Statistics and Its Interface Volume 7 (2014) 241–250

Joint modeling of survival data and mismeasured longitudinal data using the proportional odds model

Juan Xiong*, Wenqing He, and Grace Y. Yi

CONTENTS

Joint modeling of longitudinal and survival data has been studied extensively, where the Cox proportional hazards model has frequently been used to incorporate the relationship between survival times and covariates. Although the proportional odds model is an attractive alternative to the Cox proportional hazards model by featuring the dependence of survival times on covariates via cumulative covariate effects, this model is rarely discussed in the joint modeling context. To fill up this gap, we investigate joint modeling of survival and longitudinal data where the proportional odds model is employed to feature survival data, and longitudinal covariates are postulated using measurement error models. An estimation method based on the expectation maximization algorithm is developed. In addition, the impact of naive analyses, which fail to address errors occurring in longitudinal measurements, is assessed. The performance of the proposed method is evaluated through simulation studies, and a real example is invoked for illustration.

AMS 2000 subject classifications: 62N01.
Keywords and phrases: Joint modeling, Proportional odds model, Measurement error, Survival analysis, Time-varying covariates.

*Corresponding author.

1. INTRODUCTION

In biomedical studies and clinical trials, it is often the case that data consist of times to an event of interest and covariates that are repeatedly measured over a time period. A typical example is that in HIV/AIDS clinical trials, the time to death is of interest, while the biomarker CD4 lymphocyte counts are measured at regularly scheduled intervals over a certain follow-up period. We are interested in not only identifying the risk factors to death, and the trend of the CD4 counts, but also the relationship between time to death and longitudinal CD4 counts measurements. Regression models are usually employed to describe the relationship between response and covariate variables. For time to event data, popularly used regression models include the Cox proportional hazards (PH) model [7], and the accelerated failure time (AFT) model [8]. When longitudinal covariates are time-varying, it is typical to require that time dependent covariates are observed at all failure times in order to carry out valid inferences. This requirement is rarely possible in practice, most longitudinal covariates are only measured at intermediate scheduled times. In addition, longitudinal covariates may be subject to substantial measurement error. Naive methods which ignore these issues often yield biased estimates.

Two stage approaches [9, 27] and joint modeling methods [11, 18] have been developed to overcome the difficulties discussed above. With the two stage approaches, missing co-

variate values are first imputed through a longitudinal prediction model, and in the second stage, a survival model is fitted to the imputed data. This two stage method can only partially correct for bias as opposed to the naive method which ignores missing observations. Further difficulties arise from the informative dropout which incurred in longitudinal covariates due to the occurrence of the event time. The joint modeling method, on the other hand, uses information from both longitudinal and survival processes to improve the efficiency of parameter estimates and to correct for the bias. There has been extensive research on addressing joint modeling of survival data and longitudinal covariate measurements. See [11, 18, 22, 23, 26, 29, 32], among many others.

A large number of inference methods have been developed for joint modeling of survival data with longitudinal covariates, in which the Cox PH model [7, 31] is frequently employed to modulate the survival process, and longitudinal covariates are treated as time varying covariates. This treatment basically requires time varying covariates to be observed at all time points at which failures occur. However, in reality those observations of longitudinal covariates are often not available at failure times. To get around this problem, it is a common practice that covariate measurements at the latest time points are used as approximated versions. This approach can produce approximate analysis results, but it is not ideal. In this paper, we use a different approach to handle this issue. We particularly take into account the differences of the true covariates and observed measurements by embedding the problem into the framework with covariates measurement error models.

The Cox proportional hazards model has been popularly used in survival data analysis by both researchers and analysts. The main reason is that the baseline hazard function in the proportional hazards model does not need to be specified, and the partial likelihood inferential procedure makes its implementation easy. Moreover, elegantly theoretical justifications are available to support its use. The availability of various statistical software packages makes inferential procedures based on the Cox proportional hazards models ready to be used. Although the Cox proportional hazards model enjoys wide applications, this model can be restrictive for real applications. A key condition in using such a model is the proportionality assumption for the hazard functions. This is a strong assumption that often fails in practice. To overtime this inflexibility, many alternative models have been proposed to handle data with distinct features. Among them, the proportional odds (PO) model is a useful tool, and this model has been widely used in areas such as epidemiological and biomedical studies [33].

The proportional odds model has often been suggested as an attractive alternative to the Cox PH model when the ratio of the two hazard functions is not constant over time, but is changing along the time [10, 24]. For instance, when the treatment effects diminish over time, the ratio of the hazard functions of different treatment groups is typically not a constant, and using the proportional hazards model to feature such a scenario is apparently not feasible. The proportional odds model can be, however, a useful alternative to characterizing this case.

Regression parameters in the proportional odds model have a clear interpretation in terms of log odds ratios. Inference procedures for the proportional odds model have been studied by many authors [1, 6, 20, 21, 33]. Despite the usefulness of such a model, there is no investigation on joint modeling of survival data that are posited by the proportional odds model, along with longitudinal covariates subject to measurement error. Aiming to fill up this gap to provide an alternative to the Cox model, we develop inferential procedures to jointly analyze survival and longitudinal data, with survival data being postulated directly to the proportional odds model, and the longitudinal covariates being subject to measurement error. Furthermore, we assess the impact of naive analyses that fail to address error incurring in longitudinal measurements on the estimation of the model parameters of interest.

Our work is partially motivated by the Mediterranean fruit fly egg-laying data [4]. The data set consists of the lifetime and complete records of the numbers of eggs produced daily until death from 1,000 female Mediterranean fruit flies. We are interested in investigating the relationship between reproduction and longevity.

The remainder of the paper is organized as follows. Section 2 presents the notation and model formulation. Section 3 describes the joint modeling estimation procedures. In Section 4, the performance of the proposed method is evaluated by simulation studies. The proposed method is applied to the Mediterranean fruit fly egg-laying data in Section 5 for illustration. Discussion and concluding remarks are provided in the last section.

2. NOTATION AND MODEL FORMULATION

Let T_i and C_i be the survival time and censoring time for the ith subject, respectively, $i = 1, \ldots, n$. The observed time is denoted by $V_i = \min(T_i, C_i)$, and δ_i represents the censoring indicator, taking 1 if $T_i \le C_i$ and 0 otherwise. Let $\mathbf{Z}_i$ be the vector of time independent covariates and $\mathbf{X}_i(t)$ be the vector of time-varying covariates. Assume that T_i and C_i are independent conditional on the covariates.

In the absence of time-varying covariates $\mathbf{X}_i(t)$, the proportional odds model for the survival time associated with time independent covariate $\mathbf{Z}_i$ is defined as

$$\frac{1 - S(t|\mathbf{Z}_i)}{S(t|\mathbf{Z}_i)} = \exp\left(\boldsymbol{\beta}_z^T \mathbf{Z}_i\right) \frac{1 - S_0(t)}{S_0(t)}, \tag{1}$$

where $S(t|\mathbf{Z}_i)$ is the survival function given covariate $\mathbf{Z}_i$, $S_0(t)$ denotes the baseline survival function, and $\boldsymbol{\beta}_z$ is the parameter vector. Equivalently, (1) can be written as a form by differentiating both sides of (1) with respect to time t.

That is, let $r(t;\mathbf{Z}_i) = (\mathrm{d}/\mathrm{d}t)\big[\{1-S(t|\mathbf{Z}_i)\}/S(t|\mathbf{Z}_i)\big]$ and $r_0(t) = (\mathrm{d}/\mathrm{d}t)[\{1-S_0(t)\}/S_0(t)]$, then (1) is equivalent to

$$r(t;\mathbf{Z}_i) = \exp\left(\boldsymbol{\beta}_z^T\mathbf{Z}_i\right) r_0(t). \tag{2}$$

To accommodate time varying covariates $\mathbf{X}_i(t)$ into the PO model, it is more convenient to use (2) then (1) [25]. Specifically, we set the model

$$r(t;\mathbf{X}_i(t),\mathbf{Z}_i) = \exp\{\boldsymbol{\beta}_x^T\mathbf{X}_i(t) + \boldsymbol{\beta}_z^T\mathbf{Z}_i\} r_0(t), \tag{3}$$

and this model yields the survival function linking T_i and covariates $\{\mathbf{X}_i(t),\mathbf{Z}_i\}$:

$$S(t|\mathbf{X}_i(t),\mathbf{Z}_i) = \frac{1}{1+\int_0^t r(u;\mathbf{X}_i(u),\mathbf{Z}_i)du}, \tag{4}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_x^T,\boldsymbol{\beta}_z^T)^T$ is the vector of regression parameters. Under model (3), the cumulative distribution function of T_i is given by

$$F(t) = \frac{\int_0^t \exp\{\boldsymbol{\beta}_x^T\mathbf{X}_i(s) + \boldsymbol{\beta}_z^T\mathbf{Z}_i\}h_0(s)/S_0(s)\mathrm{d}s}{1+\int_0^t \exp\{\boldsymbol{\beta}_x^T\mathbf{X}_i(s) + \boldsymbol{\beta}_z^T\mathbf{Z}_i\}h_0(s)/S_0(s)\mathrm{d}s}, \tag{5}$$

where $h_0(t)$ is the baseline hazard function. The corresponding density function under model (3) is then

$$f(t) = \frac{\exp\{\boldsymbol{\beta}_x^T\mathbf{X}_i(t) + \boldsymbol{\beta}_z^T\mathbf{Z}_i\}h_0(t)}{S_0(t)\left[1+\int_0^t \exp\{\boldsymbol{\beta}_x^T\mathbf{X}_i(s) + \boldsymbol{\beta}_z^T\mathbf{Z}_i\}\frac{h_0(s)}{S_0(s)}\mathrm{d}s\right]^2}. \tag{6}$$

As noted in [17], under the joint modeling setup, leaving the baseline survival function completely unspecified may result in inexplicit maximum likelihood estimates of the associated parameters. A viable strategy is to use a parametric or weakly parametric model to postulate the baseline survival function. While step functions or regression splines can be used to model baseline functions, here we employ a parametric form to posit the baseline survival function, and use $\boldsymbol{\alpha}$ to denote the associated parameter vector.

To complete modeling, we modulate the longitudinal covariate vector $\mathbf{X}_i(t)$ by

$$\mathbf{X}_i(t) = \mathbf{U}_i^T\boldsymbol{\rho}_i(t), \tag{7}$$

where $\mathbf{U}_i = (u_{i1},\ldots,u_{ir})^T$ is a r-dimensional vector of random effects, and $\boldsymbol{\rho}_i(t) = (\rho_{i1}(t),\ldots,\rho_{ir}(t))^T$ is a vector of functions of t. Different specification of $\boldsymbol{\rho}_i(t)$ allows flexibility in modeling various time trajectories. Other modeling schemes can be used to model longitudinal covariates as well. For instance, [32] introduced the simple linear model; [16] and [30] considered Gaussian processes; [2] and [22] discussed representing the time trajectory via the polynomial model, and [3] and [12] proposed the B-splines approach. Here we assume the random effects $\mathbf{U}_i$ to follow a multivariate normal distribution $\mathrm{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}_u)$ as often discussed in the literature [15, 26, 32], and let $\boldsymbol{\zeta}$ denote the vector of associated parameters.

In practice, longitudinal processes can not be fully observed, but are measured intermittently at certain time points and they are often subject to measurement error. Let $t_{i1} < \cdots < t_{im_i} \le V_i$ denote the time points at which longitudinal covariates are assessed, and write $\mathbf{t}_i = (t_{i1},\ldots,t_{im_i},V_i)^T$. Let $\mathbf{X}_i(\mathbf{t}_i) = \{\mathbf{X}_i(t_{ij}), j=1,\ldots,m_i\}$ be the covariates $\mathbf{X}_i(t)$ at all the longitudinal assessment points, and $\mathbf{X}_{ij}^*$ be the observed version of covariate $\mathbf{X}_i(t_{ij}), j=1,\ldots,m_i$. $\mathbf{X}_{ij}^*$ and $\mathbf{X}_i(t_{ij})$ are assumed to follow a classical additive measurement error model[5].

$$\mathbf{X}_{ij}^* = \mathbf{X}_i(t_{ij}) + \mathbf{e}_{ij}, \tag{8}$$

where the measurement error $\mathbf{e}_{ij}$, independent of other variables, is assumed to follow a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}_e = [\sigma_{kl}]$ with σ_{kk} denoted as σ_e^2. We use $\boldsymbol{\Delta}$ to denote the vector of associated parameters.

3. ESTIMATION PROCEDURES

Let $\boldsymbol{\eta} = (\boldsymbol{\beta}^T,\boldsymbol{\alpha}^T,\boldsymbol{\gamma}^T)^T$ be the vector of the unknown parameters, where $\boldsymbol{\beta}$ is the vector of regression parameters, $\boldsymbol{\alpha}$ is the vector of parameters associated with the baseline survival function $S_0(t)$, and $\boldsymbol{\gamma} = (\boldsymbol{\zeta}^T,\boldsymbol{\Delta}^T)^T$ denotes the parameters characterizing the random effect model and the measurement error process.

We estimate $\boldsymbol{\eta}$ by maximizing the joint observed likelihood which can be constructed by integrating out random effects with the assumption of nondifferential measurement error. The observed likelihood function contributed by subject i by integrating out random effects is obtained:

$$\begin{aligned} L_i(\boldsymbol{\eta}) \;=\; & \int_u \Bigg[f(V_i,\delta_i|\mathbf{X}_i(\mathbf{t}_i),\mathbf{Z}_i,\mathbf{U}_i)\cdot \\ & \left\{\prod_{j=1}^{m_i} f(\mathbf{X}_{ij}^*|\mathbf{U}_i)\right\} f(\mathbf{U}_i)\Bigg]\mathrm{d}\mathbf{U}_i, \end{aligned}$$

where $f(\mathbf{X}_{ij}^*|\mathbf{U}_i)$ is determined by (7) and (8), and with the undifferential measurement error mechanism, the survival density function is derived from (5) and (6)

$$\begin{aligned} & f(V_i,\delta_i|\mathbf{X}_i(\mathbf{t}_i),\mathbf{Z}_i,\mathbf{U}_i) \\ = \;& f(V_i|\mathbf{X}_i(\mathbf{t}_i),\mathbf{Z}_i,\mathbf{U}_i)^{\delta_i}(1-F(V_i|\mathbf{X}_i(\mathbf{t}_i),\mathbf{Z}_i,\mathbf{U}_i))^{1-\delta_i} \\ = \;& \left[\frac{\exp\{\boldsymbol{\beta}_x^T\mathbf{X}_i(V_i) + \boldsymbol{\beta}_z^T\mathbf{Z}_i\}h_0(V_i)}{S_0(V_i)}\right]^{\delta_i}\cdot \\ & \left[\frac{1}{1+\int_0^{V_i}\exp\{\boldsymbol{\beta}_x^T\mathbf{X}_i(s) + \boldsymbol{\beta}_z^T\mathbf{Z}_i\}\frac{h_0(s)}{S_0(s)}\mathrm{d}s}\right]^{1+\delta_i}. \end{aligned} \tag{9}$$

Let $L(\boldsymbol{\eta}) = \prod_{i=1}^n L_i(\boldsymbol{\eta})$. Then maximizing $L(\boldsymbol{\eta})$ gives the maximum likelihood estimator of $\boldsymbol{\eta}$, denoted by $\widehat{\boldsymbol{\eta}}$.

Under regularity conditions, $\sqrt{n}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})$ has an asymptotic normal distribution with mean $\mathbf{0}$ and covariance matrix $[E\{(\partial \ell_i(\boldsymbol{\eta})/\partial \boldsymbol{\eta}^T)(\partial \ell_i(\boldsymbol{\eta})/\partial \boldsymbol{\eta})\}]^{-1}$, where $\ell_i(\boldsymbol{\eta}) = \log L_i(\boldsymbol{\eta})$.

The observed likelihood method is conceptually straightforward, but its implementation can be computational difficult due to complex integrals involved. Alternatively, an expectation-maximization method can be developed for inferences. Below we describe this method in detail, proceeded with two naive methods.

3.1 Naive method

In naive analyses, measurement error is often ignored and $\mathbf{X}_i(t)$ is treated as constant. Here we discuss two naive methods that are used in practice.

3.1.1 Naive method 1

A naive method is to directly replace the true covariate $\mathbf{X}_i(t)$ with the last observed surrogate $\mathbf{X}^*_{im_i}$, and treat it as constant over the integral, therefore, leading to the approximation

$$\begin{aligned} & \int_0^t \exp\{\boldsymbol{\beta}_x^T \mathbf{X}_i(s) + \boldsymbol{\beta}_z^T \mathbf{Z}_i\} \frac{h_0(s)}{S_0(s)} \mathrm{d}s \\ \approx \quad & \exp\{\boldsymbol{\beta}_x^T \mathbf{X}^*_i(t_{im_i}) + \boldsymbol{\beta}_z^T \mathbf{Z}_i\} \int_0^t \frac{h_0(s)}{S_0(s)} \mathrm{d}s. \end{aligned}$$

3.1.2 Naive method 2

A less naive method is to approximate $\mathbf{X}_i(t)$ with constants over small intervals formed by the time points $\{t_{i1}, t_{i2}, \ldots, t_{im_i}\}$. Specifically, we use the following approximation to simplify the likelihood function $L_i(\boldsymbol{\eta})$:

$$\begin{aligned} & \int_0^t \exp\{\boldsymbol{\beta}_x^T \mathbf{X}_i(s) + \boldsymbol{\beta}_z^T \mathbf{Z}_i\} \frac{h_0(s)}{S_0(s)} \mathrm{d}s \\ \approx \quad & \exp\{\boldsymbol{\beta}_x^T \mathbf{X}^*_{i1} + \boldsymbol{\beta}_z^T \mathbf{Z}_i\} \int_0^{t_{i2}} \frac{h_0(s)}{S_0(s)} \mathrm{d}s \\ + \quad & \sum_{j=3}^{m_i} \exp\{\boldsymbol{\beta}_x^T \mathbf{X}^*_{i(j-1)} + \boldsymbol{\beta}_z^T \mathbf{Z}_i\} \int_{t_{i(j-1)}}^{t_{ij}} \frac{h_0(s)}{S_0(s)} \mathrm{d}s \\ + \quad & \exp\{\boldsymbol{\beta}_x^T \mathbf{X}^*_{im_i} + \boldsymbol{\beta}_z^T \mathbf{Z}_i\} \int_{t_{im_i}}^{V_i} \frac{h_0(s)}{S_0(s)} \mathrm{d}s. \end{aligned}$$

3.2 EM algorithm

To fully incorporate the measurement error in the longitudinal covariate measurements, we now develop an expectation-maximization (EM) algorithm to obtain an estimate of $\boldsymbol{\eta}$. With models (6), (7), and 8, the complete data likelihood for subject i is given by

$$\begin{aligned} & L_i(\boldsymbol{\eta}) \\ = \quad & f(V_i, \delta_i | \mathbf{X}_i(\mathbf{t}_i), \mathbf{Z}_i, \mathbf{U}_i) \Big\{ \prod_{j=1}^{m_i} f(\mathbf{X}^*_{ij} | \mathbf{X}_i(t_{ij}), \mathbf{Z}_i, \mathbf{U}_i) \Big\} \cdot \\ & f(\mathbf{X}_i(\mathbf{t}_i), \mathbf{U}_i | \mathbf{Z}_i) \\ = \quad & f(V_i, \delta_i | \mathbf{X}_i(\mathbf{t}_i), \mathbf{Z}_i, \mathbf{U}_i) \Big\{ \prod_{j=1}^{m_i} f(\mathbf{X}^*_{ij} | \mathbf{X}_i(t_{ij}), \mathbf{U}_i) \Big\} \cdot \\ & f(\mathbf{X}_i(\mathbf{t_i}), \mathbf{U}_i), \end{aligned}$$

where nondifferential measurement error is assumed.

Since $\mathbf{X}_i(t)$ is a function of $\mathbf{U}_i$, $\mathbf{X}_i(t) = \mathbf{U}_i^T \boldsymbol{\rho}_i(t)$, an alternative way to view this likelihood is to replace $\mathbf{X}_i(t)$ with model (7), leading to

$$L^*_i(\boldsymbol{\eta}) = f(V_i, \delta_i | \mathbf{Z}_i, \mathbf{U}_i) \Big\{ \prod_{j=1}^{m_i} f^*(\mathbf{X}^*_{ij} | \mathbf{U}_i) \Big\} f(\mathbf{U}_i) \tag{10}$$

for some functions $f^*(\mathbf{X}^*_{ij}|\mathbf{U}_i)$. In the formulation (10), random effect $\mathbf{U}_i$ is the only unobserved variable.

Let $L^*(\boldsymbol{\eta}) = \prod_{i=1}^n L^*_i(\boldsymbol{\eta})$ and $\ell^*(\boldsymbol{\eta}) = \log L^*(\boldsymbol{\eta})$. At the E-step, we compute the expected log-likelihood of the complete data, $E\{\ell^*(\boldsymbol{\eta})\}$, conditional on the observed data and the estimate $\boldsymbol{\eta}^{(k)}$ of the parameters from the previous iteration k, where the conditional density function is given by

$$\begin{aligned} & f(\mathbf{U}_i | V_i, \delta_i, \mathbf{X}^*_{ij}, \mathbf{Z}_i; \boldsymbol{\eta}^{(k)}) \\ = \quad & \frac{f(V_i, \delta_i, \mathbf{U}_i | \mathbf{X}^*_{ij}, \mathbf{Z}_i; \boldsymbol{\eta}^{(k)})}{f(V_i, \delta_i | \mathbf{X}^*_{ij}, \mathbf{Z}_i; \boldsymbol{\eta}^{(k)})} \\ = \quad & \frac{f(V_i, \delta_i | \mathbf{U}_i, \mathbf{X}^*_{ij}, \mathbf{Z}_i; \boldsymbol{\eta}^{(k)}) f(\mathbf{U}_i | \mathbf{X}^*_{ij}, \mathbf{Z}_i, ; \boldsymbol{\eta}^{(k)})}{\int_u f(V_i, \delta_i | \mathbf{U}_i, \mathbf{X}^*_{ij}, \mathbf{Z}_i; \boldsymbol{\eta}^{(k)}) f(\mathbf{U}_i | \mathbf{X}^*_{ij}, \mathbf{Z}_i; \boldsymbol{\eta}^{(k)}) \mathrm{d}\mathbf{U}_i}. \end{aligned}$$

To evaluate the conditional expectation $E\{\ell^*(\boldsymbol{\eta})|\boldsymbol{\eta}^{(k)}\}$, we use the Gauss-Hermite quadrature approximation [19]. Specifically if a function $g(x)$ has the form $g(x) = h(x)\phi(x; a, b)$, where $\phi(x; a, b)$ is a Gaussian density function with mean a and standard deviation b, then

$$\int_{-\infty}^{\infty} g(x) \mathrm{d}x \approx \sum_{i=1}^{M} \sqrt{2} b w_i \exp(c_i^2) g(\sqrt{2} b c_i + a),$$

where M is the number of sample points used for the approximation, c_i are the roots of the Hermite polynomial, and w_i are associated weights.

In the M-step, we maximize the resulting approximation of $E\{\ell^*(\boldsymbol{\eta})|\boldsymbol{\eta}^{(k)}\}$ and obtain a new updated estimate $\boldsymbol{\eta}^{(k+1)}$ of the parameter $\boldsymbol{\eta}$. Iterate between the E-step and M-step until the parameter estimate converges. The associated standard errors are calculated by the bootstrap method.

In implementing the EM iterations, convergence is declared if either of the following two conditions is satisfied:

$$E\{\ell^*(\boldsymbol{\eta}^{(k+1)})\} - E\{\ell^*(\boldsymbol{\eta}^{(k)})\} < \mathrm{tol}_1 \Big\{ |E\{\ell^*(\boldsymbol{\eta}^{(k)})\}| + \mathrm{tol}_1 \Big\}$$

or

$$\max\left\{|\boldsymbol{\eta}^{(k+1)}-\boldsymbol{\eta}^{(k)})|/(\boldsymbol{\eta}^{(k)}+\text{tol}_2)\right\}<\text{tol}_3,$$

where $\text{tol}_1 = 1.490116\cdot 10^{-8}$ is the square root of the machine precision, and we set $\text{tol}_2 = 10^{-3}$ and $\text{tol}_3 = 10^{-4}$.

4. SIMULATION STUDIES

We carry out extensive simulation studies to assess the performance of the proposed method. We focus on the situation of a single time-varying covariate $X_i(t)$ and a single time independent covariate Z_i. The sample size is $n = 200$. Five hundred simulations are run at each parameter setting. The covariate Z_i is generated from the binomial distribution BIN$(1, 0.5)$. The longitudinal component $X_i(t)$ is generated from the linear growth curve model $X_i(t) = U_{i0} + U_{i1}t$.

Random effects $\mathbf{U}_i = (U_{i0}, U_{i1})^T$ follow a bivariate normal distribution with mean $(\mu_1, \mu_2)^T = (4.173, -0.103)^T$ and covariance matrix $\boldsymbol{\Sigma}_u$, $(\sigma_{11}, \sigma_{12}, \sigma_{22})^T = (1.24, -0.0114, 0.0003)^T$ to mimic an HIV clinical trial as described at [9, 28]. Longitudinal covariate $X_i(t)$ are measured at $t_{ij} = 0, 2, 4, 8, 16, 24, 32, 40, 56, 64, 72, 80$ weeks with constant probability 0.1 of missing observation at any time point except the baseline. The surrogate version X^*_{ij} is generated from the classic measurement error model $X^*_{ij} = X_i(t_{ij}) + e_{ij}$ with $e_{ij} \sim N(0, \sigma^2_e)$ and $\sigma^2_e = 0.6$.

Survival times are generated based on the proportional odds model (1). That is, for the ith subject, $i = 1, \ldots, n$, we first generate a random variate $v_i \sim U[0, 1]$, and calculate the survival time T_i by solving equation

$$\frac{v_i}{1-v_i} = \int_0^{T_i} \exp\{\boldsymbol{\beta}_x^T X_i(s) + \boldsymbol{\beta}_z^T Z_i\} h_0(s)/S_0(s)\mathrm{d}s,$$

where $S_0(t)$ is the baseline survival function. We particularly consider two parametric modeling for $S_0(t)$: (1) an exponential distribution $S_0(t) = \exp(-t/\alpha_1)$ with $\alpha_1 = 30$, and (2) a log-logistic distribution $S_0(t) = 1/1 + (\alpha_1 t)^{\alpha_2}$ with $\alpha_1 = 0.05$ and $\alpha_2 = 2$. We set $\beta_x = -\log(2)$ and $\beta_z = 0.5$.

Censoring times are generated independently from the exponential distribution with three different means: 1) a fixed mean of 110 weeks, 2) a fixed mean resulting in roughly 20% censoring rate and 3) a fixed mean resulting in about 40% censoring rate.

We compare the performance of the three methods: (1) the naive method 1 with the true covariate $X_i(t)$ directly replaced by the last observed surrogate $X^*_{im_i}$; (2) the naive method 2 with the integration involved in the survival function approximated by treating $X_i(t)$ as constant over the intervals formed by the grid $\{t_{i1}, t_{i2}, \ldots, t_{im_i}\}$; and (3) the EM method. Here we report on the results of the biases of the estimates (Bias), the empirical standard errors (SEE), the bootstrap standard errors (SEB), and the coverage rates (CR) for 95% confidence intervals of the parameters. Tables 1 and 2 report the results for the case that the baseline function follows an exponential distribution. In

Table 1. Simulation Results on the Regression Coefficients and Baseline Function Parameters Obtained from the Naive Methods and Proposed Method. The True Baseline Distribution Follows an Exponential Distribution

Censoring		Naive Method 1				Naive Method 2				Proposed Method			
Rate	Parameter	Bias	SEM	SEE	CR	Bias	SEM	SEE	CR	Bias	SEB	SEE	CR
20%	β_x	0.821	0.063	0.045	0.000	0.177	0.055	0.053	0.120	−0.011	0.075	0.075	0.952
	β_z	−1.134	0.236	0.159	0.000	−0.290	0.249	0.234	0.806	0.011	0.273	0.275	0.938
	$\log(\alpha_1)$	0.169	0.113	0.082	0.734	−0.099	0.085	0.073	0.806	0.002	0.092	0.085	0.966
28%	β_x	0.778	0.064	0.045	0.000	0.173	0.058	0.054	0.150	−0.012	0.078	0.078	0.968
	β_z	−1.052	0.243	0.164	0.000	−0.279	0.258	0.242	0.840	0.016	0.284	0.288	0.940
	$\log(\alpha_1)$	0.216	0.122	0.086	0.586	−0.093	0.090	0.077	0.846	0.003	0.098	0.089	0.976
40%	β_x	0.719	0.068	0.050	0.000	0.165	0.063	0.058	0.250	−0.015	0.085	0.084	0.956
	β_z	−0.943	0.256	0.187	0.006	−0.268	0.274	0.260	0.854	0.014	0.303	0.310	0.934
	$\log(\alpha_1)$	0.282	0.138	0.101	0.438	−0.084	0.100	0.086	0.894	0.002	0.107	0.098	0.974

Table 2. Simulation Results on the Random Effect Parameters and Variance of Measurement Error Obtained from the Proposed Method. The True Baseline Distribution Follows an Exponential Distribution

Censoring Rate	Parameter	Proposed Method			
		Bias	SEB	SEE	CR
20%	μ_1	0.005	0.085	0.084	0.954
	μ_2	0.000	0.003	0.003	0.940
	σ_{11}	0.002	0.148	0.149	0.946
	σ_{12}	0.000	0.003	0.003	0.920
	σ_{22}	0.000	0.000	0.000	0.924
	σ_e^2	0.002	0.029	0.027	0.952
28%	μ_1	0.006	0.086	0.084	0.954
	μ_2	0.000	0.003	0.003	0.934
	σ_{11}	0.003	0.148	0.149	0.946
	σ_{12}	−0.001	0.003	0.003	0.926
	σ_{22}	0.000	0.000	0.000	0.926
	σ_e^2	0.001	0.029	0.029	0.946
40%	μ_1	0.006	0.086	0.084	0.954
	μ_2	0.000	0.003	0.003	0.940
	σ_{11}	0.001	0.149	0.151	0.942
	σ_{12}	−0.001	0.004	0.004	0.934
	σ_{22}	0.000	0.000	0.000	0.924
	σ_e^2	0.000	0.030	0.029	0.946

Table 1, we report the results for the regression coefficients and the baseline function parameters, while Table 2 reports the results for the estimates of parameters associated with the distribution function of the random effect and measurement error. Similarly, Tables 3 and 4 display the results for the case that the baseline function follows a log-logistic distribution.

The estimates from the naive method 1 have the largest bias and the worst coverage of the 95% confidence interval compared to the other two methods. As the censoring rate gets higher, the biases of the estimates of the regression coefficients decrease while the associated standard errors tend to become bigger. The corresponding coverage rates for the 95% confidence intervals are far off the nominal level. For the parameters of the baseline distribution function, as the censoring rate gets higher, the biases of the estimates and the standard errors increase and consequently, the corresponding coverage rates for the 95% confidence intervals decrease.

Naive method 2, as expected, leads to more accurate approximation than naive method 1. The estimates from naive method 2 have smaller biases and standard errors than those from naive method 1. As the censoring rate gets higher, the biases of the point estimates decrease while the associated standard errors increase, thus the coverage rates for the 95% confidence intervals increase.

In implementing the EM integration algorithm, we use the bootstrap method with 100 runs for each configuration to calculate standard errors. This method remarkably

Table 3. Simulation Results on the Regression Coefficients and Baseline Function Parameters Obtained from the Naive Methods and Proposed Method. The True Baseline Distribution Follows a Log-logistic Distribution

Censoring Rate	Parameter	Naive Method 1				Naive Method 2				Proposed Method			
		Bias	SEM	SEE	CR	Bias	SEM	SEE	CR	Bias	SEB	SEE	CR
20%	β_x	1.435	0.091	0.088	0.000	0.396	0.080	0.084	0.008	−0.015	0.137	0.138	0.942
	β_z	−0.193	0.263	0.258	0.909	−0.013	0.265	0.268	0.944	0.020	0.280	0.281	0.940
	$\log(\alpha_1)$	−0.709	0.047	0.040	0.000	−0.322	0.094	0.096	0.120	0.027	0.227	0.209	0.934
	$\log(\alpha_2)$	0.775	0.066	0.078	0.000	0.346	0.090	0.103	0.062	−0.001	0.149	0.151	0.918
26%	β_x	1.399	0.094	0.091	0.000	0.392	0.083	0.086	0.010	−0.018	0.143	0.139	0.948
	β_z	−0.182	0.270	0.265	0.917	−0.013	0.272	0.278	0.942	0.021	0.289	0.291	0.942
	$\log(\alpha_1)$	−0.711	0.050	0.042	0.000	−0.318	0.098	0.099	0.125	0.030	0.233	0.212	0.946
	$\log(\alpha_2)$	0.756	0.068	0.080	0.000	0.338	0.093	0.106	0.077	−0.002	0.153	0.152	0.934
40%	β_x	1.318	0.099	0.097	0.000	0.384	0.091	0.092	0.028	−0.021	0.157	0.154	0.954
	β_z	−0.164	0.287	0.289	0.909	−0.016	0.291	0.303	0.936	0.018	0.311	0.318	0.930
	$\log(\alpha_1)$	−0.712	0.057	0.047	0.000	−0.310	0.107	0.108	0.191	0.034	0.250	0.231	0.942
	$\log(\alpha_2)$	0.712	0.075	0.089	0.000	0.322	0.100	0.110	0.129	−0.001	0.159	0.161	0.934

Table 4. Simulation Results on the Random Effect Parameters and Variance of Measurement Error Obtained from the Proposed Method. The True Baseline Distribution Follows a Log-logistic Distribution

Censoring Rate	Parameter	Proposed Method			
		Bias	SEB	SEE	CR
20%	μ_1	0.006	0.085	0.084	0.952
	μ_2	0.000	0.003	0.003	0.938
	σ_{11}	0.001	0.149	0.149	0.934
	σ_{12}	0.000	0.003	0.003	0.942
	σ_{22}	0.000	0.000	0.000	0.938
	σ_e^2	0.002	0.029	0.028	0.940
26%	μ_1	0.005	0.085	0.084	0.952
	μ_2	0.000	0.003	0.003	0.938
	σ_{11}	0.002	0.150	0.148	0.942
	σ_{12}	0.000	0.004	0.003	0.944
	σ_{22}	0.000	0.000	0.000	0.942
	σ_e^2	0.002	0.029	0.029	0.940
40%	μ_1	0.007	0.085	0.084	0.954
	μ_2	0.000	0.003	0.003	0.928
	σ_{11}	0.002	0.150	0.149	0.950
	σ_{12}	−0.001	0.004	0.004	0.942
	σ_{22}	0.000	0.000	0.000	0.936
	σ_e^2	0.001	0.031	0.030	0.950

Table 5. Simulation Results for Evaluating the Effect on Misspecifying the Baseline Survival Function

σ_e^2	Bias							
	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2	β_x	β_z
0.25	−0.049	0.000	0.013	−0.005	0.000	0.023	−0.888	−0.330
0.60	−0.045	−0.001	0.013	−0.008	0.000	0.023	−1.019	−0.252

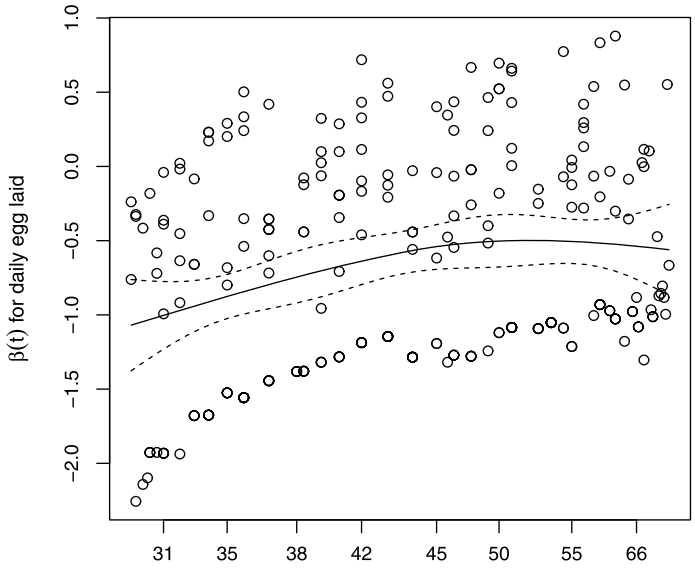

Figure 1. Scaled Schoenfeld Residuals Plot of the Cox PH Model.

outperforms the two naive methods. It produces reasonably small finite sample biases for the estimates of both the regression coefficients and the baseline parameters. The standard errors obtained from the EM approach are larger than those obtained from the naive methods, but the corresponding coverage rates are reasonably close to the nominal value. The biases of the estimates of regression coefficients and the associated standard errors increase as the censoring rate gets higher. Meanwhile the biases of the estimates of the baseline parameters are quite stable regardless of the change of censoring rate. In addition to estimating $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, we also assess the performance of the estimates of the nuisance parameters $\boldsymbol{\gamma} = (\boldsymbol{\mu}^T, \boldsymbol{\Sigma}_u, \sigma_e^2)$. As shown, in Tables 2 and 4, the estimates have reasonably small finite sample biases and the coverage rates of the 95% confidence intervals are close to the nominal value. The point estimates and standard errors of $\boldsymbol{\gamma}$ from the proposed method do not appear to be obviously affected by varying the censoring rate.

We comment that the proposed joint modeling method relies on the parametric modeling of the baseline survival function. A natural concern is on the effects of misspecifying baseline survival functions. Our simulation study demonstrates that biased estimates of covariate effects can be yielded, if the baseline survival function is misspecified, as seen in Table 5 where the true Gumbel distribution is misspecified as Exponential distribution. To increase modeling flexibility, one scheme is to approximate the baseline survival function through weakly parametric approaches, such as regression splines [14]. The weakly parametric methods can approximate the baseline survival function reasonably well for many applications, and this modeling scheme allows a finite number of parameters thus retaining the validity of the proposed method.

In summary, the naive analyses would often lead to biased results due to the inherent nature of ignoring the measurement error. Naive method 2 partially improves the estimation compare to the naive method 1, nevertheless, the improvement is limited and not entirely satisfactory. The EM method behaves the best and outperforms the two naive methods. This method is advantageous in that it adjusts for the measurement error effects on estimation of the covariate effects and the parameters for the baseline distribution function simultaneously; it produces reasonable estimates for the model parameters.

5. APPLICATION TO MEDITERRANEAN FRUIT FLY FECUNDITY DATA

As an illustration, we apply the proposed method to the egg-laying data originated from [4]. The data set consists of the lifetime and complete records of the numbers of eggs produced daily until death from 1,000 female Mediterranean fruit flies. We are interested in investigating the relationship

Table 6. Analysis of Mediterranean Fruit Fly Data: Estimation from the Exponential Baseline Function

	Exponential Distribution							
	β_x	$\log(\alpha_1)$	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2
Estimate	−0.238	3.675	2.142	−0.132	0.144	−0.019	0.003	0.253
Bootstrap SD	0.017	0.237	0.026	0.004	0.016	0.002	0.000	0.013

Table 7. Analysis of Mediterranean Fruit Fly Data: Estimation from the Log-logistic Baseline Function

	Log-logistic Distribution								
	β_x	$\log(\alpha_1)$	$\log(\alpha_2)$	μ_1	μ_2	σ_{11}	σ_{12}	σ_{22}	σ_e^2
Estimate	−0.393	−3.607	1.454	2.126	−0.130	0.105	−0.014	0.002	0.261
Bootstrap SD	0.091	0.173	0.501	0.042	0.005	0.023	0.003	0.000	0.020

between reproduction and longevity. Such an investigation is much needed, given that reproduction is a fundamental life history trait of uttermost importance in the study of evolutionary biology.

A key to the proposed procedure is a suitable parametric longitudinal model. The following parametric longitudinal process is suggested by [4, 26] to describe the individual fly's fecundity profile

$$X_i(t_{ij}) = U_{i0}\log(t_{ij}) + U_{i1}(t_{ij} - 1),$$

where random effect $U_i = (U_{i0}, U_{i1})^T$ is assumed to follow a bivariate normal distribution with mean $(\mu_1, \mu_2)^T$ and a covariance matrix with parameters $(\sigma_{11}, \sigma_{12}, \sigma_{22})$.

It is acknowledged that the daily egg production X_{ij}^* is subject to random daily fluctuations. The classic measurement error model (8) provides a good way to link the underlying reproductive process to the actual observed daily egg-laying. The measurement error is assumed to follow a normal distribution $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$. To overcome the problem that in some days there are eggs laid, it is common to take the logarithmic transformation to the daily egg-laying plus one

$$\log(X_{ij}^* + 1) = X_i(t_{ij}) + e_{ij}.$$

Following the criteria of other authors who analyzed this data set [26], we include only the flies that produced more than 1,150 eggs in their lifetime. The effective sample size is 251 flies, with lifetimes ranging from 22 to 99 days. The first two days contain zero counts for all flies and therefore were left out from the analysis. To test our procedure in the presence of irregular sampling plans on the longitudinal data, we randomly set the measurement at any of the days to be missing with probability 0.1 for all flies except the baseline at day three.

The Cox PH regression model assumption is evaluated by the plot of scaled Schoenfeld residuals versus time [12]. The PH model was rejected at p-value = 0.0004. [26] explored this data by the joint modeling approach with the AFT assumption for the survival component. [13] studied this data by the Bayesian approach for joint models. Specifically, they

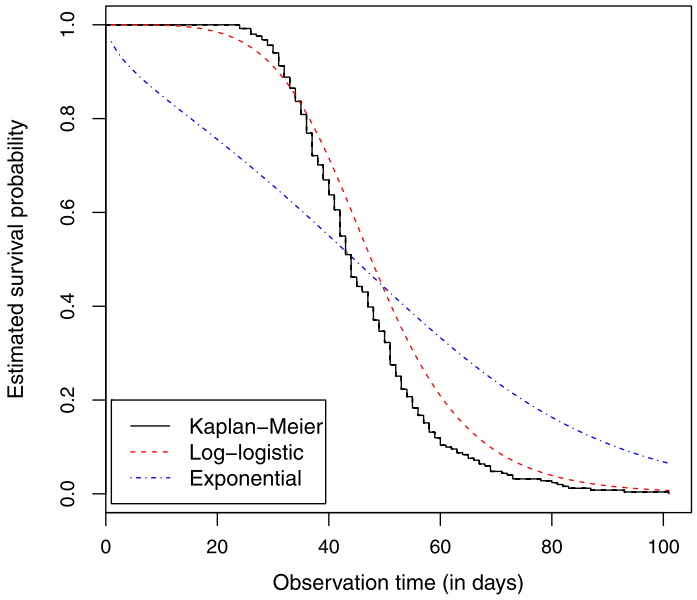

Figure 2. Mediterranean Fruit Fly Data Survival Curves.

compared the performance of Cox PH, AFT and PO survival models and suggested that the PO model predictively outperforms the other survival specifications. Here we fit the data to the proportional odds model in conjunction with longitudinal data, with baseline survival function following either exponential distribution and log-logistic distribution.

The estimated model parameters with baseline function postulated by an exponential model or a log-logistic model are reported in Tables 6 and 7. The negative regression coefficient suggests that the odds of survival increases with increasing egg counts, i.e, reproduction activity is positively associated with longevity.

We compared the estimated survival functions from the exponential baseline function and log-logistic baseline function with the Kaplan-Meier estimate of the survival function. Figure 2 shows that the fitted survival curve for the log-logistic is much closer to the nonparametric Kaplan-Meier curve than the exponential baseline model, which

suggests that log-logistic baseline survival function better describe the baseline survival function. Other weakly parametric models, such as spline based approach, may be worth investigating for this data set.

6. DISCUSSION

In this paper, we explore the joint modeling of longitudinal and survival data under the proportional odds model. Longitudinal covariates subject to measurement error are assumed to follow a linear mixed effects model. The procedure is based on maximizing the joint likelihood of both longitudinal and survival processes. The EM algorithm is used to estimate the parameters. Our simulation studies and real data analysis show that the performance of the proposed method is reasonably satisfactory. Our method can not only correct for the measurement error effects on estimation of the covariate and the parameters for the baseline distribution function, but also produce accurate estimates for the parameters of random effect.

As seen in the simulation studies, the proposed joint modeling method relies on the parametric modeling of the baseline survival function, and the estimates may be biased if the baseline survival function is misspecified. Weakly parametric approaches, such as step functions or regression splines [14], can be invoked to approximate baseline survival function to both increasing modeling flexibility and retain the validity of the proposed method.

When estimating the standard error of the parameters, we encounter the difficulty that the exact information matrix of parameters of interest cannot be obtained directly in the EM algorithm. In such a case, the use of the bootstrap estimates of standard deviations is desirable. This technique works well in the simulation studies and data illustration. Nevertheless, it is computationally intensive. Another difficulty that needs to be further addressed in the future study is the requirement for numerical integration with respect to the random effects. The integration is conventionally performed using Gaussian quadrature rules or the Monte Carlo algorithm. However due to its increasing computational complexity with the dimension of the random-effects vector, this algorithm is time consuming.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge two anonymous reviewers for their helpful comments which improved the presentation of this manuscript. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada.

Received 1 March 2013

REFERENCES

[1] Bennett, S. (1983). Log-logistic regression models for survival data. *Applied Statistics* **32** 165–171.

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

[3] Brown, E. R., Ibrahim, J. G. and DeGruttola, V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics* **61** 64–73. MR2129202

[4] Carey, J. R., Liedo, P., Müller, H. G., Wang, J. L. and Chiou, J. M. (1998). Relationship of age patterns of fecundity to mortality, longevity, and lifetime reproduction in a large cohort of Mediterranean fruit fly females. *Journal of Gerontology – Biological Sciences* **53** 245–251.

[5] Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Mesurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. Chapman and Hall, London. MR2243417

[6] Cheng, SC., Wei, LJ. and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82** 835–845. MR1380818

[7] Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34** 187–220. MR0341758

[8] Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London. MR0751780

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

[10] Dauxois, J. Y. and Kirmani, S. N. U. A. (2003). Testing the proportional odds model under random censoring. *Biometrika* **90** 913–922. MR2024766

[11] Ding, J. and Wang, J. L. (2008). Modelling longitudinal data with nonparametric multiplicative random effects jointly with surival data. *Biometrics* **64** 546–556. MR2432425

[12] Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81** 515–526. MR1311094

[13] Hanson, T. E., Branscum, A. J. and Johnson, W. O. (2011). Predictive comparison of joint longitudinal-survival modeling: A case study illustrating competing approaches. *Lifetime Data Analysis* **17** 3–28. MR2764577

[14] He, W. Q. and Lawless, J. F. (2003). Flexible maximum likelihood methods for bivariate proportional hazards models. *Biometrics* **59** 837–848. MR2025107

[15] He, W. Q., Xiong, J. and Yi, G. Y. (2011). Analysis of error-contaminated survival data under the proportional odds model. *Journal of Statistical Research* **45** 111–130. MR2934365

[16] Henderson, R., Diggle, P. and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1** 465–480.

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

[18] Li, L., Hu, B. and Greene, T. (2009). A semiparametric joint model for longitudinal and survival data with application to hemodialysis study. *Biometrics* **65** 737–745. MR2649846

[19] Liu, Q. and Pierce, D. A. (1994). A note on gauss-hermite quadrature. *Biometrika* **81** 624–629. MR1311107

[20] Murphy, S. A., Rossini, A. J. and Van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association* **92** 968–976. MR1482127

[21] Pettitt, A. N. (1984). Proportional odds model for survival data and estimates using ranks. *Applied Statistics* **33** 169–175.

[22] Song, X., Davidian, M. and Tsiatis, A. A. (2002). An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. *Biostatistics* **3** 511–528.

[23] Song, X. and Wang, C. Y. (2008). Semiparametric approaches for joint modeling of longitudinal and survival data with time-varying coefficients. *Biometrics* **64** 557–566. MR2432426

[24] Sun, J., Sun, R. L. and Zhu, C. (2007). Testing the proportional odds model for interval-censored data. *Lifetime Data Analysis* **13** 37–50. MR2355295
[25] Sundaram, S. (2006). Semiparametric inference in proportional odds model with time-dependent covariates. *Journal of Statistical Planning and Inference* **136** 320–334. MR2211343
[26] Tseng, Y. K., Hsieh, F. and Wang, J. L. (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika* **92** 587–603. MR2202648
[27] Tsiatis, A. A., DeGruttola, V. and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* **90** 27–37.
[28] Tsiatis, A. A. and Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* **88** 447–458. MR1844844
[29] Tsiatis, A. A. and Davidian, M. (2004). Joint modelling of longitudinal and time-to-event data: An overview. *Statistica Sinica* **14** 809–834. MR2087974
[30] 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** 895–9 MR1946362
[31] Wu, L., Liu, W., Yi, G. Y. and Huang, Y. X. (2012). Analysis of longitudinal and survival data: Joint modeling, inference methods, and issues. *Journal of Probability and Statistics* Article ID **640153**. MR2862470
[32] Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53** 330–339. MR1450186
[33] Yang, S. and Prentice, R. L. (1999). Semiparametric inference in the proportional odds model. *Journal of the American Statistical Association* **94** 125–136. MR1689218

Juan Xiong
Department of Statistical and Actuarial Sciences
University of Western Ontario
1151 Richmond Street North, London
Ontario, N6A 5B7
Canada
E-mail address: xiongjuan2000@gmail.com

Wenqing He
Department of Statistical and Actuarial Sciences
University of Western Ontario
1151 Richmond Street North, London
Ontario, N6A 5B7
Canada
E-mail address: whe@stats.uwo.ca

Grace Y. Yi
Department of Statistics and Actuarial Science
University of Waterloo
200 University Avenue West, Waterloo
Ontario, N2L 3G1
Canada
E-mail address: yyi@uwaterloo.ca