\gamma}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \theta) \\ &= \int \left\{ \prod_{i=1}^m \lambda_i(t|\mathbf{b}_i, W_{2i}^T)^{\Delta_i} S_i(t|\mathbf{b}_i, W_{2i}^T) f(x_{i1}, \ldots, x_{in_i}|\mathbf{b}_i) \right\} \\ &\quad \times p(\mathbf{b}|V(\phi))d\mathbf{b} \\ &= \int \prod_{i=1}^m \left(\lambda_0(Y_i) exp\{\boldsymbol{\alpha}\mathbf{L_i}(Y_i)\mathbf{b} + W_{2i}^T\boldsymbol{\gamma}\}\right)^{\Delta_i} \\ &\quad \times e^{-\int_0^{Y_i} \lambda_0(u) exp\{\boldsymbol{\alpha}\mathbf{L_i}(u)\mathbf{b} + W_{2i}^T\boldsymbol{\gamma}\} du} \\ &\times \prod_{i=1}^m \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} exp\left\{-\frac{(X_{ij} - \mathbf{Z_i}(t_{ij})\mathbf{b})^2}{2\sigma_\varepsilon^2}\right\} \times p(\mathbf{b}|\boldsymbol{\mu}, V(\phi))d\mathbf{b} \end{aligned} \tag{3}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$, $\mathbf{b} = (b_{01}, b_{11}, b_{02}, b_{12}, \ldots, b_{0m}, b_{1m})$, $\boldsymbol{\mu} = (\mu_{01}, \mu_{11}, \mu_{02}, \mu_{12}, \ldots, \mu_{0m}, \mu_{1m})$, and $\theta = (\sigma_1^2, \sigma_2^2, \sigma_{12}, \sigma_\varepsilon^2)$. $\mathbf{L_i}(u)$ is a $2 \times 2m$ matrix with its $(1, i \times 2 - 1)^{th}$ and $(2, i \times 2)^{th}$ element equals to 1, its $(1, i \times 2)^{th}$ element equals to u, and its $(2, i \times 2 - 1)^{th}$ elements equals to zero. $p(\mathbf{b}|\boldsymbol{\mu}, V(\phi))$ is the density function of $\mathbf{b}$ given $\boldsymbol{\mu}$ and $V(\phi)$, which is a $2m \times 2m$ block diagonal matrix with all the blocks equal to $D(\phi)$. $\mathbf{Z_i}(t_{ij})$ is a $1 \times 2m$ matrix with its $(i \times 2 - 1)^{th}$ element equals to 1 and its $(1, i \times 2)^{th}$ element equals to t_{ij}.

We can rewrite equation (3) in the form of $I = c \cdot |V(\phi)|^{-\frac{1}{2}} \cdot \sigma_\varepsilon^{-\sum_{i=1}^m n_i} \cdot \int e^{K(\mathbf{b})} d\mathbf{b}$, where

$$\begin{aligned} K(\mathbf{b}) = -\sum_{i=1}^m \Bigg(&\Delta_i \left(log(\lambda_0(Y_i)) + \boldsymbol{\alpha}\mathbf{L_i}(Y_i)\mathbf{b} + W_{2i}^T\boldsymbol{\gamma}\right) \\ &- \int_0^{Y_i} \lambda_0(u) exp\{\boldsymbol{\alpha}\mathbf{L_i}(u)\mathbf{b} + W_{2i}^T\boldsymbol{\gamma}\} du \Bigg) \\ &+ \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{(X_{ij} - \mathbf{Z_i}(t_{ij})\mathbf{b})^2}{2\sigma_\varepsilon^2} \\ &+ \frac{1}{2}(\mathbf{b} - \boldsymbol{\mu})^T V(\phi)^{-1}(\mathbf{b} - \boldsymbol{\mu}) \end{aligned}$$

Then applying the first order Laplace approximation (Breslow and Clayton, 1993) and ignoring the multiplicative c, the approximation yields

$$\begin{aligned} &log(L(\lambda_0(t), \alpha_1, \alpha_2, \boldsymbol{\gamma}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \theta)) \\ &\approx -\frac{m}{2} log|D(\phi)| - \sum_{i=1}^m \frac{n_i}{2} log(\sigma_\varepsilon^2) - \frac{1}{2} log|K''(\hat{\mathbf{b}})| - K(\hat{\mathbf{b}}) \end{aligned} \tag{4}$$

where $\hat{\mathbf{b}} = argmax_{\mathbf{b}} K(\hat{\mathbf{b}})$.

Ripatti and Palmgren (2000) used a similar Laplace approximation in a frailty model setting in survival analysis and named the approximate likelihood function as Penalized Partial Likelihood (PPL). They noted that if the variance components θ are known and $\mathbf{b}$ were considered a fixed effects parameter, then the last term in (4), $-K(\mathbf{b})$, would be a penalized log likelihood (Green, 1987), where $-\frac{1}{2}(\mathbf{b}-\boldsymbol{\mu})^T V(\phi)^{-1}(\mathbf{b}-\boldsymbol{\mu})$ is the penalty term penalizing extreme values of $\mathbf{b}$. Further, to avoid estimating the nonparametric baseline function, we replace the full survival likelihood in $K(\mathbf{b})$ with a partial likelihood and obtain

(5)

$$\begin{aligned} K_{PJL}(\mathbf{b}) \quad = \quad & \sum_{i=1}^{m} \Delta_i \Bigg\{ \boldsymbol{\alpha}\mathbf{L_i}(Y_i)\mathbf{b} + W_{2i}^T\boldsymbol{\gamma} \\ & - log \sum_{j\in R(Y_i)} exp\{\boldsymbol{\alpha}\mathbf{L_j}(Y_i)\mathbf{b} + W_{2i}^T\boldsymbol{\gamma}\} \Bigg\} \\ & - \sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{(X_{ij} - \mathbf{Z_i}(t_{ij})\mathbf{b})^2}{2\sigma_\varepsilon^2} \\ & - \frac{1}{2}(\mathbf{b}-\boldsymbol{\mu})^T V^{-1}(\mathbf{b}-\boldsymbol{\mu}) \end{aligned}$$

where $R(Y_i)$ is the set of subjects at risk at the event time Y_i. The corresponding term $log|K''(\mathbf{b})|$ in (4) is also replaced by $log|K''_{PJL}(\mathbf{b})|$. However, since omitting this complicated term $log|K''_{PJL}(\mathbf{b})|$ has negligible affect on the parameter estimate results, we omit it, and estimate $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\gamma}$ and $\mathbf{b}$ by maximizing $-K_{PJL}(\mathbf{b})$, which is referred to as Penalized Joint Likelihood (PJL) in the rest of the paper.

For given $D(\phi)$, σ_ε^2, and $\boldsymbol{\alpha}$ the estimating equations based on the first partial derivatives of the PJL are the following. For $\mathbf{b}$,

(6)

$$\begin{aligned} & \sum_{k=1}^{m} \Delta_i \Bigg((\boldsymbol{\alpha}\mathbf{Z_i}(Y_i))^T \\ & \qquad - \frac{\sum_{j\in R(Y_i)} exp\{\boldsymbol{\alpha}\mathbf{L_j}(Y_i)\mathbf{b} + W_{2j}^T\boldsymbol{\gamma}\}(\boldsymbol{\alpha}\mathbf{L_j}(Y_i))^T}{\sum_{j\in R(Y_i)} exp\{\boldsymbol{\alpha}\mathbf{L_j}(Y_i)\mathbf{b} + W_{2j}^T\boldsymbol{\gamma}\}} \Bigg) \\ & \quad + \frac{1}{\sigma_\varepsilon^2}\sum_{j=1}^{n_i}(X_{ij} - \mathbf{Z_i}(t_{ij})\mathbf{b})\mathbf{Z_i}^T(t_{ij}) - V^{-1}(\phi)(\mathbf{b}-\boldsymbol{\mu}) = 0 \end{aligned}$$

for $\boldsymbol{\gamma}$,

(7)

$$\sum_{k=1}^{m} \Delta_i \left(W_{2i} - \frac{\sum_{j\in R(Y_i)} exp\{\boldsymbol{\alpha}\mathbf{L_j}(Y_i)\mathbf{b} + W_{2j}^T\boldsymbol{\gamma}\} W_{2j}}{\sum_{j\in R(Y_i)} exp\{\boldsymbol{\alpha}\mathbf{L_j}(Y_i)\mathbf{b} + W_{2j}^T\boldsymbol{\gamma}\}} \right) = 0$$

and for $\boldsymbol{\beta} = \{\boldsymbol{\beta}_0, \boldsymbol{\beta}_1\}$,

(8) $$\mathbf{W}_1 V^{-1}(\phi)\mathbf{b} - \mathbf{W}_1 V^{-1}(\phi)\mathbf{W}_1^T\boldsymbol{\beta} = 0$$

where

$$\mathbf{W}_1^T = \begin{pmatrix} W_{101} & \mathbf{0} \\ \mathbf{0} & W_{111} \\ W_{102} & \mathbf{0} \\ \mathbf{0} & W_{112} \\ \vdots & \vdots \\ W_{10m} & \mathbf{0} \\ \mathbf{0} & W_{11m} \end{pmatrix}$$

The estimates $(\hat{\mathbf{b}}(\boldsymbol{\alpha},\theta), \hat{\boldsymbol{\gamma}}(\boldsymbol{\alpha},\theta), \hat{\boldsymbol{\beta}}(\boldsymbol{\alpha},\theta))$ can be found by alternating between solving (6), (7), and (8). Note that, the solution to equation (8) is of the simple form $\hat{\boldsymbol{\beta}} = (\mathbf{W}_1 V^{-1}(\phi)\mathbf{W}_1^T)^{-1}\mathbf{W}_1 V^{-1}(\phi)\mathbf{b}$. Corresponding information matrices can be found in Appendix B.

3.2 Risk coefficient and variance component estimation

We estimate the risk coefficient $\boldsymbol{\alpha}$ and each parameter $(\sigma_1^2, \sigma_2^2, \sigma_{12}, \sigma_\varepsilon^2)$ in the variance components θ by using a profile likelihood approach. The approximate profile log likelihood function for $\boldsymbol{\alpha}$ and θ is,

(9)

$$\begin{aligned} & l(\hat{\mathbf{b}}(\boldsymbol{\alpha},\theta), \hat{\boldsymbol{\gamma}}(\boldsymbol{\alpha},\theta), \hat{\boldsymbol{\mu}}(\boldsymbol{\alpha},\theta), \boldsymbol{\alpha}, \theta) \\ & \quad = -\frac{m}{2}log|D(\phi)| - \sum_{i=1}^{m}\frac{n_i}{2}log(\sigma_\varepsilon^2) - \frac{1}{2}log\left|K''(\hat{\mathbf{b}})\right| \\ & \qquad + \sum_{i=1}^{m}\Bigg(\Delta_i\left(log(\lambda_0(Y_i)) + \boldsymbol{\alpha}\mathbf{L_i}(Y_i)\hat{\mathbf{b}} + W_{2j}^T\hat{\boldsymbol{\gamma}}\right) \\ & \qquad - \int_0^{Y_i}\lambda_0(u)exp\{\boldsymbol{\alpha}\mathbf{L_i}(u)\hat{\mathbf{b}} + W_{2j}^T\hat{\boldsymbol{\gamma}}\}du\Bigg) \\ & \qquad - \sum_{i=i}^{m}\sum_{j=1}^{n_i}\frac{(X_{ij} - \mathbf{Z_i}(t_{ij})\hat{\mathbf{b}})^2}{2\sigma_\varepsilon^2} \\ & \qquad - \frac{1}{2}(\hat{\mathbf{b}}-\hat{\boldsymbol{\mu}})^T V(\phi)^{-1}(\hat{\mathbf{b}}-\hat{\boldsymbol{\mu}}) \end{aligned}$$

Again, we substitute $K(\hat{\mathbf{b}})$ with $K(\hat{\mathbf{b}})_{PJL}$ and $K''(\hat{\mathbf{b}})$ with $K''_{PJL}(\hat{\mathbf{b}})$, where

$$\begin{aligned} K_{PJL}(\hat{\mathbf{b}}) \quad = \quad & -\sum_{i=1}^{m}\Delta_i\Bigg\{\boldsymbol{\alpha}\mathbf{L_i}(Y_i)\hat{\mathbf{b}} + W_{2j}^T\hat{\boldsymbol{\gamma}} \\ & - log\sum_{j\in R(Y_i)} exp\{\boldsymbol{\alpha}\mathbf{L_j}(Y_i)\hat{\mathbf{b}} + W_{2j}^T\hat{\boldsymbol{\gamma}}\}\Bigg\} \\ & + \sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{(X_{ij} - \mathbf{Z_i}(t_{ij})\hat{\mathbf{b}})^2}{2\sigma_\varepsilon^2} \\ & + \frac{1}{2}(\hat{\mathbf{b}}-\hat{\boldsymbol{\mu}})^T V(\phi)^{-1}(\hat{\mathbf{b}}-\hat{\boldsymbol{\mu}}) \end{aligned}$$

and the corresponding

$$(10)\quad K''_{PJL}(\hat{\mathbf{b}}) = \sum_{i=1}^{m}\Delta_i\left(\frac{Sb2}{S0} - \frac{Sb1^{\otimes 2}}{S0^2}\right) + \frac{1}{\sigma_\varepsilon^2}\sum_{j=1}^{n_i}\mathbf{Z}_i^T(t_{ij})\mathbf{Z}_i(t_{ij}) + V(\phi)^{-1}$$

$S0$, $Sb1$, $Sb2$, and the meaning of the symbol $\otimes$ are given in Appendix A. In contrast to the situation in the objective function for estimating $\mathbf{b}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$, to estimate variance components θ and $\boldsymbol{\alpha}$, $log|K''_{PJL}(\mathbf{b})|$ is an important term and can not be omitted. Therefore, the objective function to maximize is

$$(11)\quad -\frac{m}{2}log|D(\phi)| - \sum_{i=1}^{m}\frac{n_i}{2}log(\sigma_\varepsilon^2) - \frac{1}{2}log|K''_{PJL}(\hat{\mathbf{b}})| - K_{PJL}(\hat{\mathbf{b}})$$

The corresponding estimating equation for $\boldsymbol{\alpha}$ is

$$(12)\quad -\frac{1}{2}tr\left(K''_{PJL}(\hat{\mathbf{b}})^{-1}\frac{\partial K''_{PJL}(\hat{\mathbf{b}})}{\partial\boldsymbol{\alpha}}\right) + \sum_{i=1}^{m}\Delta_i\left(\mathbf{L_i}(Y_i)\hat{\mathbf{b}} - \frac{\sum_{j\in R(Y_i)} exp\{\boldsymbol{\alpha}\mathbf{L_j}(Y_i)\hat{\mathbf{b}} + W_{2j}^T\boldsymbol{\gamma}\}\mathbf{L_j}(Y_i)\hat{\mathbf{b}}}{\sum_{j\in R(Y_i)} exp\{\boldsymbol{\alpha}\mathbf{L_j}(Y_i)\hat{\mathbf{b}} + W_{2j}^T\boldsymbol{\gamma}\}}\right) = 0$$

Because of the restriction $\sigma_{12}^2 \le \sigma_1^2\sigma_2^2$ on parameters $\phi = \{\sigma_1^2, \sigma_2^2, \sigma_{12}\}$, to simplify computation, we reparameterize ϕ as: σ_1^2, σ_2^2, and ρ, where $\rho = \sigma_{12}/\sigma_1\sigma_2$. The estimating equation for ϕ can be written in the general form

$$-\frac{1}{2}\left[tr\left(V^{-1}(\phi)\frac{\partial V(\phi)}{\partial\phi}\right) + tr\left(K''_{PJL}(\hat{\mathbf{b}})^{-1}\frac{\partial V^{-1}(\phi)}{\partial\phi}\right) + (\hat{\mathbf{b}} - \hat{\boldsymbol{\mu}})^T\frac{\partial V^{-1}(\phi)}{\partial\phi}(\hat{\mathbf{b}} - \hat{\boldsymbol{\mu}})\right] = 0$$

For σ_ε^2, the estimating equation is

$$(13)\quad -\sum_{i=1}^{m}\frac{n_i}{2\sigma_\varepsilon^2} - \frac{1}{2}tr\left(K''_{PJL}(\hat{\mathbf{b}})^{-1}\frac{-1}{(\sigma_\varepsilon^2)^2}\sum_{j=1}^{n_i}\mathbf{Z}_i^T(t_{ij})\mathbf{Z}_i(t_{ij})\right) + \sum_{i=i}^{m}\sum_{j=1}^{n_i}\frac{(X_{ij} - \mathbf{Z_i}(t_{ij})\hat{\mathbf{b}})^2}{2(\sigma_\varepsilon^2)^2} = 0$$

Note that equation (13) has a closed form solution.

3.3 Computation

Initial values for all the parameters where obtained using a two-stage method (Dafni and Tsiatis, 1998). Then each parameter is updated using a one-step Newton-Raphson iteration or closed form solution to the estimating equation, if the closed form exists. Updating steps for the parameters are iterated until convergence. After convergence, the baseline hazard function can be estimated by using Breslow's estimator (Breslow, 1974).

3.4 Inference

Standard errors for $\mathbf{b}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be approximated by inverting the information matrix of the penalized joint partial likelihood (5). It is an $(m \times 2 + 2 + p_{10} + p_{11} + p_2) \times (m \times 2 + 2 + p_{10} + p_{11} + p_2)$ matrix

$$\begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} & \mathcal{I}_{13} \\ \mathcal{I}_{21} & \mathcal{I}_{22} & \mathcal{I}_{23} \\ \mathcal{I}_{31} & \mathcal{I}_{32} & \mathcal{I}_{33} \end{pmatrix}$$

the components of which are given in Appendix C.

The standard errors for the variance components $\sigma_1^2, \sigma_2^2, \rho$, and σ_ε^2 and $\boldsymbol{\alpha}$ can be obtained by inverting the 6×6 observed information matrix $\mathcal{J}$ given in the Appendix D. The delta method is used to further obtain variance estimates for σ_{12}.

4. SIMULATION STUDY

We carried out a simulation study to evaluate the proposed penalized joint partial likelihood estimator by comparing it to a two-stage estimator (ordinary regression calibration estimator) and a full joint likelihood estimator which is estimated using the EM algorithm (EM). The three main questions we address are:

a. How do measurement error, censoring rate, and the true value of the risk coefficient affect the quality of the PJL estimator and its inference?
b. An attractive feature of the EM joint modeling approach is its robustness against departure from the normal random effects assumption. Hsieh, Tseng and Wang (2006) provide a theoretical explanation for this feature and demonstrate it numerically. Is the PJL estimator also robust to departure of the assumed Gaussian random effect distribution of $\mathbf{b_i}$?
c. Is the PJL method computationally faster than the EM method?

The two-stage method we use is similar to the regression calibration procedure in Dafni and Tsiatis (1998). But we only fit one linear mixed model to the longitudinal data of all subjects and the predicted longitudinal covariate values for each subject still at risk at each failure time are based on this single fit. The EM algorithm we use is the version described in Wulfsohn and Tsiatis (1997). We use ten grid points for all the Gaussian quadrature integrations.

For simplicity, we assume that the subject-specific intercept and slope in function (1) do not depend on any time-independent covariate. In addition, we assume the hazard in

Table 1. Simulation Results: Effect of Measurement Error on Parameter Estimates. Entries Are Relative Bias of the Monte Carlo Average of Estimates (RB), Monte Carlo Standard Deviation (SD), Monte Carlo Average of Estimated Standard Errors (SE), Monte Carlo Coverage Probability of Nominal 95% Confidence Interval (CP), and Average Computing Time for Each Simulation Repetition (ACT)

	Two-stage				PJL				EM	
	RB	SD	SE	CP	RB	SD	SE	CP	RB	SD
$\sigma_\varepsilon^2 = 0.09$										
μ_0	-0.6	0.156	0.159	0.94	-0.3	0.156	0.158	0.94	-0.4	0.156
μ_1	129.1	0.009	0.009	0.67	-4.3	0.01	0.009	0.95	1.1	0.01
$\boldsymbol{\alpha}$	4.6	0.071	0.071	0.88	-0.7	0.085	0.075	0.89	0.3	0.081
σ_1^2	1.1	0.006	0.006	0.91	0	0.006	0.006	0.91	0.1	0.006
σ_ε^2	0.4	0.494	0.506	0.94	0.2	0.492	0.475	0.92	0.2	0.492
σ_2^2	-2.8	0.002	0.002	0.92	-0.8	0.002	0.001	0.85	-1.2	0.002
σ_{12}	-40.8	0.026	0.026	0.91	2.3	0.026	0.017	0.84	1.1	0.026
ACT					5 min				45 min	
$\sigma_\varepsilon^2 = 0.3$										
μ_0	-0.9	0.159	0.16	0.95	-0.3	0.159	0.16	0.94	-0.3	0.159
μ_1	250.9	0.01	0.01	0.27	-11.5	0.011	0.01	0.95	6.2	0.011
$\boldsymbol{\alpha}$	10.1	0.069	0.069	0.67	-2	0.097	0.078	0.87	0.9	0.087
σ_1^2	2.2	0.021	0.02	0.93	-0.2	0.021	0.018	0.91	-0.1	0.021
σ_ε^2	-0.3	0.51	0.518	0.94	0.1	0.508	0.474	0.91	0.2	0.508
σ_2^2	-8.8	0.002	0.002	0.85	0.3	0.002	0.001	0.81	-1.1	0.002
σ_{12}	-67.2	0.028	0.027	0.81	1.9	0.029	0.018	0.79	-1.5	0.028
ACT					7 min				57 min	
$\sigma_\varepsilon^2 = 0.6$										
μ_0	-1.1	0.162	0.162	0.96	-0.2	0.163	0.163	0.94	-0.3	0.163
μ_1	354.8	0.011	0.01	0.03	-22	0.012	0.011	0.93	11.1	0.012
$\boldsymbol{\alpha}$	15.7	0.067	0.068	0.39	-3.8	0.113	0.081	0.83	1.6	0.093
σ_1^2	3	0.042	0.04	0.94	-0.2	0.041	0.035	0.91	-0.2	0.041
σ_ε^2	-1.3	0.53	0.533	0.93	0	0.527	0.474	0.93	0.2	0.527
σ_2^2	-13.7	0.002	0.002	0.76	1.3	0.002	0.001	0.81	-1.2	0.002
σ_{12}	-87.9	0.029	0.028	0.68	2.8	0.03	0.018	0.77	-3.4	0.03
ACT					8 min				69 min	

survival model (2) is only determined by the current longitudinal covariate value $x_i(t)$, i.e. $\alpha_2 = 0$ and $\boldsymbol{\gamma} = \mathbf{0}$, therefore

$$\lambda\{t; b_i\} = \lambda_0(t) exp\{\alpha(b_{0i} + b_{1i}t)\} \tag{14}$$

We first describe the basic setting of the model considered in our simulation, then explain the variation of the relevant factors in each of these simulation studies. Our basic setting was similar to that described in Tsiatis and Davidian (2004), except that $\sigma_\varepsilon^2 = 0.3$. Following their paper, we set $\{\mu_0, \mu_1\} = \{4.173, -0.0103\}$, $\{\sigma_1^2, \sigma_{12}, \sigma_2^2\} = \{4.96, -0.0456, 0.012\}$, and $\alpha = -1$. For each simulation study, we simulated 100 data sets of sample size N = 200 based on the joint model defined by (1) and (14). The properties examined include relative bias (RB), the empirical standard error (EMP_SE), the average of the estimated standard error (EST_SE), and the coverage rate of the nominal 95% confidence interval (CP).

The impact of measurement error is shown in Table 1. The parameter estimates from the two-stage method are all severely biased. Both the EM and the PJL methods greatly reduce the bias of the two-stage estimators, especially improving the estimate of μ_1. PJL method provides estimates for α, μ_0, σ_ε^2, σ_1^2, σ_2^2, and σ_{12} as good as those from the EM method. However the PJL estimator for μ_1 seems to have slightly more bias than the corresponding EM estimator. But both have substantially less bias than the two-stage estimator. Even when the measurement error is relatively large, both the EM and the PJL methods yield acceptable results. By comparing the empirical standard errors between the PJL and the EM methods, one can see that the PJL method is as efficient as the EM method with regard to all the parameter estimates except α. The relative efficiency of the PJL estimator for α compared to the EM estimator is 92.6%, 89.7%, and 81.7% when $\sigma_\varepsilon^2 = 0.09$, 0.3, and 0.6, respectively. The Monte Carlo average of the estimated stan-

Table 2. Simulation Results: Effect of Censoring Rate on Parameter Estimates. Entries Are Relative Bias of the Monte Carlo Average of Estimates (RB), Monte Carlo Standard Deviation (SD), Monte Carlo Average of Estimated Standard Errors (SE), Monte Carlo Coverage Probability of Nominal 95% Confidence Interval (CP), and Average Computing Time for Each Simulation Repetition (ACT)

	Two-stage				PJL				EM	
	RB	SD	SE	CP	RB	SD	SE	CP	RB	SD
Censoring Rate = 30% (Censor only at the end of the study)										
μ_0	-0.5	0.168	0.158	0.93	0.3	0.168	0.159	0.93	0.2	0.168
μ_1	265	0.01	0.009	0.19	-12.8	0.011	0.01	0.91	3.7	0.011
$\boldsymbol{\alpha}$	10.3	0.067	0.064	0.59	-2.7	0.085	0.073	0.94	0.2	0.079
σ_ε^2	2.5	0.016	0.017	0.96	0.6	0.015	0.015	0.98	0.1	0.015
σ_1^2	-2.4	0.465	0.505	0.91	-1.8	0.46	0.467	0.91	-1.8	0.46
σ_2^2	-9.8	0.002	0.002	0.81	1.1	0.002	0.001	0.84	-2.3	0.002
σ_{12}	-63.7	0.026	0.025	0.82	4.9	0.027	0.017	0.8	5	0.027
ACT					7 min				58 min	
Censoring Rate = 40%										
μ_0	-0.4	0.168	0.159	0.93	0.3	0.168	0.159	0.93	0.2	0.169
μ_1	257.9	0.011	0.01	0.24	-3.8	0.012	0.01	0.91	13	0.012
$\boldsymbol{\alpha}$	10.4	0.066	0.07	0.62	-2.5	0.087	0.079	0.96	0.8	0.079
σ_ε^2	3.1	0.02	0.02	0.95	0.8	0.019	0.018	0.94	0.2	0.019
σ_1^2	-2.3	0.467	0.508	0.93	-1.8	0.462	0.465	0.91	-1.8	0.462
σ_2^2	-7.7	0.002	0.002	0.86	1.3	0.002	0.001	0.82	-1.8	0.002
σ_{12}	-67	0.026	0.027	0.84	1.6	0.027	0.017	0.83	1.6	0.027
ATC					5 min				45 min	
Censoring Rate = 70%										
μ_0	-0.1	0.168	0.16	0.92	0.2	0.169	0.16	0.91	0.1	0.169
μ_1	210.7	0.014	0.013	0.61	5.8	0.015	0.013	0.93	19.2	0.015
$\boldsymbol{\alpha}$	9.1	0.08	0.089	0.78	-2.6	0.109	0.1	0.97	1.4	0.097
σ_ε^2	2.9	0.035	0.035	0.96	-0.2	0.035	0.026	0.84	-0.5	0.034
σ_1^2	-1.9	0.468	0.515	0.93	-1.9	0.465	0.461	0.91	-1.8	0.464
σ_2^2	0	0.003	0.003	0.91	2	0.003	0.002	0.74	0.2	0.003
σ_{12}	-70	0.035	0.034	0.87	-0.7	0.036	0.018	0.64	-1.4	0.036
ATC					5 min				21 min	

dard errors in the PJL method shows that the PJL method slightly underestimates the standard error of the parameter estimates, especially for the variance component parameters. This is not unexpected, as Ripatti and Palmgren (2000) have shown a similar phenomena. This is probably due to both the Laplace approximation and replacing the full survival likelihood function with the partial likelihood function. Also, we use profile likelihood functions to estimate $\mathbf{b}$ and $\boldsymbol{\mu}$, and the variance components θ and α separately. Hence variance estimation for the variance components doesn't incorporate the variation of $\mathbf{b}$ and $\boldsymbol{\mu}$ and vice versa. The resultant coverage rate of 95% confidence interval is slightly lower than the nominal 95%. However, since our main interest is in α and $\boldsymbol{\mu}$, but not θ, this underestimation is not a big concern.

Table 2 examines the effect of censoring. In the basic setting, censoring rate is approximately 40%. We change this by first setting all subjects censored at the end of the study. This results in an 30% marginal censoring rate. In the third situation, the parameter λ in the censoring Weibull function is adjusted to obtain a 70% marginal censoring rate. Under all the three scenarios the PJL and the EM methods are much better than the two-stage method. As censoring rate increases, bias in the μ_1 and α estimates using the EM algorithm both increase. Bias in μ_1 estimate using the PJL method is negative when censoring rate is 30%, but it changes to positive as censoring rate increases to 70%. Bias in the α estimate using the PJL method does not change as censoring rate changes. But in general, the quality of the EM and PJL estimators are comparable under all the three settings.

We also examine the effect of the risk coefficient $\boldsymbol{\alpha}$ by varying its value from the basic setting $\alpha = -1$ to -0.5 and -1.5. The corresponding censoring rates are approximately 15%, 40%, and 55% for $\alpha = -0.5, -1$, and -1.5, respectively. Table 3 shows that the two-stage estimator for

Table 3. Simulation Results: Effect of True Value of Risk Coefficient α on Parameter Estimates. Entries Are Relative Bias of the Monte Carlo Average of Estimates (RB), Monte Carlo Coverage Probability of Nominal 95% Confidence Interval (CP), and Average Computing Time for Each Simulation Repetition (ACT)

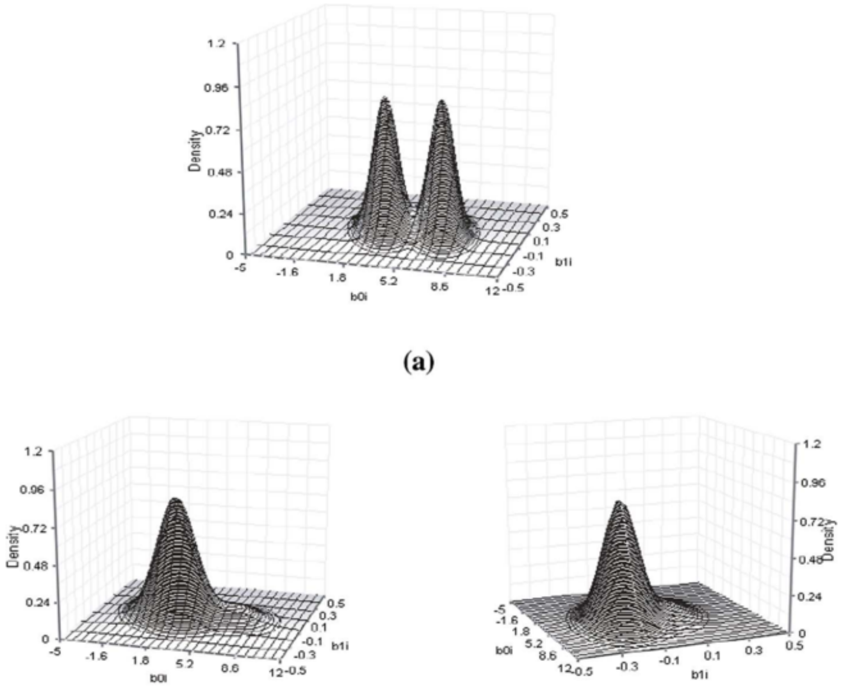
	Two-stage				PJL				EM	
	RB	SD	SE	CP	RB	SD	SE	CP	RB	SD
$\alpha = -0.5$										
μ_0	0.2	0.153	0.16	0.98	0.7	0.154	0.161	0.98	0.5	0.154
μ_1	344.5	0.019	0.014	0.34	6.4	0.017	0.014	0.92	7.5	0.018
$\boldsymbol{\alpha}$	0.9	0.053	0.046	0.92	−1.3	0.058	0.047	0.86	−1.1	0.058
σ_ε^2	3.3	0.039	0.036	0.91	0.7	0.037	0.028	0.84	1.5	0.037
σ_1^2	−1.4	0.527	0.518	0.94	−0.6	0.53	0.463	0.92	−0.8	0.526
σ_2^2	−1.1	0.003	0.003	0.92	−2.1	0.003	0.002	0.74	−3.4	0.003
σ_{12}	−73.3	0.04	0.039	0.82	−10.5	0.037	0.017	0.62	−8.3	0.04
ACT					13 min				97 min	
$\alpha = -1$										
μ_0	−0.1	0.155	0.159	0.97	0.6	0.155	0.16	0.98	0.5	0.155
μ_1	255.2	0.012	0.01	0.27	−6.9	0.012	0.01	0.9	9.1	0.012
$\boldsymbol{\alpha}$	10.1	0.078	0.07	0.63	−3	0.11	0.079	0.87	0.4	0.096
σ_ε^2	1.5	0.019	0.02	0.95	−0.9	0.018	0.018	0.97	−1.5	0.018
σ_1^2	−1.2	0.529	0.513	0.93	−0.8	0.531	0.468	0.88	−0.8	0.532
σ_2^2	−8.7	0.002	0.002	0.83	0.3	0.002	0.001	0.86	−2.8	0.002
σ_{12}	−71.8	0.026	0.027	0.79	−3	0.028	0.017	0.78	−2.9	0.028
ACT					5 min				47 min	
$\alpha = -1.5$										
μ_0	−0.1	0.156	0.159	0.97	0.5	0.156	0.159	0.96	0.4	0.156
μ_1	191.6	0.01	0.009	0.43	0.7	0.011	0.009	0.9	19.1	0.011
$\boldsymbol{\alpha}$	28.1	0.114	0.098	0.2	−7.8	0.227	0.132	0.83	7.2	0.152
σ_ε^2	1.3	0.017	0.017	0.96	−0.9	0.016	0.016	0.95	−1.5	0.016
σ_1^2	−0.7	0.532	0.514	0.92	−0.7	0.531	0.469	0.88	−0.6	0.534
σ_2^2	−10.2	0.001	0.002	0.8	−0.1	0.002	0.001	0.86	−3.3	0.001
σ_{12}	−65.2	0.023	0.024	0.78	−7.7	0.023	0.017	0.82	−6.4	0.023
ACT					4 min				33 min	

α behaves worse as α increases, but the bias in μ_1 estimate decreases slightly. Both the PJL and the EM estimators are much better than the two-stage estimator. When $\alpha = -1.5$, both the EM α estimate and PJL α estimate have more bias.

To investigate the second question, we study the behavior of the PJL estimator under a non-normal scenario. Similar to Tsiatis and Davidian (2004) we also consider a model in which $\mathbf{b_i}$ is distributed as a bivariate mixture of normals. The random effects densities were generated by mixing two normal distributions $N(\nu, \Xi)$ and $N(-\nu, \Xi)$ with the mixing proportion $\pi = \{\varpi, 1-\varpi\}$ (Davidian and Gallant, 1993). We consider three cases. The first case mimics the scenario in Tsiatis and Davidian (2004), $\nu = \{2, 0\}$ and $\varpi = 0.5$, which gives a bimodal distribution for the random intercept. In the first of the two additional cases, $\nu = \{2.5, 0\}$ and $\varpi = 0.1$. This generates a skewed distribution for the random intercepts. In the last case $\nu = \{0, 0.12\}$ and $\varpi = 0.1$, which generates a skewed distribution for the random slopes. In all the three cases, the covariance matrix Ξ were adjusted accordingly such that the covariance matrix of the random effects has distinct elements as described in the above normal cases. Figure 1 shows the true density of $\mathbf{b_i}$ in these three situations.

Table 4 shows the simulation results. Compared to the results under the basic setting ($\sigma_\varepsilon^2 = 0.3$) in Table 1, we find that the behavior of the two-stage estimators does not change much in all the three non-normal situations. All the three α estimates are slightly more attenuated than the one in Table 1. The bias in the other parameter estimates can be slightly better or slightly worse. In all the three situations, both the PJL and the EM estimators behave better than the two-stage estimator. However, compared to the normal situation, both these two methods yield more bias. The magnitude of the bias is similar between all the PJL and EM parameter estimates.

Finally, to answer the last question, for each of the above simulation studies, we show (Tables 1–4) the average computing time for both the PJL algorithm and the EM algo-



(a)

(b) **(c)**

Figure 1. The True Density of b_i *for the Three Bimodal Cases. (a)* $\nu = \{2, 0\}$ *and* $\varpi = 0.5$*; (b)* $\nu = \{2.5, 0\}$ *and* $\varpi = 0.1$*; (c)* $\nu = \{0, 0.12\}$ *and* $\varpi = 0.1$.

rithm. For all of our simulation studies, the PJL takes much less time than the EM method. In addition, for the EM method calculating the asymptotic variance-covariance matrix for the EM estimator requires a separate Supplementing EM (SEM) procedure which is very time consuming. Therefore, we didn't calculate the estimated standard errors of the parameter estimates. In contrast, the PJL procedure gives variance estimates by simply inverting the observed information matrix. If one would like the EM procedure to include computation of the variance estimates computation, it would take significantly more computational time.

In the above simulation studies we have considered several important factors that affect the quality of PJL estimator and its inference. There are certainly other factors we could consider. For example, the result may be affected by inclusion of covariates for the fixed effects in the longitudinal model and in the survival model. In addition, the range in the values for the mean and variance components of the subject-specific intercept and slope may also affect the result.

5. APPLICATION TO A PROSTATE CANCER STUDY

To demonstrate this new approach, we apply it to a study of cancer recurrence in prostate cancer patients after radiation therapy (Taylor, Yu and Sandler, 2005). A major interest of this study is to estimate the dependence of the risk of cancer recurrence on the prostate specific antigen (PSA) level and the rate of change of the PSA level. A total of 934 patients were enrolled from July 1987 to February 2000.

Since the start of radiation therapy, post-treatment longitudinal PSA values were measured about every 6 months to the time of clinical recurrence of cancer, lost of follow-up, death from non-prostate cancer disease, or censoring at the end of the study. The median follow-up from entry of the study was 44.5 months (range: 0.6–144.5 months). Clinical recurrence (local recurrence, distant metastasis or regional recurrence) was observed in 140 (15.0%) of the 934 patients. All other patients are treated as censored. During the study follow-up, a total of 59 patients received salvage hormone therapy (HT), among whom 15 (25.4%) experienced clinical failure. PSA measurements after HT in these patients were not included in the data analysis. The average number of PSA measurements is 6.6 ± 4.1 (range: $1 \sim 29$). Tumor Stage III and above are used as the reference group against Tumor stage I and II. The baseline covariates include patient age at the entry of the study (Age), baseline PSA (B_PSA), tumor stage (T-stage), Gleason score (GS) and total dose of radiation therapy (Dose). The PSA values measured at the baseline and the follow-up were log e-base transformed as LNPSA = ln(PSA+1), B_LNPSA = ln(B_PSA+1), to make the normality assumption more plausible.

Table 4. Simulation Results: Robustness of Parameter Estimates. Entries Are Relative Bias of the Monte Carlo Average of Estimates (RB), Monte Carlo Standard Deviation (SD), Monte Carlo Average of Estimated Standard Errors (SE), Monte Carlo Coverage Probability of Nominal 95% Confidence Interval (CP), and Average Computing Time for Each Simulation Repetition (ACT)

	Two-stage				PJL				EM	
	RB	SD	SE	CP	RB	SD	SE	CP	RB	SD
Bimodal distribution for intercept: $\nu = \{2, 0\}$; $\varpi = 0.5$										
μ_0	-0.6	0.145	0.16	0.96	-0.1	0.146	0.16	0.97	-0.2	0.146
μ_1	229.3	0.011	0.01	0.33	-63.5	0.013	0.01	0.82	-45.4	0.012
$\boldsymbol{\alpha}$	13.7	0.074	0.07	0.5	2	0.102	0.08	0.85	4.7	0.091
σ_ε^2	1.2	0.018	0.02	0.98	-1.1	0.018	0.018	0.95	-1.1	0.018
σ_1^2	-1.1	0.304	0.513	0.99	-0.8	0.307	0.48	0.99	-0.9	0.306
σ_2^2	-7.9	0.002	0.002	0.88	0.8	0.002	0.001	0.88	-0.7	0.002
σ_{12}	-51.6	0.027	0.028	0.9	43.5	0.03	0.017	0.66	39.1	0.03
ACT					4 min				29 min	
Skewed distribution for intercept: $\nu = \{2.5, 0\}$; $\varpi = 0.1$										
μ_0	-1	0.17	0.16	0.93	-0.3	0.168	0.16	0.94	-0.3	0.169
μ_1	208	0.01	0.01	0.38	-55.1	0.011	0.01	0.91	-36.6	0.011
$\boldsymbol{\alpha}$	12.5	0.067	0.073	0.63	-0.6	0.084	0.082	0.96	2.2	0.076
σ_ε^2	0.9	0.02	0.02	0.96	-1.2	0.02	0.018	0.92	-1.1	0.02
σ_1^2	-0.2	0.667	0.515	0.87	0	0.66	0.483	0.85	0	0.661
σ_2^2	-10	0.002	0.002	0.85	1.5	0.002	0.001	0.87	-0.1	0.002
σ_{12}	-28.5	0.022	0.024	0.95	28.9	0.023	0.018	0.77	26	0.023
ACT					4 min				43 min	
Skewed distribution for slope: $\nu = \{0, 0.12\}$; $\varpi = 0.1$										
μ_0	-0.3	0.144	0.16	0.98	0.4	0.143	0.16	0.97	0.4	0.143
μ_1	223.1	0.011	0.01	0.42	-51	0.011	0.011	0.91	-32	0.011
$\boldsymbol{\alpha}$	10.9	0.072	0.069	0.59	-1.9	0.093	0.078	0.92	1	0.084
σ_ε^2	0.7	0.022	0.02	0.92	-1.1	0.021	0.018	0.86	-1.1	0.021
σ_1^2	-0.5	0.489	0.516	0.97	-0.2	0.487	0.476	0.94	-0.2	0.487
σ_2^2	6.2	0.002	0.002	0.88	14.1	0.003	0.002	0.72	12.6	0.002
σ_{12}	-67.7	0.029	0.029	0.76	8	0.03	0.019	0.77	4.5	0.03
ACT					4 min				38 min	

The major interest of this study is to estimate the dependence of the risk of cancer recurrence on the PSA level and the rate of change of the PSA level ($LNPSA'(t)$). Since the increase of PSA level after remission is a signal for cancer recurrence, and the PSA curve of most patients reached their nadir around 20 weeks or earlier after radiation therapy, we only consider the post-twenty-week data and assume that LNPSA follows the linear model in (1). The covariates in W_{10}^T include GS, T-stage, Age, B_LNPSA, and Dose. The covariates in W_{11}^T include GS, T-stage, and B_LNPSA.

We also assume the hazard for cancer recurrence follows model (2), where the covariates in W_2^T include GS, T-stage, Age, Dose, and B_LNPSA. HT was included in the model as a time-dependent covariate, where HT(t) equals to 1 if a subject has had the hormone therapy at time t, 0 otherwise.

The post-twenty-week data contains 658 subject, among which 86 (13%) have experienced cancer recurrence. Forty-six of these patients received Hormone therapy and nine of them experienced cancer recurrence. The average number of PSA measurements is 4.5 ± 3.4 (range: $1 \sim 26$).

The results are shown in Table 5. LNPSA level is a significant predictor of cancer recurrence. If a patient's PSA level increases ten folds, the risk of cancer recurrence increases by approximately 230%. The rate of change of LNPSA level does not turn out to be significant. A plausible explanation is as follows. For subjects who experienced cancer recurrence, the rate of change of PSA often increases. Although this increase is rather small, we probably still underestimate the slope of increase of PSA at the time of cancer recurrence by assuming linear PSA trajectory. This might lead to the underestimation of the effect of $LNPSA'(t)$. Other important survival pre-

Table 5. Data Analysis Results: Application to the Prostate Cancer Study

Model	Parameter	Estimates	Standard Error
Longitudinal	Intercept	1.512	0.525
	Gleason Score	−0.051	0.021
	Stage1	0.269	0.120
	Stage2	0.105	0.114
	Age	−0.00254	0.00330
	Dose	−0.0176	0.0060
	Baseline lPSA	0.230	0.037
	Time	−0.0166	0.0082
	Gleason Score × Time	0.00289	0.00089
	Stage1 × Time	−0.0171	0.0051
	Stage2 × Time	−0.0127	0.0048
	Baseline lPSA × Time	0.0118	0.0016
	σ_ε^2	0.0322	0.0010
	σ_1^2	0.253	0.012
	σ_2^2	0.000567	0.000027
	σ_{12}	−0.00523	0.00050
Survival	lPSA	1.066	0.122
	Slope of lPSA	−0.421	4.734
	Gleason Score	0.488	0.102
	Stage1	−1.560	0.429
	Stage2	−0.675	0.317
	Age	−0.0493	0.0183
	Dose	−4.453	4.196
	Basline lPSA	−0.114	0.123
	Hormone Therapy	−1.938	0.439

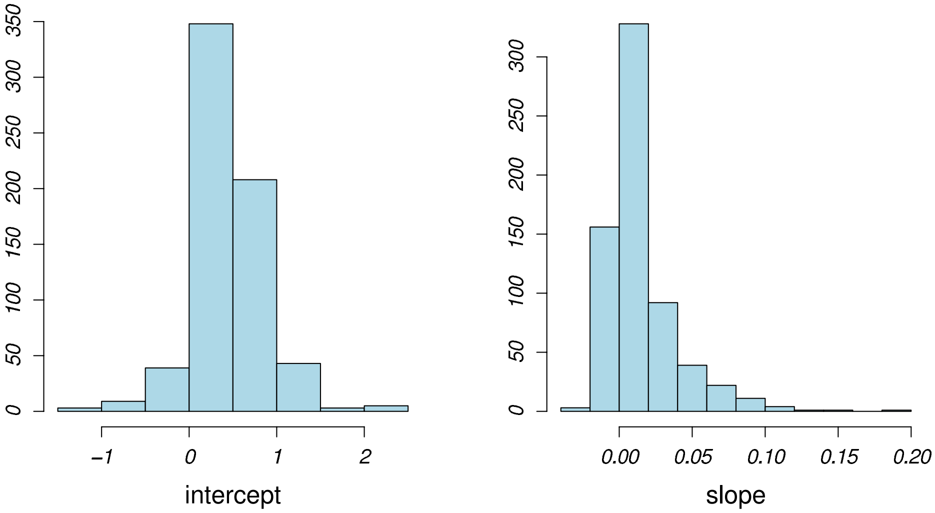


Figure 2. Histograms of Subject-Specific Intercept and Slope Estimates of lPSA Longitudinal Profile from PJL Method.

dictors include Gleason score, tumor stage, and age at radiation therapy. Hormone therapy has been shown to be able to effectively decrease the risk of cancer recurrence. Both baseline PSA level and dose of radiation are shown not to be important predictors. Figure 2 shows the histogram of the predicted subject-specific intercept and slope of LNPSA longitudinal profile using the PJL method. Although the distribution of the intercepts is symmetric and close to a bell shape, the distribution of the slopes is skewed to the right. From our simulation study we know that the PJL method is quite robust to this non-normal feature of the data.

6. DISCUSSION

In this paper, we propose a penalized likelihood approach to jointly model longitudinal and survival data. Simulation results show that under a variety of scenarios, the parameter estimators from the PJL procedure behave as well as the EM estimators. Under all of our simulated scenarios, the PJL method also takes less computing time than the EM algorithm. In addition, it avoids estimating the baseline hazard function, thus it simplifies the computation. Under non-normal scenario, both the PJL and the EM estimators are fairly robust.

The PJL method slightly underestimates the variance, especially for the variance component parameters and α. This is because the variance estimation of the variance components and α does not take account of the variability in the estimation of $\mathbf{b}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\mu}$. Similar phenomenon was reported by Ripatti and Palmgren (2000) in their penalized partial likelihood approach for multivariate frailty models. Consequently, the 95% coverage rate of the PJL estimators can be less than the nominal level.

For the EM method, calculating the asymptotic variance-covariance matrix for the EM estimator is very time consuming. Since calculating the variance estimates in the PJL method only requires evaluating the observed low dimensional information matrix at convergence, a by-product of the Newton-Ralphson algorithm, the calculation is much simpler and takes much less time than the EM method. For the risk coefficient α, a quick way (Tsiatis and Davidian, 2004) to estimate the standard error of the risk coefficient α for the EM method is to use a profile information matrix, which is derived by taking the partial derivative of an implicit profile score equation with respect to α, by treating the conditional expectation in the E step as if it does not involve α. Hsieh, Tseng and Wang (2006) pointed out that this working SE is smaller than the real SE. In their simulation study, the ratios of the two SEs range from 61% to 95% for settings with normal random effects and range from 50% to 90% for settings with non-normal random effects, respectively. Although our simulation settings are different than that of Hsieh, Tseng and Wang (2006), it is still reasonable to conclude that the variance estimate for α in the PJL method is comparable to the variance estimate using profile information matrix in the EM method.

APPENDIX A. DEFINITIONS AND NOTATIONS

For any matrix A, the symbol $\otimes$ is define as

$$A^{\otimes 2} = A^T A$$

Define

$$\begin{aligned}
S0 &= \sum_{j\in R(Y_i)} exp\{\boldsymbol{\alpha}\mathbf{L_j}(Y_i)\mathbf{b} + W_{2j}^T\boldsymbol{\gamma}\} \\
Sb1 &= \sum_{j\in R(Y_i)} exp\{\boldsymbol{\alpha}\mathbf{L_j}(Y_i)\mathbf{b} + W_{2j}^T\boldsymbol{\gamma}\}\boldsymbol{\alpha}\mathbf{L_j}(Y_i) \\
Sb2 &= \sum_{j\in R(Y_i)} exp\{\boldsymbol{\alpha}\mathbf{L_j}(Y_i)\mathbf{b} + W_{2j}^T\boldsymbol{\gamma}\}(\boldsymbol{\alpha}\mathbf{L_j}(Y_i))^{\otimes 2} \\
Sw1 &= \sum_{j\in R(Y_i)} exp\{\boldsymbol{\alpha}\mathbf{L_j}(Y_i)\mathbf{b} + W_{2j}^T\boldsymbol{\gamma}\}W_{2j}^T \\
Sw2 &= \sum_{j\in R(Y_i)} exp\{\boldsymbol{\alpha}\mathbf{L_j}(Y_i)\mathbf{b} + W_{2j}^T\boldsymbol{\gamma}\}W_{2j}^{\otimes 2} \\
S\alpha 1 &= \sum_{j\in R(Y_i)} exp\{\boldsymbol{\alpha}\mathbf{L_j}(Y_i)\hat{\mathbf{b}} + W_{2j}^T\hat{\boldsymbol{\gamma}}\}\mathbf{L_j}(Y_i)\hat{\mathbf{b}} \\
S\alpha 1 &= \sum_{j\in R(Y_i)} exp\{\boldsymbol{\alpha}\mathbf{L_j}(Y_i)\hat{\mathbf{b}} + W_{2j}^T\hat{\boldsymbol{\gamma}}\}(\mathbf{L_j}(Y_i)\hat{\mathbf{b}})^{\otimes 2}
\end{aligned}$$

APPENDIX B. INFORMATION MATRICES USED IN SECTION 3.1

The observed information for $\mathbf{b}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$ are calculated as follows. For $\mathbf{b}$,

$$\begin{aligned}
-\frac{\partial^2 l}{\partial\mathbf{b}\partial\mathbf{b}^T} &= \sum_{i=1}^{m}\Delta_i\left(\frac{Sb2}{S0} - \frac{Sb1^{\otimes 2}}{S0^2}\right) \\
&\quad + \frac{1}{\sigma_\varepsilon^2}\sum_{j=1}^{n_i}\mathbf{Z}_i^T(t_{ij})\mathbf{Z}_i(t_{ij}) + V^{-1}(\phi)
\end{aligned} \tag{B.1}$$

for $\boldsymbol{\gamma}$,

$$-\frac{\partial^2 l}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T} = \sum_{i=1}^{m}\Delta_i\left(\frac{Sw2}{S0} - \frac{Sw1^{\otimes 2}}{S0^2}\right) \tag{B.2}$$

and for $\boldsymbol{\beta}$,

$$-\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = \mathbf{W}_1 V^{-1}(\phi)\mathbf{W}_1^T \tag{B.3}$$

APPENDIX C. INFORMATION MATRICES IN SECTION 3.2

The observed information matrix for $\boldsymbol{\alpha}$ is

(C.4)

$$\begin{aligned}
&-\frac{\partial^2 l}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T} \\
&= \frac{1}{2}tr\Bigg(- K''_{PJL}(\hat{\mathbf{b}})^{-1}\frac{\partial K''_{PJL}(\hat{\mathbf{b}})}{\partial\boldsymbol{\alpha}} \\
&\quad \times K''_{PJL}(\hat{\mathbf{b}})^{-1}\frac{\partial K''_{PJL}(\hat{\mathbf{b}})}{\partial\boldsymbol{\alpha}} + K''_{PJL}(\hat{\mathbf{b}})^{-1}\frac{\partial^2 K''_{PJL}(\hat{\mathbf{b}})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}\Bigg) \\
&\quad + \sum_{i=1}^{m}\Delta_i\left(\frac{S\alpha 2}{S1} - \frac{S\alpha 1^{\otimes 2}}{S0^2}\right)
\end{aligned}$$

The general form of the elements in the observed information matrix for σ_1^2, σ_2^2, ρ is

(C.5)

$$\begin{aligned} I_{i,j} &= \frac{1}{2} tr \left(\frac{\partial V^{-1}}{\partial \phi_j} \frac{\partial V}{\partial \phi_i} + V^{-1} \frac{\partial^2 V}{\partial \phi_i \partial \phi_j} \right) \\ &\quad + \frac{1}{2} tr \Bigg(- K''_{PJL}(\hat{\mathbf{b}}, \hat{\boldsymbol{\mu}})^{-1} \frac{\partial V^{-1}}{\partial \phi_j} K''_{PJL}(\hat{\mathbf{b}}, \hat{\boldsymbol{\mu}})^{-1} \frac{\partial V^{-1}}{\partial \phi_i} \\ &\quad + K''_{PJL}(\hat{\mathbf{b}}, \hat{\boldsymbol{\mu}})^{-1} \frac{\partial^2 V^{-1}}{\partial \phi_i \partial \phi_j} \Bigg) + (\hat{\mathbf{b}} - \hat{\boldsymbol{\mu}})^T \frac{\partial V^{-1}}{\partial \phi_i \partial \phi_j} (\hat{\mathbf{b}} - \hat{\boldsymbol{\mu}}) \end{aligned}$$

$\frac{\partial V^{-1}}{\partial \phi_i}$, $\frac{\partial V}{\partial \phi_i}$, $\frac{\partial^2 V^{-1}}{\partial \phi_i \partial \phi_j}$, $\frac{\partial^2 V}{\partial \phi_i \partial \phi_j}$ are block diagonal matrices with blocks equal to $\frac{\partial D^{-1}}{\partial \phi_i}$, $\frac{\partial D}{\partial \phi_i}$, $\frac{\partial^2 D^{-1}}{\partial \phi_i \partial \phi_j}$, $\frac{\partial^2 D}{\partial \phi_i \partial \phi_j}$, respectively.

APPENDIX D. INFORMATION MATRICES USED IN SECTION 3.4

In the $(m\times 2+2+p_{10}+p_{11}+p_2)\times(m\times 2+2+p_{10}+p_{11}+p_2)$ matrix for estimating standard errors for $\mathbf{b}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, the $(m \times 2) \times (m \times 2)$ upper left matrix $\mathcal{I}_{11}$ is (B.1), and the $(p_{10} + p_{11} + 2) \times (p_{10} + p_{11} + 2)$ center matrix $\mathcal{I}_{22}$ is (B.3). The lower right $p_2 \times p_2$ matrix is (B.2). All are evaluated at the parameter estimates at convergence. For the rest of the components, $\mathcal{I}_{21} = \mathcal{I}_{12}^T = W_1 V^{-1}$, $\mathcal{I}_{31} = \mathcal{I}_{13}$ is given by (D.7), and $\mathcal{I}_{32} = \mathcal{I}_{23}^T$ is a $p_2 \times (p_{10} + p_{11} + 2)$ matrix with all its elements equal to 0.

For σ_ε^2, the observed information is

(D.6)

$$\begin{aligned} & - \frac{\partial^2 l}{\partial \sigma_\varepsilon^2 \partial \sigma_\varepsilon^2} \\ &= - \sum_{i=1}^m \frac{n_i}{2(\sigma_\varepsilon^2)^2} + \frac{1}{(\sigma_\varepsilon^2)^3} tr \left(K''_{PJL}(\hat{\mathbf{b}})^{-1} \sum_{j=1}^{n_i} \mathbf{Z}_i^T(t_{ij}) \mathbf{Z}_i(t_{ij}) \right) \\ &\quad - \frac{1}{2(\sigma_\varepsilon^2)^4} tr \Bigg(K''_{PJL}(\hat{\mathbf{b}})^{-1} \sum_{j=1}^{n_i} \mathbf{Z}_i^T(t_{ij}) \mathbf{Z}_i(t_{ij}) K''_{PJL}(\hat{\mathbf{b}})^{-1} \\ &\quad \times \sum_{j=1}^{n_i} \mathbf{Z}_i^T(t_{ij}) \mathbf{Z}_i(t_{ij}) \Bigg) + \sum_{i=i}^m \sum_{j=1}^{n_i} \frac{(X_{ij} - \mathbf{Z_i}(t_{ij})\hat{\mathbf{b}})^2}{(\sigma_\varepsilon^2)^3} \end{aligned}$$

The partial deferential equation for $\boldsymbol{\gamma}$ and $\mathbf{b}$ is

(D.7)

$$\begin{aligned} & - \frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \mathbf{b}^T} \\ &= \sum_{i=1}^m \Delta_i \frac{\sum_{j \in R(Y_i)} exp\{\boldsymbol{\alpha} \mathbf{L_j}(Y_i)\mathbf{b} + W_{2j}^T \boldsymbol{\gamma}\} W_{2j} \mathbf{L_j}(Y_i) \boldsymbol{\alpha}}{S0} \\ &\quad - \sum_{i=1}^m \Delta_i \frac{(Sw1)^T Sb1}{S0^2} \end{aligned}$$

The components of the information matrix $\mathcal{J}$ for calculating the standard errors for the variance components $\sigma_1^2, \sigma_2^2, \rho$, and σ_ε^2, and $\boldsymbol{\alpha}$ are described as follows. The upper left 3×3 block of the $\mathcal{J}$ matrix is calculated with (C.5), $\mathcal{J}_{4,4}$ with (D.6) and $\mathcal{J}_{5:6,5:6}$ with (C.4). The rest of the elements in the $\mathcal{J}$ matrix are:

$$\begin{aligned} \mathcal{J}_{4,1} &= \mathcal{J}_{1,4} \\ &= \frac{1}{2\sigma_\varepsilon^2} tr \left(K''_{PJL}(\hat{\mathbf{b}})^{-1} \sum_{j=1}^{n_i} \mathbf{Z}_i^T(t_{ij}) \mathbf{Z}_i(t_{ij}) K''_{PJL}(\hat{\mathbf{b}})^{-1} \frac{D^{-1}}{\partial(\sigma_1^2)} \right) \\ \mathcal{J}_{4,2} &= \mathcal{J}_{2,4} \\ &= \frac{1}{2\sigma_\varepsilon^2} tr \left(K''_{PJL}(\hat{\mathbf{b}})^{-1} \sum_{j=1}^{n_i} \mathbf{Z}_i^T(t_{ij}) \mathbf{Z}_i(t_{ij}) K''_{PJL}(\hat{\mathbf{b}})^{-1} \frac{D^{-1}}{\partial(\sigma_2^2)} \right) \\ \mathcal{J}_{4,3} &= \mathcal{J}_{3,4} \\ &= \frac{1}{2\sigma_\varepsilon^2} tr \left(K''_{PJL}(\hat{\mathbf{b}})^{-1} \sum_{j=1}^{n_i} \mathbf{Z}_i^T(t_{ij}) \mathbf{Z}_i(t_{ij}) K''_{PJL}(\hat{\mathbf{b}})^{-1} \frac{D^{-1}}{\partial(\rho)} \right) \\ \mathcal{J}_{5:6,1} &= \mathcal{J}_{1,5:6} \\ &= \frac{1}{2\sigma_\varepsilon^2} tr \left(K''_{PJL}(\hat{\mathbf{b}})^{-1} \frac{\partial D^{-1}}{\partial(\sigma_1^2)} K''_{PJL}(\hat{\mathbf{b}})^{-1} \frac{\partial K''_{PJL}(\hat{\mathbf{b}})}{\partial \boldsymbol{\alpha}} \right) \\ \mathcal{J}_{5:6,2} &= \mathcal{J}_{2,5:6} \\ &= \frac{1}{2\sigma_\varepsilon^2} tr \left(K''_{PJL}(\hat{\mathbf{b}})^{-1} \frac{\partial D^{-1}}{\partial(\sigma_2^2)} K''_{PJL}(\hat{\mathbf{b}})^{-1} \frac{\partial K''_{PJL}(\hat{\mathbf{b}})}{\partial \boldsymbol{\alpha}} \right) \\ \mathcal{J}_{5:6,3} &= \mathcal{J}_{3,5:6} \\ &= \frac{1}{2\sigma_\varepsilon^2} tr \left(K''_{PJL}(\hat{\mathbf{b}})^{-1} \frac{\partial D^{-1}}{\partial(\rho)} K''_{PJL}(\hat{\mathbf{b}})^{-1} \frac{\partial K''_{PJL}(\hat{\mathbf{b}})}{\partial \boldsymbol{\alpha}} \right) \\ \mathcal{J}_{5:6,4} &= \mathcal{J}_{4,5:6} \\ &= - \frac{1}{2\sigma_\varepsilon^2} tr \Bigg(K''_{PJL}(\hat{\mathbf{b}})^{-1} \sum_{j=1}^{n_i} \mathbf{Z}_i^T(t_{ij}) \mathbf{Z}_i(t_{ij}) \\ &\quad \times K''_{PJL}(\hat{\mathbf{b}})^{-1} \frac{\partial K''_{PJL}(\hat{\mathbf{b}})}{\partial \boldsymbol{\alpha}} \Bigg) \end{aligned}$$

Received 15 December 2007

REFERENCES

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89–99.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Brown, E. R. and Ibrahim, J. G. (2003a). Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. *Biometrics* **59**, 686–693. MR2004274

Brown, E. R. and Ibrahim, J. G. (2003b). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* **59**, 221–228. MR1987388

Dafni, U. G. and Tsiatis, A. A. (1998). Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics* **54**, 1445–1462.

Davidian, M. and Gallant, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika* **80**, 475–488. MR1248015

Faucett, C. L. and Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Statistics in Medicine* **15**, 1663–1685.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review* **55**, 245–259. MR0963142

Hsieh, F., Tseng, Y.-K. and Wang, J.-L. (2006). Joint modeling of survival and longitudinal data: Likelihood approach revisited. *Biometrics* **62**, 1037–1043. MR2297674

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* **90**, 1112–1121. MR1354029

Pauler, D. K. and Finkelstein, D. M. (2002). Predicting time to prostate cancer recurrence based on joint models for non-linear longitudinal biomarkers and event time outcomes. *Statistics in Medicine* **21**, 3897–3911.

Raboud, J., Reid, N., Coates, R. A. and Farewell, V. T. (1993). Estimating risks of progressing to AIDS when covariates are measured with error. *Journal of the Royal Statistical Society, Series A: Statistics in Society* **156**, 393–406.

Ratcliffe, S. J., Guo, W. and Ten Have, T. R. (2004). Joint modeling of longitudinal and survival data via a common frailty. *Biometrics* **60**, 892–899. MR2133541

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* **56**, 1016–1022. MR1806744

Taylor, J. M., Yu, M. and Sandler, H. M. (2005). Individualized predictions of disease progression following radiation therapy for prostate cancer. *Journal of Clinical Oncology* **23**, 816–825.

Tseng, Y.-K., Hsieh, F. and Wang, J.-L. (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika* **92**, 587–603. MR2202648

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica* **14**, 809–834. MR2087974

Wang, Y. and Taylor, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association* **96**, 895–905. MR1946362

Wu, L. (2002). A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to AIDS Studies. *Journal of the American Statistical Association* **97**, 955–964. MR1951254

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330–339. MR1450186

Xu, J. and Zeger, S. L. (2001a). The evaluation of multiple surrogate endpoints. *Biometrics* **57**, 81–87. MR1833292

Xu, J. and Zeger, S. L. (2001b). Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society, Series C: Applied Statistics* **50**, 375–387. MR1856332

Yu, M., Law, N. J., Taylor, J. M. G. and Sandler, H. M. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica* **14**, 835–862. MR2087975

Wen Ye
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48109, USA
E-mail address: `wye@umich.edu`

Xihong Lin
Department of Biostatistics
Harvard School of Public Health
Boston, MA 02115, USA
E-mail address: `xlin@hsph.harvard.edu`

Jeremy M. G. Taylor
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48109, USA
E-mail address: `jmgt@umich.edu`