Biometrics 61, 64–73
March 2005

A Flexible B-Spline Model for Multiple Longitudinal Biomarkers and Survival

Elizabeth R. Brown,[1,*] Joseph G. Ibrahim,[2] and Victor DeGruttola[3]

[1]Department of Biostatistics, University of Washington, Campus Mail Stop 357232, Seattle, Washington 98195-7232, U.S.A.
[2]Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A.
[3]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, U.S.A.
[*]*email:* elizab@u.washington.edu

Summary. Often when jointly modeling longitudinal and survival data, we are interested in a multivariate longitudinal measure that may not fit well by linear models. To overcome this problem, we propose a joint longitudinal and survival model that has a nonparametric model for the longitudinal markers. We use cubic B-splines to specify the longitudinal model and a proportional hazards model to link the longitudinal measures to the hazard. To fit the model, we use a Markov chain Monte Carlo algorithm. We select the number of knots for the cubic B-spline model using the Conditional Predictive Ordinate (CPO) and the Deviance Information Criterion (DIC). The method and model selection approach are validated in a simulation. We apply this method to examine the link between viral load, CD4 count, and time to event in data from an AIDS clinical trial. The cubic B-spline model provides a good fit to the longitudinal data that could not be obtained with simple parametric models.

Key words: Bayesian methods; Cubic B-splines; Joint longitudinal and survival model; Model assessment.

1. Introduction

Methods for jointly modeling longitudinal and survival data have recently gained popularity in the statistical literature. A common interest in modeling is to make inferences about the effect of the longitudinal measures on the time to event, but not to make inferences about the longitudinal measures or their projected change over time. The longitudinal model is important for accounting for measurement error and defining the individual's longitudinal trajectory between times of measurement. Because a complex model for the longitudinal measures may be too complicated to estimate in the joint model, a linear mixed model that is parametric with respect to time is most commonly used. More recently, attention has focused on relaxing assumptions on the model for the longitudinal data, but this research has focused mainly on the distributional assumptions of the random effects and has not addressed the shape of the trajectory, often making simple parametric assumptions for complicated patterns of data. If we are only interested in modeling the longitudinal component, we may often judge these simple models to be inappropriate. Also, when the true relationship between time and the longitudinal biomarker is nonlinear, the effect that assuming a simple linear longitudinal model has on the estimate of the relationship of the marker and the time-to-event outcome is unclear.

Several models have been proposed for jointly modeling longitudinal and survival data, most assuming a mixed effects model for the longitudinal component with the random effects or subject-specific parameters arising from normal distributions (DeGruttola and Tu, 1994; Faucett and Thomas, 1996; Wulfsohn and Tsiatis, 1997; Xu and Zeger, 2001; Ibrahim, Chen, and Sinha, 2004). An overview is given by Ibrahim, Chen, and Sinha (2001, Chapter 7). Wang and Taylor (2001) also used a mixed effects model but proposed an integrated Ornstein–Uhlenbeck (IOU) process for longitudinal CD4 count data in a joint model. Using random intercepts and a fixed slope, they used the IOU process to allow the path of an individual's biomarker, also known as the trajectory, to fluctuate around a straight line. Although this is an improvement in the fit of the longitudinal markers, we may want more flexibility than this model provides. Several papers have also recently been published specifying nonparametric distributions for the random effects, both from the frequentist (Tsiatis and Davidian, 2001; Song, Davidian, and Tsiatis, 2002a,b) and Bayesian (Brown and Ibrahim, 2003) perspectives. Lin et al. (2002) developed a latent class model that allowed the polynomial trajectory to depend on class membership. While these approaches do allow for more flexible modeling of the longitudinal data, they still impose parametric assumptions on the path of an individual's longitudinal marker.

We motivate the need for a more flexible model with data from a substudy of plasma HIV RNA concentrations (viral load) from the AIDS clinical trials group (ACTG) 175 study (Hammer et al., 1996; Katzenstein et al., 1996). This substudy evaluated the viral loads of 391 randomly selected patients. In

this article, we explore how levels of viral load and CD4 counts over time affect the time to AIDS or death, whichever comes first. We limited this analysis to a subset of patients who had at least one HIV RNA level above the lower limit of detection and had CD4 counts measured concurrently with viral load at one or more of those time points. Figures 2 and 3 show viral load and CD4 counts for 25 of the subjects. The shapes of the curves for the two longitudinal markers are clearly not consistent between patients over time. However, there does appear to be some correlation in the shapes of the two curves over time within the patients. In this article, we take a different approach to producing a flexible longitudinal model than those taken by Wang and Taylor (2001), Tsiatis and Davidian (2001), Song et al. (2002a,b), or Brown and Ibrahim (2003). We do not make rigid assumptions about the path of the biomarkers over time. Instead, we develop a smooth, flexible, data-driven model for multiple longitudinal outcomes using cubic B-splines. This longitudinal model will then be incorporated in the joint longitudinal and survival model. Properties of this model, including the effect of the model specification on the estimate of the effect of the longitudinal marker on the time-to-event outcome, will be explored through simulation and a data analysis.

2. The Joint Longitudinal and Survival Model

In this section, we present a joint survival and longitudinal model in which the longitudinal markers are modeled using a cubic B-spline model for multiple longitudinal outcomes. The resulting model provides a flexible and robust approach to jointly modeling longitudinal and failure-time data.

We begin with a review of the longitudinal cubic B-spline model introduced by Rice and Wu (2001) for a single longitudinal outcome and then extend this model to the Bayesian setting. We then generalize the univariate Bayesian B-spline model to accommodate a multivariate outcome and show how to incorporate this multivariate B-spline model in a joint longitudinal and survival model.

2.1 *Longitudinal Cubic B-Spline Model*

Given N subjects with observations taken at times $0 \le t_{ij} \le T$, $j = 1, \ldots, m_i$ on the ith subject, let Y_{ij} be the ith subject's set of observed biomarkers at time t_{ij}. We first focus on the case where there is only one biomarker measured over time. We denote this biomarker's value at time t_{ij} as Y_{ij1} to distinguish it from the multivariate case, which we discuss later. We define an observation at time t_{ij} to be a function of the true underlying trajectory $\psi(t_{ij})$ plus error,

$$Y_{ij1} = \psi(t_{ij}) + e_{ij}.$$

We now focus on modeling $\psi(t_{ij})$. Rice and Wu (2001) introduced a univariate spline model for unequally sampled longitudinal data, where

$$\psi(t_{ij}) = \sum_{k=1}^{\bar{q}} b_{0k}\bar{B}_k(t_{ij}) + \sum_{k=1}^{q} \beta_{ik}B_k(t_{ij}), \tag{1}$$

and $\{\bar{B}_k(\cdot)\}$ is a $\bar{q}$-dimensional basis for spline functions on $[0, T]$ with a fixed knot sequence. The quantity $\sum_{k=1}^{q} \beta_{ik}B_k(t_{ij})$ defines a random effect curve for the ith subject with q-dimensional basis $\{B_k(\cdot)\}$ on $[0, T]$, which may be a different basis than $\{\bar{B}_k(\cdot)\}$. The β_{ik} are random coefficients with mean 0 and covariance matrix V. The notation $B_k(t_{ij})$ denotes the value of the kth basis function at measurement time t_{ij}. Rice and Wu (2001) take a frequentist approach to obtain parameter estimates using the EM algorithm.

It is also possible to approach this problem from a Bayesian perspective by specifying (1) as a hierarchical model. This model may also include other covariates such as treatment. Toward this goal, let

$$\psi(t_{ij}) = \sum_{k=1}^{q} \beta_{ik}B_k(t_{ij}) + x_i'\alpha, \tag{2}$$

where $\beta_{ik} \sim N(b_{0k}, V_{0k})$, and α is a vector of parameters linking the vector of baseline covariates x_i to the longitudinal outcome. We assume that the errors are independent and normally distributed, $e_{ij} \sim N(0, \sigma^2)$. We may also specify priors for b_{0k}, V_{0k}, and σ^2. Note that unlike the model in (1), the model in (2) imposes the same basis for modeling the individual trajectories and the mean curve. There may be situations where this is undesirable. For example, we may believe that the population curve would change slowly over time and require fewer basis functions to model. On the other hand, we may want to allow individual curves more flexibility to make more rapid changes over time. In this case, we may want to specify an individual trajectory that has more basis functions than the population trajectory. For the model presented here, we assume the simpler case in which both subject and population trajectories have the same number of basis functions, q. With large V_{0k}, $k = 1, \ldots, q$, this should allow for adequate variability between and within subjects.

We next extend this model to the multivariate case. Let p be the number of longitudinal outcomes measured at a time point. The observation, trajectory, and error vectors for subject i at time point j are then

$$Y_{ij} = \begin{pmatrix} Y_{ij1} \\ \vdots \\ Y_{ijp} \end{pmatrix}, \quad \psi_{\alpha,\beta}(t_{ij}) = \begin{pmatrix} \psi_{\alpha,\beta,1}(t_{ij}) \\ \vdots \\ \psi_{\alpha,\beta,p}(t_{ij}) \end{pmatrix}$$
$$= \begin{pmatrix} \sum_{k=1}^{q} \beta_{ik1}B_k(t_{ij}) + x_i'\alpha_1 \\ \vdots \\ \sum_{k=1}^{q} \beta_{ikp}B_k(t_{ij}) + x_i'\alpha_p \end{pmatrix},$$

and

$$\epsilon_{ij} = \begin{pmatrix} \epsilon_{ij1} \\ \vdots \\ \epsilon_{ijp} \end{pmatrix}.$$

We then specify the longitudinal multivariate model as

$$\begin{aligned} Y_{ij} &= \psi_{\alpha,\beta}(t_{ij}) + \epsilon_{ij}, \\ \beta_{ik} &\sim N_p(b_{0k}, V_{0k}), \end{aligned} \tag{3}$$

where $\beta_{ik} = (\beta_{ik1}, \ldots, \beta_{ikp})'$, $b_0 = (b_{0k1}, \ldots, b_{0kp})'$, V_{0k} is a $p \times p$ covariance matrix, $\epsilon_{ij} \sim N_p(0, \Sigma)$, and $N_p(a, b)$ is the p-dimensional multivariate normal distribution with mean

vector a and covariance matrix b. We give $\psi(t)$ the subscripts α and β to denote its dependence on the parameters α and $\beta = \{\beta_{ikl},\, i = 1, \ldots, N,\, k = 1, \ldots, q,\, l = 1, \ldots, p\}$.

The model specified in (3) describes how the biomarker changes over time and estimates the impact of another covariate or set of covariates, α, on the biomarker. One drawback of this specification is that the effect of these covariates is assumed to shift the trajectory a constant amount over time. This assumption may often be too restrictive. In these circumstances, it may be more appropriate to move the treatment effect up a level in the hierarchical model. Because the model is specified using cubic B-splines, we can easily model treatment at the group or population level without sacrificing interpretability. Using this approach to modeling treatment makes intuitive sense in a hierarchical model as well. The hierarchies in our model consist of the subject-specific effects at the lowest level, followed by covariate effects at the population level such as treatment, and then our prior beliefs about the population parameters at the top level of the hierarchy. Let $\psi_\beta(t_{ij}) = \psi_{\alpha=0,\beta}(t_{ij})$ and define the multivariate longitudinal model as

$$\begin{aligned} Y_{ij} &= \psi_\beta(t_{ij}) + \epsilon_{ij}, \\ \beta_{ik} &\sim N_p(b_{0k} + x_i'\alpha, V_{0k}). \end{aligned} \tag{4}$$

With this specification, the covariates shift the means of the spline coefficients equally over time; however, because the effect of the covariates is now specified in the prior, the estimate of the subject-specific trajectory is more flexible with respect to the effect of the covariates. This is in contrast to the trajectory specified in (3), which strictly imposes a constant shift in the estimates over time.

The expected value of the trajectory for an individual measurement in (3) is $\mathrm{E}(\psi_\beta(t_{ij}) \,|\, b_0, \alpha) = \sum_{k=1}^{q} b_{0k} B_k(t_{ij}) + x_i'\alpha$. We can show that the trajectory specified in (4) has the same prior mean as (3). We have

$$\begin{aligned} &\mathrm{E}\left[\sum_{k=1}^{q} \beta_{ik} B_k(t_{ij}) \,\middle|\, b_0, \alpha\right] \\ &= \sum_{k=1}^{q} \mathrm{E}[\beta_{ik} \,|\, b_0, \alpha] B_k(t_{ij}) \\ &= \sum_{k=1}^{q} (b_{0k} + x_i'\alpha) B_k(t_{ij}) \\ &= \sum_{k=1}^{q} b_{0k} B_k(t_{ij}) + x_i'\alpha \sum_{k=1}^{q} B_k(t_{ij}) \\ &= \sum_{k=1}^{q} b_{0k} B_k(t_{ij}) + x_i'\alpha. \end{aligned} \tag{5}$$

Although we can use either model in practice, the model specified in (4) exhibits better convergence properties.

We next compare the effect of the priors on the subject-specific parameters in the trajectory function. Specifically, we see that the variance of the trajectory is dependent on the parameterization of the trajectory. The variance of the trajectory defined in (4) at a point in time can be written as

$$\begin{aligned} \mathrm{var}(\psi_{\beta,\alpha}(t_{ij}) \,|\, b_0, \alpha, V_0) &= \sum_{k=1}^{q} B_k(t_{ij})^2 V_{0k} \\ &\leq \sum_{k=1}^{q} V_{0k}. \end{aligned}$$

However, if the trajectory is defined as a polynomial function of time, $g(t_{ij}) = \sum_{k=1}^{q} \beta_{ik} t_{ij}^{k-1}$ and the β's have the same distributional assumptions as (4), then the variance of the trajectory at a point in time is expressed as

$$\mathrm{var}(g(t_{ij}) \,|\, b_0, \alpha, V_0) = \sum_{k=1}^{q} t^{2*(k-1)} V_{0k}.$$

In this case, the variance of the trajectory at a point in time is unbounded as $t \to \infty$ and is also sensitive to the units of time (e.g., days, weeks). These are properties that are undesirable and can be avoided using cubic B-splines.

Using (4) for the longitudinal model and defining

$$Y_i = \begin{pmatrix} Y_{i10} & \cdots & Y_{im_i0} \\ \vdots & & \vdots \\ Y_{i1p} & \cdots & Y_{im_ip} \end{pmatrix},$$

an individual's contribution to the likelihood of the longitudinal marker is

$$\begin{aligned} &p(Y_i \,|\, \Sigma, \beta) \\ &\propto \frac{1}{|\Sigma|^{\frac{m_i}{2}}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{m_i} (Y_{ij} - \psi_\beta(t_{ij}))' \Sigma^{-1} (Y_{ij} - \psi_\beta(t_{ij}))\right\}. \end{aligned} \tag{6}$$

We can then specify priors for the parameters in (6) and obtain estimates for the Bayesian semiparametric multivariate longitudinal model. Because this model is only a component of the joint longitudinal and survival model, we will postpone further discussion of appropriate priors until the next section where we present the joint longitudinal and survival model.

2.2 *The Joint Model*

To model the relationship between the longitudinal measures and event time, we construct the joint likelihood as the product of the time-to-event likelihood conditional on the longitudinal measures multiplied by the likelihood of the longitudinal measure in (6).

We define the hazard at time t to be a function of the trajectory at time t. The hazard function given the longitudinal measures is thus expressed as

$$h(t \,|\, Y) = \lambda(t) \exp(\gamma \psi_\beta(t) + z'\zeta), \tag{7}$$

where $\gamma = (\gamma_1, \ldots, \gamma_p)'$ is a vector of parameters linking the trajectory to the hazard function, $\lambda(t)$ is the baseline hazard, and ζ is a parameter vector linking a vector z of baseline covariates to the failure time. These covariates may be categorical, such as treatment, or continuous, such as age.

The specification of the hazard in (7) leads to the following distribution for an individual's time to event, s_i, given the trajectory function:

$$f(s_i, \nu_i \,|\, Y_i) = \lambda(s_i)^{\nu_i} \exp\{\nu_i(\gamma'\psi_\beta(s_i) + z_i'\zeta)\} \times \exp\left\{-\int_0^{s_i} \lambda(u) e^{\gamma'\psi_\beta(u)+z_i'\zeta}\, du\right\},$$

where ν_i is the censoring indicator for subject i.

If we assume the baseline hazard function is piecewise constant so that

$$\lambda(u) = \lambda_j, \quad u_{j-1} \le u < u_j, \quad j = 1, \ldots, J,$$

then the cumulative hazard,

$$\int_0^{s_i} \lambda(u) e^{\gamma'\psi_\beta(u)+z_i'\zeta}\, du,$$

can be rewritten as

$$e^{z_i'\zeta} \sum_{j=1}^{J} H_{ij}(\beta, \gamma, \lambda),$$

where

$$H_{ij}(\beta, \gamma, \lambda) = I\{s_i \ge u_{j-1}\}\lambda_j \int_{u_{j-1}}^{\min(u_j, s_i)} e^{\gamma'\psi_\beta(u)+z_i'\zeta}\, du, \quad (8)$$

and $I\{s_i \ge u_{j-1}\}$ is an indicator function that equals 1 if the event time occurs in or later than the jth interval and 0 otherwise. The integral in (8) does not have an analytical solution for the trajectory defined by cubic B-splines. Instead, for computational ease and speed, we use the trapezoidal rule to approximate the integral in (8).

With this approximation, we can now express the ith subject's contribution to the joint likelihood function as

$$\begin{aligned} f(Y_i, s_i, \nu_i) &= f(s_i, \nu_i \,|\, Y_i) \times f(Y_i), \\ f(Y_i, s_i, \nu_i) &\propto \lambda(s_i)^{\nu_i} \exp\{\nu_i(\gamma'\psi_\beta(s_i) + z_i'\zeta)\} \\ &\quad \times \exp\left\{-e^{z_i'\zeta} \sum_{j=1}^{J} H_{ij}(\beta, \gamma, \lambda)\right\} \\ &\quad \times \frac{1}{|\Sigma|^{\frac{m_i}{2}}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{m_i} (Y_{ij} - \psi_\beta(t_{ij}))' \right. \\ &\quad \left. \times \Sigma^{-1}(Y_{ij} - \psi_\beta(t_{ij}))\right\}. \end{aligned}$$

We then specify the following priors on the parameters in (9), $\gamma_l \sim N_2(g_{l0},\ g_{l1})$, $\Sigma^{-1} \sim \text{Wishart}(S^{-1},\ \nu_\epsilon)$, and $\beta_{il} = (\beta_{il1} \cdots \beta_{ilp})' \sim N_2(b_{0l} + x_i'\alpha,\ V_{0l})$, $b_{0l} = (b_{0l1} \cdots b_{0lp})'$, and V_{0l} is a $p \times p$ covariance matrix. Here, $\text{Wishart}(S^{-1},\ \nu_\epsilon)$ denotes the Wishart distribution with ν_ϵ degrees of freedom and scale matrix S^{-1}. We may also specify priors on the hyperparameters of β, $\alpha \sim N_p(C_0,\ C_1)$, $b_{0l} \sim N_p(A_0,\ A_1)$, and $V_{0l}^{-1} \sim \text{Wishart}_{\nu_{v_0}}(S_{v_0}^{-1})$. The prior distributions were chosen to be as general as possible while still being proper and conjugate to the likelihood when possible.

3. Estimation

We use the Gibbs sampler (Gelfand and Smith, 1990) to obtain samples from the posterior distribution. The full conditionals are available on the first author's website (`http://faculty.washington.edu/elizab`). Many of the parameters in the model (γ, β, ζ) do not have conjugate priors. However, after selecting the prior distributions shown in the previous section, the full conditional distributions are log concave. Therefore, we can use adaptive rejection sampling (ARS; Gilks and Wild, 1992) to obtain samples from these full conditionals. Defining the model for the trajectory as (4) results in a closed-form full conditional for α; whereas, the definition in (3) does not. If we had specified a parametric time-dependent trajectory, we could not have obtained this result. The sampling scheme was implemented in `C` and is available from the authors as an `R` (Ihaka and Gentleman, 1996) library.

4. Model Comparison

We examine two statistics for model comparison, the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002) and the Conditional Predictive Ordinate (CPO; Gelfand, Dey, and Chang, 1992).

The DIC is the sum of the deviance estimated using the posterior estimates of the parameters, $D(\bar{\Theta})$, and twice the effective number of parameters, p_D. The effective number of parameters is estimated by $p_D = \overline{D(\Theta)} - D(\bar{\Theta})$, where $\overline{D(\Theta)}$ is the posterior mean of the deviance (the average of the deviances calculated using the estimated parameters at each step of the Markov chain Monte Carlo [MCMC] sampler). For the model presented in this article, the DIC can be expressed as

$$\text{DIC} = 2\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{N} \log\left\{f\left(s_i, \nu_i, Y_i \,\middle|\, \Theta^{(g)}\right)\right\} - \sum_{i=1}^{N} \log\{f(s_i, \nu_i, Y_i \,|\, \bar{\Theta})\},$$

where $\Theta^{(g)}$ denotes the parameter samples at the gth, $g = 1, \ldots, G$, iteration of the Gibbs sampler and $\bar{\Theta}$ represents the means of the posterior samples. A smaller DIC indicates a better fit when comparing models.

For the ith observation, the CPO-statistic is defined as

$$\begin{aligned} \text{CPO}_i &= f\left(s_i, \nu_i, Y_i \,|\, D^{(-i)}\right) \\ &= \int f(s_i, \nu_i, Y_i \,|\, \Theta, D_i)\pi(\Theta \,|\, D^{(-i)})\, d\Theta, \end{aligned} \quad (9)$$

where Θ denotes the model parameters, D_i denotes the ith patient's covariate data, and $D^{(-i)}$ denotes the covariate data for all the patients except the ith patient. We cannot obtain a closed form solution for (9); however, Chen et al. (2000, Chapter 10) show that a Monte Carlo approximation of (9) is

$$\widehat{\text{CPO}}_i = \left(\frac{1}{G}\sum_{g=1}^{G} \frac{1}{f\left(s_i, \nu_i, Y_i \,|\, \Theta^{(g)}\right)}\right)^{-1},$$

where $\Theta^{(g)}$ denotes the parameter samples at the gth, $g = 1, \ldots, G$, iteration of the Gibbs sampler. A large CPO value

indicates a better fit. We can compare different models using the sums of the logs of the CPOs of the individual observations. Models with greater $\sum \log(\widehat{\text{CPO}}_i)$ values will indicate a better fit.

Computing DIC and $\sum \log(\widehat{\text{CPO}}_i)$ is straightforward, requiring only the samples generated by the Gibbs sampler. In this article, we examine the performance of both measures in a simulation as well as a real data set.

5. Simulations

We conducted simulations to evaluate the performance of the model, the DIC, and the CPO. We specified a model with three evenly spaced interior knots and an intercept ($q = 7$). The parameter values were set as follows:

$$\gamma = (-0.9, -1.0)', \quad \alpha = (-0.1, -0.2)', \quad \zeta = -0.5,$$

$$b_{01} = (0.22, 0.14, 0.13, 0.06, 0.09, 0.01, -0.12)',$$

$$b_{02} = (-1.00, -0.95, -1.10, -1.20, -1.50, -1.60, -1.80)',$$

$$\lambda(t) = \left(\frac{\left(\frac{t}{142}\right)^{3/2}}{50} + .01 \right) \Big/ 2, \quad \Sigma = \begin{pmatrix} 0.04 & 0.03 \\ 0.03 & 0.06 \end{pmatrix},$$

and

$$V_{0l} = \begin{pmatrix} 0.20, & 0.25 \\ 0.25 & 0.75 \end{pmatrix}.$$

Using $S(s) = \exp\left\{-\int_0^s h(u)\,du\right\}$ and $S(s) \sim U(0, 1)$, we generated event times by randomly generating $S(s_i)$, then solving for s_i using the `uniroot()` and `integrate()` functions in `R` (Ihaka and Gentleman, 1996). Observations were censored with a probability of 0.9. A censored subject's censoring time was chosen uniformly over the interval $(0,\ s_i)$. We chose 20 time points per subject, indexed as t_{ij}, for the longitudinal measure observation times. About 50 data sets with 500 observations each were simulated. Five different models were fit to each data set with $q = 6$, 7, 8 and models that were linear ($\psi(t) = \beta_{i1} + \beta_{i2}t$) and quadratic ($\psi(t) = \beta_{i1} + \beta_{i2}t + \beta_{i3}t^2$) in time. For comparability with the B-spline models, these polynomial models were fit with a slightly modified prior on the β's, $\beta_{i1} \sim N_2(b_{01} + x_i\alpha,\ V_{01})$ and $\beta_{ik} \sim N_2(b_{0k},\ V_{0k})$, $k = 2, 3$.

Because we are interested in both how well the model estimates the parameters of interest and how well the model fits the data, we took two approaches to evaluating the results. First, to determine which model produced the least biased parameter estimates overall, we evaluated the simulations using the sums of the squared errors, $\text{SSE} = \sum_{i=1}^{50} (\theta_i - \hat{\theta}_i)' \times (\theta_i - \hat{\theta}_i)$, where $\theta = (\gamma_0,\ \gamma_1,\ \alpha_0,\ \alpha_1,\ \zeta)'$ and $\hat{\theta}$ represents the posterior mean, and the 95% credible intervals. The values are shown in Table 1 along with the mean bias in the estimates and the mean standard error of the estimates. In general, the SSE decreases as q increases. Also, the frequency that the 95% posterior credible interval contains the true value increases as q increases. The linear and quadratic models both perform worse than the spline models when estimating γ. The simple linear model, however, performs better than the quadratic model. The estimate of ζ from the linear model appears to be comparable to the spline models. Closer examination of the results from these simulations indicates that the quadratic linear model tends to positive or negative infinity more quickly than the simple linear model. This often results in a significant increase in the bias of the estimates of the trajectory at the event time. This bias in the trajectory may translate into larger biases in the estimates of γ_0 and γ_1. The fit of the linear model at the event time usually falls between the spline and quadratic fits. This may explain why there is closer agreement between the parameter estimates of the linear and spline models than between the quadratic and spline models. The coverage probabilities for γ and ζ from the quadratic and linear models are lower than for the spline models. Again, the linear model performs better than the quadratic model and is closer to the spline models with respect to the coverage probabilities. The poorer performance of the linear and quadratic models with respect to SSE and the coverage probabilities appears to be due to bias in the estimates and not the variability of the estimates (which appears to be similar for all the models as shown in Table 1). Based on these simulations, it appears that choosing a first-order (linear) approximation to the trajectory is preferable to a second-order (quadratic) approximation.

There is also an unusual twist in the setup of this simulation. The parameters γ and ζ indicate that increasing the trajectory and being on treatment decrease the hazard. However, α indicates that treatment decreases the trajectory. So, although the treatment's direct effect on the hazard is to decrease it, its effect through the trajectory is to increase it. This situation could occur in a medical setting. A desirable feature of this model is that it can identify such phenomena.

Because we also wanted to compare models to determine which may best fit the data, we used the CPO summary ($\sum_{i=1}^{N} \log(\widehat{\text{CPO}}_i)$) and the DIC as model evaluation tools. A

Table 1
Sum of squared errors (SSE), the percentage of simulations in which the 95% posterior credible interval contained the true parameter (%), the average bias, and the average standard error

		α_0	α_1	γ_0	γ_1	ζ
Linear	SSE	0.04	0.15	37.1	11.5	5.3
	%	96	98	74	80	80
	Bias	0.021	0.043	0.729	0.389	0.264
	SE	0.031	0.062	0.54	0.29	0.22
Quadratic	SSE	0.04	0.15	62.1	55.1	10.8
	%	96	98	70	20	60
	Bias	0.021	0.044	0.94	0.96	0.39
	SE	0.031	0.063	0.62	0.28	0.24
$q = 6$	SSE	0.03	0.10	12.73	5.93	3.66
	%	98	100	94	86	88
	Bias	0.023	0.035	0.38	0.26	0.21
	SE	0.023	0.041	0.52	0.26	0.21
$q = 7$	SSE	0.04	0.09	10.80	5.15	3.42
	%	88	98	96	84	92
	Bias	0.022	0.035	0.40	0.28	0.22
	SE	0.025	0.046	0.53	0.26	0.21
$q = 8$	SSE	0.03	0.09	10.21	4.58	3.25
	%	92	96	96	86	94
	Bias	0.022	0.032	0.36	0.25	0.21
	SE	0.022	0.038	0.52	0.26	0.22

second goal of using the CPO summary and the DIC with the simulated data was to determine how each could be used to select the correct model. If we chose the model with the largest CPO summary, we would select the correct model ($q = 7$) 48% of the time and a larger model 52% of the time. If we chose the model with the smallest DIC, we would select the correct model ($q = 7$) 76% of the time and a larger model 24% of the time. Neither the CPO summary nor the DIC selected a smaller model. The two measures both select the correct model 44% of the time.

6. Application

We use the methodology developed in this article to examine the relationship between CD4 count, viral load, and time to AIDS or death in a substudy of 391 patients from ACTG 175 (Katzenstein et al., 1996). Patients were randomized to one of four treatments: zidovudine (AZT), AZT plus zalcitabine (ddC), AZT plus didanosine (ddI), or ddI alone. For this analysis, we examined 198 of these patients who had biomarker measures taken after the start of treatment. To simplify the analysis for ease of exposition, we compare the AZT alone treatment to the remaining three treatments combined. This grouping is justified by the results in the original analysis. Because the first viral load and CD4 measures were taken at 8 weeks, we set time 0 for the survival analysis to be at 8 weeks. This did not exclude any patients from the subset of 198 already under investigation. We examined bivariate longitudinal measurements of CD4 count and viral load. Patients had a median of four measurements with a maximum number of five measurements and a minimum of one measurement. The endpoint was the time of the first event of either progression to AIDS or death. Of the 198 patients, 26 had an event based on this definition. Knots were placed at quantiles of the longitudinal observation times. In this analysis, we used the following noninformative priors:

$$\gamma \sim N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}50 & 0\\0 & 50\end{pmatrix}\right),$$

$$\alpha \sim N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}50 & 0\\0 & 50\end{pmatrix}\right),$$

$$\zeta \sim N(0, 50),$$

$$\lambda_j \sim \text{gamma}(1, 2), \quad j = 1, \ldots, J,$$

$$\Sigma^{-1} \sim \text{Wishart}\left(\begin{pmatrix}0.1 & 0\\0 & 0.1\end{pmatrix}, 2\right),$$

and

$$V_{0l}^{-1} \sim \text{Wishart}\left(\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}, 2\right), \quad l = 1, \ldots, q.$$

Convergence of the Gibbs sampler was assessed using standard methods. We fit B-spline models with $q = 5$ and $q = 6$ and $\alpha = 0$, $\zeta = 0$, $\gamma_0 = 0$, and $\gamma_1 = 0$. We also fit models where CD4 count and viral load were assumed to change linearly and quadratically with time as specified in Section 5. We attempted to fit a model with a cubic term in time, but the Gibbs sampler performed poorly, indicating a failure to converge. A summary of the results for the different models is shown in Table 2.

Table 2

Results for the different models examined. The parameter estimates are shown followed by their 95% *posterior credible intervals (VL = viral load, CD4 = CD4 lymphocyte count). The models where only CD4 or viral load were time-varying predictors in the survival model and q = 5 are indicated by the letters VL or CD4 after q in the first column.*

q	$\begin{matrix}e^{\gamma_{\text{VL}}}\\e^{\gamma_{\text{CD4}}}\end{matrix}$	$\begin{matrix}\alpha_{\text{VL}}\\\alpha_{\text{CD4}}\end{matrix}$	e^{ζ}	$\sum \log(\widehat{\text{CPO}}_i)$	DIC
2 (linear)	1.32 (1.04, 1.75)	0.79 (0.34, 1.25)	–	−1198.6	2130.5
	0.35 (0.22, 0.51)	−0.70 (−1.13, −0.28)			
3 (quadratic)♡	1.20 (0.95, 1.61)	0.77 (0.30, 1.19)	–	−1180.8	2146.1
	0.38 (0.24, 0.53)	−0.71 (−1.13, −0.29)			
5	1.46 (1.09, 2.05)	–	1.74 (0.67, 4.32)	−1023.9	1835.0
	0.30 (0.17, 0.45)	–			
5	1.48 (1.11, 2.09)	0.83 (0.48, 1.18)	1.25 (0.44, 3.31)	−1029.1	1834.2
	0.30 (0.18, 0.46)	−0.75 (−1.10, −0.40)			
5♠	1.50 (1.14, 2.13)	0.83 (0.49, 1.16)	–	−1015.7	1834.0
	0.30 (0.17, 0.46)	−0.77 (−1.11, −0.42)			
5VL	1.08 (0.79, 2.08)	0.82 (0.50, 1.15)	–	−1083.9	1948.4
	–	–			
5VL	1.12 (0.75, 1.92)	0.83 (0.51, 1.14)	1.98 (0.82, 4.59)	−1080.5	1951.1
	–	–			
5CD4	–	–	1.87 (0.74, 4.46)	−1066.3	1908.6
	0.37 (0.22, 0.53)	−0.78 (−1.10, −0.46)			
5CD4	–	–	–	−1056.2	1912.1
	0.38 (0.23, 0.56)	−0.79 (−1.13, −0.48)			
6	1.55 (1.14, 2.24)	0.79 (0.47, 1.12)	1.16 (0.43, 2.96)	−1031.3	1838.2
	0.32 (0.19, 0.48)	−0.77 (−1.10, −0.42)			
6♣	1.49 (1.09, 2.14)	–	1.71 (0.66, 4.29)	−1046.4	1838.1
	0.34 (0.20, 0.51)	–			
6	1.57 (1.18, 2.19)	0.79 (0.47, 1.11)	–	−1032.5	1839.0
	0.32 (0.18, 0.47)	−0.77 (−1.10, −0.44)			

The model marked by ♠ has the largest $\sum \log(\widehat{\text{CPO}}_i)$ and smallest DIC, indicating that it provides the best fit for the data. As shown in Table 2, changing values of q did not dramatically alter the estimates of γ, α, or ζ. However, the cubic spline models suggest a larger impact of both viral load and CD4 count on the hazard.

The results from the selected model (♠) tell us that a one-unit increase in the log of viral load corresponds to a 50% increase in the hazard and an increase in the CD4 cell count of 100 corresponds to a 70% reduction in the hazard. Also, AZT monotherapy results in an average increase over time equal to 0.83 log viral load with a 95% posterior interval equal to (0.49, 1.16) and a decrease of 75 (40, 110) in CD4 count. However, treatment did not have a statistically significant effect directly on the hazard as seen by the estimate of ζ.

Figure 1 shows Kaplan–Meier estimates with predicted survival curves for the models identified by ♡, ♠, and ♣ in Table 2. These were calculated according to the algorithm in the Appendix. The results represented by ♠ and ♣ were obtained using the cubic B-spline approach presented in this article. Also, the predicted survival curve from the model selected by the CPO and DIC statistics (♠) appears to give the best approximation to the Kaplan–Meier curve, indicating the best fit of these three models to the observed data.

We also fit models with γ_{VL} and γ_{CD4} set to 0. This corresponds to assessing the marginal effects of viral load and CD4 count on the time-to-event endpoint. The results for these models with $q = 5$ are shown in Table 2. The results for CD4 count are similar to the CD4 results from the ♠ model. However, the results for viral load are slightly different. The posterior credible intervals for the hazard ratios of viral load include 1.0, suggesting that higher viral load may not significantly decrease a patient's time to event. However, when adjusted for CD4 count as in ♠, viral load appears to be a better predictor of the endpoint. According to this model, treatment does have a statistically significant impact on viral load and CD4 count but does not impact the time-to-event significantly. None of the models with viral load or CD4 count alone as predictors was selected over the models with CD4 count and viral load as joint predictors by CPO or DIC. It appears that modeling viral load and CD4 count together gives more information about the endpoint. The results also suggest that CD4 is a more informative predictor of survival. The estimated mean and the endpoints of the 95% credible intervals for the hazard ratio for CD4 count are far from 1.0; however, for viral load, the lower endpoint of the 95% credible intervals are often close to or below 1.0, suggesting that it may not be a strong predictor.

Figures 2 and 3 show the trajectory fits for the ♡ and ♠ models for 25 subjects. The estimated treatment group curves are also shown on these plots in gray. These are equivalent to $\sum_{k=1}^{q} \hat{b}_{0k} B_k(t) + x\alpha$ for the spline model and $\sum_{k=1}^{3} \widehat{b_{0k}} t^{(k-1)} +$

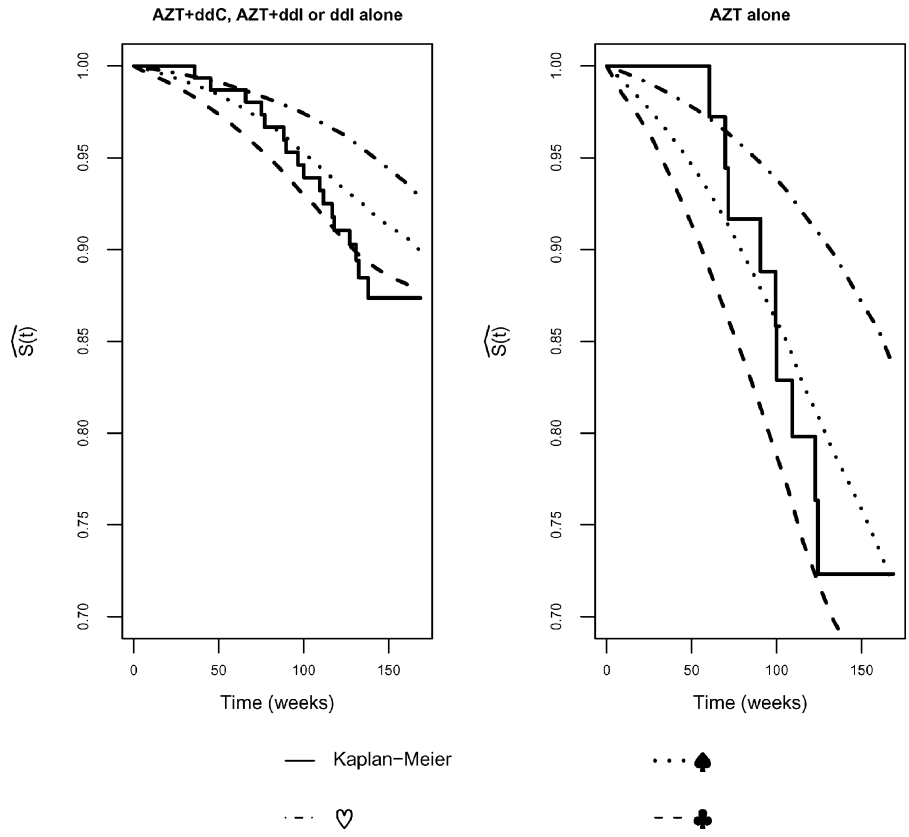

Figure 1. Predicted survival curves plotted against the Kaplan–Meier curve for three different models. The models are identified by symbols which link them to the results in Table 2.

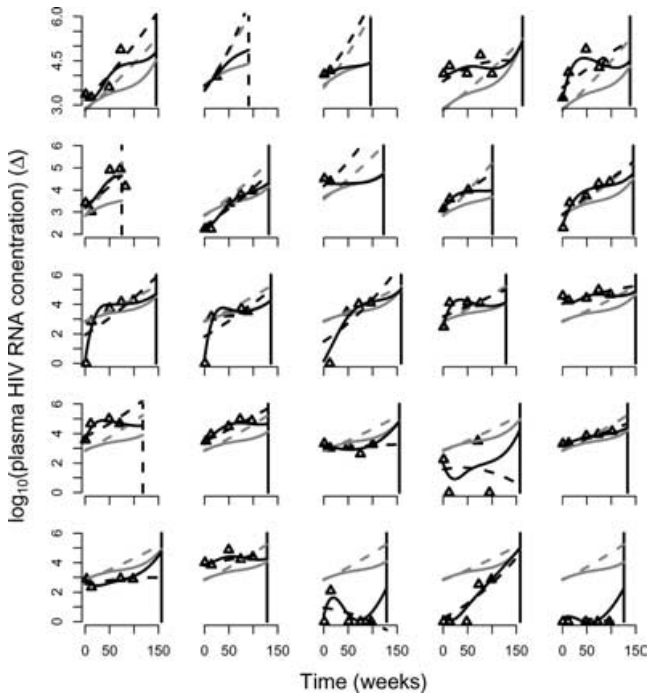

Figure 2. Observed and fitted longitudinal viral load measures for 25 patients. The black lines show the subject's fitted values for the ♡ (dashed) and ♠ (solid) models listed in Table 2. The gray lines indicate the same estimates for the population. The vertical line represents the censoring (solid) or event (dashed) time. In each row, the plots share the same vertical axis, which is labeled in the first plot of the row.

$x\alpha$ for the cubic model, where $\widehat{b_{0k}}$ indicates the posterior mean of b_{0k}. We see that using cubic B-splines to model the CD4 count and VL trajectories results in a flexible and dynamic model capable of modeling the variability in CD4 count and VL over time. Also, at later points in time, when fewer data are available, the spline model has the advantage of being weighted back to the population mean instead of continuing to increase or decrease and take on values outside of the range of the data.

Due to the differences between the simulation and the application, it is difficult to compare the results from the simulation to the results from this application. The simulation was not designed to mimic the ACTG data. The simulated longitudinal data fluctuate more over time than the ACTG data; therefore, we may expect to see more agreement between the spline and linear and quadratic estimates in the application than in the simulation. Also, because the two simulated longitudinal measures do not affect the hazard similarly to viral load and CD4 count, it is difficult to directly compare the estimates of γ in the two settings. The estimates of γ from the spline models agree more closely with the linear model than with the quadratic model. The same was true in the simulation; however, it does not appear that the difference between the spline and linear and quadratic models is as pronounced in the application as it is in the simulation. In the controlled treatment setting of this clinical trial, the variability in viral load over time is not as high as we might expect in a clinical setting where patients change treatments and viral load changes in response. In this setting, a linear model may estimate viral load and CD4 count adequately over time when the interest is estimating their impacting on time-to-event. In a setting where treatment changes over time and subjects have more advanced disease, we may expect more fluctuation in viral load and CD4 count. We might then expect to see a greater difference between the linear and spline models. The application results do agree with simulation results by selecting the spline models over the linear and quadratic models with respect to CPO and DIC. The application results suggest, as does the simulation, that a linear model is preferable to a quadratic model.

7. Discussion

In this article, we presented an approach to jointly modeling failure time and multivariate longitudinal data. Incorporating cubic B-splines results in a flexible and robust nonparametric longitudinal model. Because the model is able to change rapidly to reflect changes in the biomarkers over time, it may be preferable to parametric models for many types of data. A parametric trajectory lacks the flexibility offered by the cubic B-spline model. Other flexible models like the IOU model presented by Wang and Taylor (2001) would not be able to reflect significant changes in the level or direction of the trajectory due to the underlying parametric constraints of the

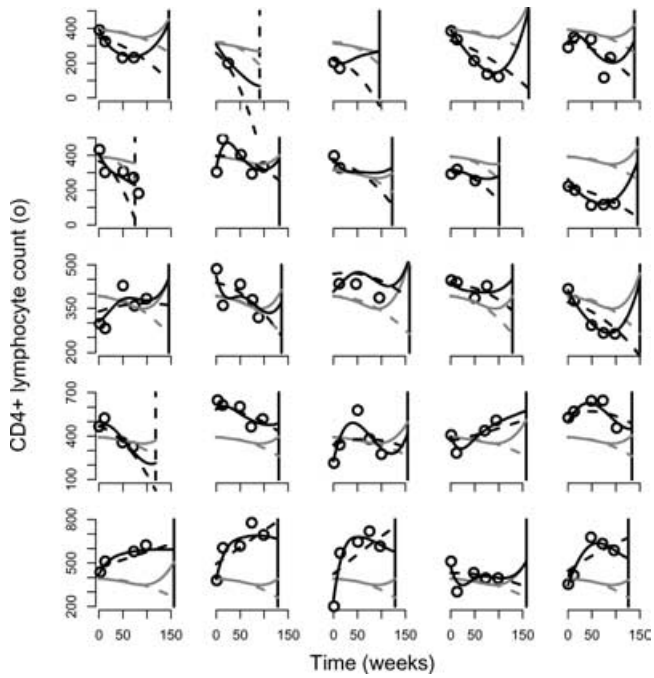

Figure 3. Observed and fitted longitudinal CD4 count measures for 25 patients. The black lines show the subject's fitted values for the ♡ (dashed) and ♠ (solid) models listed in Table 2. The gray lines indicate the same estimates for the population. The vertical line represents the censoring (solid) or event (dashed) time. In each row, the plots share the same vertical axis, which is labeled in the first plot of the row.

model. Another benefit of the longitudinal cubic B-spline trajectory model is the bound on the variance of the trajectory over time. This can be very important in the stability of the Gibbs sampler, and may better reflect uncertainty about the trajectory.

We were also able to parameterize the model to allow for a more flexible effect of treatment on the trajectory with respect to time. However, it may be possible to relax the constraints on the effect of treatment further. For example, we could include a separate parameter for treatment effect for each spline function, reflecting a time-varying effect of treatment on the marker. In this model, we assume the link between the biomarkers and survival (γ) is constant over time. In future research, we may also want to relax this assumption.

This modeling approach can also be useful to identify instances in which more dynamic parametric models may be useful. We have also proposed model selection techniques using the DIC and CPO that behave well in simulations and in practice. Other techniques for choosing the number of knots in the spline may be explored in the future. For example, we may want to specify a prior for q and then select the number of knots through MCMC sampling. We could also produce a more flexible model by investigating other distributions for the β_{ij}'s, like the t-distribution or a nonparametric distribution.

Acknowledgements

Research was supported in part by the University of Washington Department of Biostatistics Career Development award and NIH grants AI-51164, CI-70101, and CI-74015. The authors would like to thank the associate editor for helpful comments on the manuscript.

Résumé

Souvent, quand on utilise une modélisation simultanée de données longitudinales et de données de survie, on s'intéresse à une mesure longitudinale multivariée qui n'est pas forcément bien adaptée pour des modèles linéaires. Pour résoudre ce problème, nous proposons un modèle de survie et longitudinal joints, avec une composante non paramétrique pour les marqueurs longitudinaux. Des B-spline cubiques sont utilisés pour spécifier le modèle longitudinal et un modèle à risques proportionnels pour lier la mesure longitudinale à la mesure du risque. Pour estimer le modèle, nous utilisons un algorithme de Monte Carlo à chaîne de Markov. Le nombre de nœuds pour le B-spline cubique est sélectionné en utilisant l'ordonnée prédictive conditionnelle (CPO) et le critère d'information de la déviance (DIC). La méthode et l'approche de sélection du modèle sont validées par une simulation. On applique cette méthode pour examiner le lien entre la charge virale, la quantité de CD4 et le délai d'apparition des évènements à partir de données issues d'essais cliniques sur le SIDA. Le modèle à B-spline cubique fournit une bonne adéquation pour les données longitudinales qui ne peuvent pas être obtenues avec des modèles paramétriques simples.

References

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

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer-Verlag.

DeGruttola, V. and Tu, X. (1994). Modeling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics* **50,** 1003–1014.

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.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In *Bayesian Statistics*, Volume 4, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds), 147–159. Oxford: Oxford University Press.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85,** 398–409.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41,** 337–348.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S., and Merigan, T. C. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine* **335,** 1081–1090.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer Series in Statistics. New York: Springer-Verlag.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine studies. *Statistica Sinica* **14,** 863–883.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5,** 299–314.

Katzenstein, D. A., Hammer, S. M., Hughes, M. D., Gundacker, H., Jackson, J. B., Fiscus, S., Rasheed, S., Elbeik, T., Reichman, R., Japour, A., Merigan, T. C., and Hirsch, M. S., and The AIDS Clinical Trials Group Study 175 Virology Study Team (1996). The relation of virologic and immunologic markers to clinical outcomes after nucleoside therapy in HIV-infected adults with 200 to 500 CD4 cells per cubic millimeter. *New England Journal of Medicine* **335,** 1091–1098.

Lin, H., Turnbull, B. W., McCulloch, C. E., and Slate, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association* **97,** 53–65.

Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57,** 253–259.

Song, X., Davidian, M., and Tsiatis, A. (2002a). An estimator for the proportional hazards model with multiple

longitudinal covariates measured with error. *Biostatistics* **3,** 511–528.

Song, X., Davidian, M., and Tsiatis, A. (2002b). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics* **58,** 742–753.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of complexity and fit. *Journal of the Royal Statistical Society, Series B* **64,** 583–639.

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.

Wang, Y. and Taylor, J. M. G. (2001). Jointly modelling longitudinal and event time data with application to AIDS studies. *Journal of the American Statistical Association* **96,** 895–905.

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

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

Received September 2003. *Revised April* 2004.
Accepted April 2004.

Appendix

Estimating the Predicted Survival Curve

Given the G MCMC samples, $b_0^{(g)}$, $\alpha^{(g)}$, $\lambda^{(g)}$, $\gamma^{(g)}$, and $V_0^{(g)}$, $g = 1, \ldots, G$, we can estimate the predicted survival curve at a set of time points, $u = u_1, \ldots, J$, as follows.

Repeat the following steps for $g = 1, \ldots, G$.

1. Sample $\beta_{l,1}^* \sim N_2(b_{0l}^{(g)} + \alpha^{(g)}, V_{0l}^{(g)})$ and $\beta_{l,2}^* \sim N_2(b_{0l}^{(g)}, V_{0l}^{(g)})$.
2. Calculate $S_k^{(g)}(u_j) = \exp\{-\int_0^{u_j} \lambda^{(g)}(t) e^{\gamma'(g)'\psi_{\beta^*}(t)+\zeta^*(2-k)} dt\}$, $k = 1, 2$.

The predicted survival curve at time u_j is then $S_k(u_j) = \text{median}(S_k^{(g)}(u_j), g = 1, \ldots, G)$, $k = 1, 2$, where $k = 1$ indicates the treatment group and $k = 2$ is the nontreatment group.