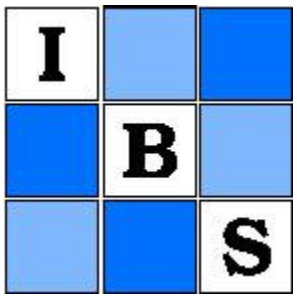

BIOMETRICS 53, 330–339
March 1997

A Joint Model for Survival and Longitudinal Data Measured with Error

Michael S. Wulfsohn

Gilead Sciences, 333 Lakeside Drive, Foster City, California 94404, U.S.A.

and

Anastasios A. Tsiatis

Department of Statistics, North Carolina State University,
Box 8203, Raleigh, North Carolina 27695-8203, U.S.A.

SUMMARY

The relationship between a longitudinal covariate and a failure time process can be assessed using the Cox proportional hazards regression model. We consider the problem of estimating the parameters in the Cox model when the longitudinal covariate is measured infrequently and with measurement error. We assume a repeated measures random effects model for the covariate process. Estimates of the parameters are obtained by maximizing the joint likelihood for the covariate process and the failure time process. This approach uses the available information optimally because we use both the covariate and survival data simultaneously. Parameters are estimated using the expectation–maximization algorithm. We argue that such a method is superior to naive methods where one maximizes the partial likelihood of the Cox model using the observed covariate values. It also improves on two-stage methods where, in the first stage, empirical Bayes estimates of the covariate process are computed and then used as time-dependent covariates in a second stage to find the parameters in the Cox model that maximize the partial likelihood.

1. Introduction

Many longitudinal studies collect information on outcomes such as infection or death, as well as covariates that vary with time. The covariates are usually measured intermittently, often at different times for each individual, with a different number of measurements for each individual, and with substantial error. The prognostic value of covariates is usually of interest in such studies, and the covariate process itself may be of interest, as it sheds light on the natural history of the disease.

In order to study the relationship of a covariate to survival, we can use the covariate as a time-dependent covariate in the proportional hazards regression model (Cox, 1972). It is necessary to have complete knowledge of the covariate history for all individuals while on study in order to maximize the partial likelihood and thereby estimate the model parameters. To do this optimally, we would need to know the covariate value as a time-continuous process without measurement error. The presence of random error in a measured covariate causes the estimated parameters to be biased towards the null (Prentice, 1982). A recent modeling approach dealing with survival as a function of a covariate that is measured repeatedly is a two-stage model, where in the first stage the covariate is modeled using growth curve models with random effects (Laird and Ware, 1982). At each event time, the individual random effects are estimated using empirical Bayes methodology. In the second stage, the modeled value is then substituted into the partial likelihood for the Cox model with time-dependent covariates, and the partial likelihood is then maximized (Dafni, 1993; Tsiatis, DeGruttola, and Wulfsohn, 1995). This modeling approach has been advocated on the basis that it reduces the bias of the parameter estimate in the Cox model.

Key words: EM algorithm; Longitudinal data; Repeated measures; Survival.

There are several drawbacks to this two-stage modeling approach. First, the assumption that the random effects are normally distributed in those at risk at each event time is probably unreasonable. If the covariate is predictive of survival, patients whose covariate trajectories have the steepest negative slopes may be at higher risk for mortality, and thus are removed from the population early on. This may result in the random effects having a distributional shift toward a nonnormal distribution as time progresses. The violation of the normality assumption possibly explains the residual bias observed in simulations of Dafni (1993). Second, a first-order approximation is required in order to use polynomial growth curve models to simplify the partial likelihood to be maximized (Tsiatis et al., 1995). The validity of this approximation depends on the scaling of the covariate. Third, the two-stage model does not use any survival information in modeling the covariate process, and thus information is not used as efficiently as it might be. Fourth, new growth curves are fitted to each individual datum at each new event so, for example, the expected baseline covariate value changes as further events happen. This is an undesirable property from the perspective of parsimony and interpretability.

We develop a methodology whereby a joint maximization of the likelihood from both the covariate process and the survival data occurs. This makes more efficient use of the data, using data from both the covariate and the survival process simultaneously. Because we estimate the parameters that describe the covariate process and those that describe the risk of failure as a function of the covariate process at the same time, our method uses not only the observed covariate data but also survival information to get estimates of the true covariate value at any time. We therefore can expect more precise and accurate estimates of the strength of the relationship between the covariate and the risk of failure. We assume that the random effects that parameterize the covariate process are normally distributed. In other words, we assume that an individual's random effects are constant over time and thus are identical at all event times when the individual is at risk. This is more reasonable than assuming that the covariates for those in the risk set will follow a normal distribution at all times. If individuals are removed from observation because of death or censoring, we do not require that the random effects for those remaining in the risk set be normally distributed. Pawitan and Self (1993) use a similar approach to ours, except they use a parametric form (a Weibull distribution) for their failure time process, and they use an exchangeable covariance matrix for the random effects, which is more restrictive than a fully parameterized covariance matrix. DeGruttola and Tu (1994) model the joint distribution of the covariate process and survival times as multivariate normal.

Our results are used to analyze the data from a completed double-blind placebo controlled trial conducted by Burroughs-Wellcome, which treated 281 patients with advanced disease due to the human immunodefiency virus, HIV. Of these, 137 patients were randomized to receive placebo and 144 patients were randomized to receive 250 mg dose of zidovudine, ZDV, every 4 hours. In this study, CD4 lymphocyte counts were determined prior to treatment and approximately every 4 weeks while on therapy. The median duration of follow-up was 18 weeks, at which point the study was stopped due to the superior results of the ZDV arm in decreasing mortality. At the termination of the study, all patients actively participating in the study were offered ZDV and subsequently followed for clinical outcomes.

The CD4 receptor is the portal of entry of the virus into the CD4 lymphocyte, and the CD4 lymphocyte count is known to be associated with clinical outcome (DeGruttola et al., 1993). The degree of this association is of interest from a prognostic viewpoint. The aim of this analysis is to understand the CD4 trajectories for the population of patients sampled in this study and to evaluate the strength of relationship between the CD4 trajectory and survival.

2. Model and Notation

For each individual, we observe survival data and covariate data. Denote by T_i the survival time for individual i. As in most clinical trials, we assume the survival data to be subject to right censoring. Therefore, we observe $X_i = \min(T_i, C_i)$, where C_i corresponds to a potential censoring time, and the failure indicator Δ_i, which is equal to 1 if the failure is observed ($T_i \leq C_i$) and 0 otherwise. Assume that censoring is independent of all other survival and covariate information.

We examine the case of a single covariate that is measured over time. The times at which the covariate is measured can be different for each individual. We assume that individual i has m_i measurements of the covariate and there are n individuals. The covariate for individual i is measured at times $\mathbf{t}_i = (t_{ij} : t_{ij} \leq X_i)$, where t_{ij} is the time from randomization and $j = 1, \ldots, m_i$. The values at those times are $\mathbf{Z}_i = (Z_{ij} : t_{ij} \leq X_i)$. The observed data available for each individual is therefore $(X_i, \Delta_i, \mathbf{Z}_i, \mathbf{t}_i)$.

There is a distribution for the potential times $\mathbf{t}_i$; however, this will be ancillary to the problem. If, however, the observation times of the covariate are predictive of the covariate trajectory or

survival, these times or functions thereof can be incorporated as covariates in the growth curve model. Previous work has shown that the pattern of delay in visits does not have a notable effect on the parameters for this dataset (Tsiatis et al., 1995).

The CD4 count distribution at baseline was skewed to the right, probably because these patients had advanced HIV disease. A log transformation normalized the data. From here on we shall refer to the transformed covariate as $\mathbf{Z}_i$.

There was a lot of heterogeneity in the baseline CD4 values between individuals and the slopes of the CD4 trajectories. It is therefore reasonable to model the covariate using a growth curve model with random effects.

For simplicity, we demonstrate this approach by modeling the covariate with a linear growth curve model with random intercept and slope. So, at times that the covariate is measured, we observe

$$Z_{ij} = \theta_{0i} + \theta_{1i} t_{ij} + e_{ij}.$$

We assume that the error e_{ij} is from an $\mathrm{N}(0, \sigma_e^2)$ distribution, that $\mathrm{cov}(e_{ij}, e_{ij'}) = 0$ (where $j \neq j'$), and that error is independent of the random intercept and slope, θ_{0i} and θ_{1i}. Assume that the individual slopes and intercepts are distributed according to a bivariate normal, i.e.,

$$\begin{pmatrix} \theta_{0i} \\ \theta_{1i} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}, \begin{pmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{01} & \sigma_{11} \end{pmatrix} \right).$$

In vector notation, we write the above as $\boldsymbol{\theta}_i \sim \mathrm{N}(\boldsymbol{\theta}, \boldsymbol{\sigma})$.

We model the hazard of failure through the proportional hazards model. We use the original Cox model formulation, where the hazard depends on the covariate through its current value. Other aspects of the trajectory can also be considered. We assume that the true covariate value is that given by the growth model with random effects that we are using. So

$$\begin{aligned} \lambda(t \mid \boldsymbol{\theta}_i, \mathbf{Z}_i, \mathbf{t}_i) &= \lambda(t \mid \boldsymbol{\theta}_i) \\ &= \lambda_0(t) \exp\left\{\beta(\theta_{0i} + \theta_{1i} t)\right\}. \end{aligned}$$

We can thus model the covariate and failure processes jointly in one larger metamodel. The model specifications outlined apply specifically to the Burroughs-Wellcome dataset, and the linear growth curve model for the covariate and the dependence of the hazard on the current value of the covariate are reasonable for modeling the data for patients randomized to placebo. The appropriate model for modeling the covariate process of patients randomized to ZDV is a piecewise linear spline with a knot at 4 weeks. This model is described by Tsiatis et al. (1995), and for simplicity we only present the model for patients randomized to placebo. These methods apply to a wide variety of models.

The observed data for each individual is $(X_i, \Delta_i, \mathbf{Z}_i, \mathbf{t}_i)$. We do not get to observe the random effects $\boldsymbol{\theta}_i$ for the growth curve model. The observed data likelihood is given by

$$\prod_{i=1}^{n} \left[\int_{-\infty}^{\infty} \left\{ \prod_{j=1}^{m_i} f(z_{ij} \mid \boldsymbol{\theta}_i, \sigma_e^2) \right\} f(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}, \mathbf{V}) f(X_i, \Delta_i \mid \boldsymbol{\theta}_i, \boldsymbol{\lambda}_0, \beta)\, d\boldsymbol{\theta}_i \right], \tag{2.1}$$

where

$$\begin{aligned} f(z_{ij} \mid \boldsymbol{\theta}_i, \sigma_e^2) &= \left(2\pi\sigma_e^2\right)^{-1/2} \exp\{-(z_{ij} - \theta_{0i} - \theta_{1i} t_{ij})^2 / 2\sigma_e^2\}, \\ f(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}, \mathbf{V}) &= (2\pi|\mathbf{V}|)^{-1/2} \exp\{-(\boldsymbol{\theta}_i - \boldsymbol{\theta})' \mathbf{V}^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\theta})/2\}, \end{aligned}$$

and

$$\begin{aligned} &f(X_i, \Delta_i \mid \boldsymbol{\theta}_i, \boldsymbol{\lambda}_0, \beta) \\ &\quad = [\lambda_0(X_i) \exp\{\beta(\theta_{0i} + \theta_{1i} X_i)\}]^{\Delta_i} \exp\left[- \int_0^{X_i} \lambda_0(u) \exp\{\beta(\theta_{0i} + \theta_{1i} u)\}\, du \right] \end{aligned} \tag{2.2}$$

The density for the survival data in equation (2.2) assumes that the current value of the covariate is the appropriate component of the covariate history to use in the model. Using the two-stage model, the current value of CD4 was the most predictive aspect of the trajectory, and other features of the path such as baseline value and slope did not add significantly to the log likelihood (Tsiatis

et al., 1995). The survival density can be generalized to be a function of several aspects of the covariate history.

The model assumes that censoring is independent of the random effects. This is often the case in many clinical trials. If, however, the censoring does depend on covariates, we would have to incorporate the appropriate density for the censoring process in the model. If the censoring process that leads to drop-out is not correctly modeled, the estimated parameters may be biased.

We estimate the parameters $\boldsymbol{\theta}, \mathbf{V}, \sigma_e^2$, and β using parametric maximum likelihood and $\lambda_0(u)$ using nonparametric maximum likelihood. We denote this set of parameters by $\boldsymbol{\Omega}$. The baseline hazard $\lambda_0(u)$ takes mass at each failure time, and the dimension of $\boldsymbol{\lambda}_0$, the vector of these baseline hazard quantities, is equal to the number of unique failure times. In order to find $\widehat{\boldsymbol{\Omega}}$, we use the expectation–maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977).

3. Parameter Estimation Using the EM Algorithm

The purpose of the EM algorithm is to estimate parameters of interest by maximizing the likelihood of the observed data. This is done by iterating between an E-step, where we compute the expected log-likelihood of the complete data conditional on the observed data and the current estimate of the parameters, and an M-step, where new parameter estimates are computed by maximizing this expected log-likelihood.

The complete data for each individual is $(X_i, \Delta_i, \mathbf{Z}_i, \mathbf{t}_i, \boldsymbol{\theta}_i)$. All components except $\boldsymbol{\theta}_i$ are observed. The complete data likelihood for each individual is the integrand of the expression given by (2.1), and the complete data likelihood is the product of these quantities over all individuals. In the maximization step, we solve the score equations. There exist closed-form maximum likelihood estimates for all parameters except β, for which we use a one-step Newton–Raphson. Most of the parameter estimates in the maximization step depend on the conditional expectation of some function of $\boldsymbol{\theta}_i$, namely $\mathrm{E}\{h(\boldsymbol{\theta}_i) \mid X_i, \Delta_i, \mathbf{Z}_i, \mathbf{t}_i, \widehat{\boldsymbol{\Omega}}\}$, which we denote by $\mathrm{E}_i\{h(\boldsymbol{\theta}_i)\}$. This expectation can be considered as a conditioning on the first four data elements after the bar and as a function of $\boldsymbol{\Omega}$ evaluated at its estimate. This expectation is evaluated in the expectation step of the EM algorithm. We iterate between the expectation step and the maximization step until the parameter estimates converge.

The closed-form maximum likelihood estimates are

$$\widehat{\boldsymbol{\theta}} = \sum_{i=1}^{n} \mathrm{E}_i(\boldsymbol{\theta}_i)/n \tag{3.1}$$

$$\widehat{\mathbf{V}} = \sum_{i=1}^{n} \mathrm{E}_i(\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}})(\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}})'/n \tag{3.2}$$

$$\widehat{\sigma}_e^2 = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \mathrm{E}_i(z_{ij} - \theta_{0i} - \theta_{1i} t_{ij})^2 / \sum_{i=1}^{n} m_i \tag{3.3}$$

$$\widehat{\lambda}_0(u) = \sum_{i=1}^{n} \frac{\Delta_i I(X_i = u)}{\sum_{j=1}^{n} \mathrm{E}_j[\exp\{\beta(\theta_{0j} + \theta_{1j} u)\}] Y_j(u)}, \tag{3.4}$$

where $Y_j(u)$ is an at risk indicator equal to $I(X_j \geq u)$.

The parameter of interest, β, is updated at each iteration using a one-step Newton–Raphson algorithm. So at the kth iteration, our estimate is

$$\hat{\beta}_k = \hat{\beta}_{k-1} + I_{\hat{\beta}_{k-1}}^{-1} S_{\hat{\beta}_{k-1}},$$

where $S_{\hat{\beta}_{k-1}}$ is the score for beta at the $k-1$st iteration, described in Appendix A, and $I_{\hat{\beta}_{k-1}}$ is the information for beta at the $k-1$st iteration and is equal to the negative derivative of the above quantity.

In the expectation step, we need to compute $\mathrm{E}\{h(\boldsymbol{\theta}_i) \mid X_i, \Delta_i, \mathbf{Z}_i, \mathbf{t}_i, \widehat{\boldsymbol{\Omega}}\}$. The new term $\widehat{\boldsymbol{\Omega}}$ denotes the parameters estimated in the maximization step, namely $\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{V}}, \widehat{\sigma}_e^2, \widehat{\boldsymbol{\lambda}}_0$, and $\hat{\beta}$. The conditional density of $\boldsymbol{\theta}_i$ given the observed data and the current estimate of the parameters is equal to

$$f(\boldsymbol{\theta}_i \mid X_i, \Delta_i, \mathbf{Z}_i, \mathbf{t}_i, \widehat{\boldsymbol{\Omega}}) = \frac{f(\boldsymbol{\theta}_i, X_i, \Delta_i \mid \mathbf{Z}_i, \mathbf{t}_i, \widehat{\boldsymbol{\Omega}})}{f(X_i, \Delta_i \mid \mathbf{Z}_i, \mathbf{t}_i, \widehat{\boldsymbol{\Omega}})}$$

$$= \frac{f(X_i, \Delta_i \mid \boldsymbol{\theta}_i, \widehat{\boldsymbol{\lambda}}_0, \hat{\beta}) f(\boldsymbol{\theta}_i \mid \mathbf{Z}_i, \mathbf{t}_i, \widehat{\boldsymbol{\theta}}, \widehat{\mathbf{V}}, \widehat{\sigma}_e^2)}{\displaystyle\int_{-\infty}^{\infty} f(X_i, \Delta_i \mid \boldsymbol{\theta}_i, \widehat{\boldsymbol{\lambda}}_0, \hat{\beta}) f(\boldsymbol{\theta}_i \mid \mathbf{Z}_i, \mathbf{t}_i, \widehat{\boldsymbol{\theta}}, \widehat{\mathbf{V}}, \widehat{\sigma}_e^2)\, d\boldsymbol{\theta}_i}.$$

The conditional expectation of any function h of the random effects is denoted by $\mathrm{E}\{h(\boldsymbol{\theta}_i) \mid X_i, \Delta_i, \mathbf{Z}_i, \mathbf{t}_i, \widehat{\boldsymbol{\Omega}}\}$. This expectation is taken with respect to the preceding density and is equal to

$$\frac{\displaystyle\int_{-\infty}^{\infty} h(\boldsymbol{\theta}_i) f(X_i, \Delta_i \mid \boldsymbol{\theta}_i, \widehat{\boldsymbol{\lambda}}_0, \hat{\beta}) f(\boldsymbol{\theta}_i \mid \mathbf{Z}_i, \mathbf{t}_i, \widehat{\boldsymbol{\theta}}, \widehat{\mathbf{V}}, \widehat{\sigma}_e^2)\, d\boldsymbol{\theta}_i}{\displaystyle\int_{-\infty}^{\infty} f(X_i, \Delta_i \mid \boldsymbol{\theta}_i, \widehat{\boldsymbol{\lambda}}_0, \hat{\beta}) f(\boldsymbol{\theta}_i \mid \mathbf{Z}_i, \mathbf{t}_i, \widehat{\boldsymbol{\theta}}, \widehat{\mathbf{V}}, \widehat{\sigma}_e^2)\, d\boldsymbol{\theta}_i}. \tag{3.5}$$

The density $f(X_i, \Delta_i \mid \boldsymbol{\theta}_i, \widehat{\boldsymbol{\lambda}}_0, \hat{\beta})$ is specified by model (2.2). The density $f(\boldsymbol{\theta}_i \mid \mathbf{Z}_i, \mathbf{t}_i, \widehat{\boldsymbol{\theta}}, \widehat{\mathbf{V}}, \widehat{\sigma}_e^2)$ can be derived from the joint distribution of $\mathbf{Z}_i$ and $\boldsymbol{\theta}_i$, which is multivariate normal, i.e.,

$$\begin{pmatrix} \mathbf{Z}_i \\ \boldsymbol{\theta}_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \theta_0 \mathbf{1}_i + \theta_1 \mathbf{t}_i \\ \boldsymbol{\theta} \end{pmatrix}, \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{pmatrix} \right),$$

where $\mathbf{1}_i$ denotes a vector of 1's of the same dimension as $\mathbf{t}_i$. The elements of the partitioned covariance matrix are

$$\mathbf{W}_{11} = \left\{ \begin{array}{ccc} (1 \quad t_{i1})\mathbf{V}\begin{pmatrix} 1 \\ t_{i1} \end{pmatrix}, & \cdots, & (1 \quad t_{i1})\mathbf{V}\begin{pmatrix} 1 \\ t_{im_i} \end{pmatrix} \\ \vdots & \ddots & \vdots \\ (1 \quad t_{im_i})\mathbf{V}\begin{pmatrix} 1 \\ t_{i1} \end{pmatrix}, & \cdots, & (1 \quad t_{im_i})\mathbf{V}\begin{pmatrix} 1 \\ t_{im_i} \end{pmatrix} \end{array} \right\} + I_{m_i}\sigma_e^2,$$

$$\mathbf{W}_{21} = \begin{pmatrix} \sigma_{00} + \sigma_{01}t_{i1}, & \cdots, & \sigma_{00} + \sigma_{01}t_{im_i} \\ \sigma_{01} + \sigma_{11}t_{i1}, & \cdots, & \sigma_{01} + \sigma_{11}t_{im_i} \end{pmatrix}$$

$$\mathbf{W}_{12} = \mathbf{W}_{21}',$$

$$\mathbf{W}_{22} = \mathbf{V}.$$

It follows from standard normal theory that the conditional distribution of $\boldsymbol{\theta}_i$ given $\mathbf{Z}_i$, the density that we require, is

$$\boldsymbol{\theta}_i \mid \mathbf{Z}_i \sim \mathrm{N}(\boldsymbol{\theta} + \mathbf{W}_{21}{\mathbf{W}_{11}}^{-1}(\mathbf{Z}_i - \theta_0\mathbf{1}_i - \theta_1\mathbf{t}_i), \mathbf{W}_{22} - \mathbf{W}_{21}{\mathbf{W}_{11}}^{-1}\mathbf{W}_{12}).$$

Introducing new notation, the distribution of $\boldsymbol{\theta}_i \mid \mathbf{Z}_i$ is $\mathrm{N}(\mathbf{W}_{\mathbf{Z}_i}, \mathbf{W}_{\mathbf{Z}_i})$. The expectation in (3.5) of any function of $\boldsymbol{\theta}_i$ was evaluated using an m-point Gauss–Hermite quadrature formula (Press et al., 1992). This is exact for all polynomials of degree $2m - 1$ or less. We experimented with various choices of m and found that the answers did not change appreciably for $m > 2$ (four grid points for two dimensions) for all expectations required in formulas (3.1) through (3.4). We consequently used two-point Gauss–Hermite quadrature for all the expectations evaluated.

The survival and growth curve densities in this model are both conditional on the unobserved random effects. The conditional estimates of the random effects and function of the random effects are required for the maximization step in which population parameters are estimated. The general form of these estimates is represented by expression (3.5), and the example given in Appendix B shows that these estimates are a function of both the longitudinal covariate data and the survival data. Thus, all the available data is used in determining the population parameters for the growth curve and the survival components of the model.

Starting values used for the growth curve parameters in the Newton–Raphson were the least squares estimates for a linear model if the person had two or more observations. If there was only one observation, the population slope was used as the slope, and the slope together with the single observation were used to calculate the intercept. These individual parameters were used to estimate covariate values at times required in the partial likelihood for the Cox model with time-dependent covariates.

In order to estimate the variance of β, we first define the profile score $S_\beta(\widehat{\boldsymbol{\Omega}}_{-\beta}(\hat{\beta}))$ to be the derivative of the log likelihood with respect to β evaluated at $\hat{\beta}$, with the remaining parameters, which we will refer to as $\boldsymbol{\Omega}_{-\beta}$, estimated using restricted maximum likelihood estimates. The profile score is given by equation (A.2) in Appendix A. The profile score uses the restricted maximum

likelihood estimates $\widehat{\boldsymbol{\Omega}}_{-\beta}(\hat{\beta})$, which were calculated using a separate EM algorithm applied to the likelihood, keeping $\hat{\beta}$ fixed. It follows from standard likelihood theory that the asymptotic variance for $\hat{\beta}$ is given by the negative inverse of the slope of the profile score at the maximum likelihood estimates.

4. Results

The parameter estimates obtained by maximizing the joint likelihood using the EM algorithm were similar to those from the two-stage model. The comparison for patients on placebo is presented below. Note that there were 23 unique event times in the placebo group, and we selected the 8th and 16th events because these correspond to the first and third terciles, respectively. We only selected two events for the two-stage model to illustrate the heterogeneity of the population parameters over time. There was, however, no time trend in these population parameters. Table 1 gives the parameters for the growth curve model. The value of $\hat{\beta}$ that describes the strength of relationship between CD4 and survival was -0.3029 with the joint maximization and -0.284 with the two-stage model.

Table 1
Parameters for growth curve model

	Event	θ_0	θ_1	σ_e^2	σ_{00}	σ_{01}	σ_{11}
Two-stage model	8th	4.23	-0.0050	0.301	1.18	0.0016	0.000032
	16th	4.27	-0.0045	0.305	1.15	0.0024	0.000029
Joint model		4.17	-0.0047	0.396	1.11	0.0027	0.000014

In Figure 1, we plot the profile scores for β as a function of β. Note that while there is a slight negative curvature, the fit is approximately linear, as expected, with a slope of -41.980. Thus, the asymptotic standard error of $\hat{\beta}$ is $(41.980)^{-1/2} = 0.154$. This is in contrast to 0.144 in the two-stage model. The standard error for $\hat{\beta}$ is greater when using the joint estimation procedure because the random effects are assumed to be influenced by the uncertainty in the estimated growth curve parameters; thus, more variability is incorporated. The standard error for pure measurement error is also greater using the joint estimation because there is no overfitting, as is the case with the two-stage model. In the two-stage model, the covariate trajectories from baseline are continually changing as more data becomes available, thus leading to overly optimistic agreement between the model and the data at early times when there is only a small amount of data. The standard errors for other parameters can be estimated by similar manipulations of the appropriate profile scores.

5. Discussion

We have presented a technique whereby we can jointly model the covariate and hazard processes. All the survival and covariate data is used to get parameter estimates for the covariate and hazard processes, and our utilization of all the available information is necessary for optimal inference. This is in contrast to the two-stage model, in which only covariate data is used to find the parameters for the covariate process and then the hazard is a function of this modeled covariate.

Simulations of Dafni (1993) indicated that the two-stage model yielded parameter estimates that were somewhat biased towards the null, even though the majority of the bias using more naive modeling approaches is eliminated. The parameter estimate of β we get from the joint model is further from the null hypothesis than that of the two-stage model, indicating that we have probably reduced the bias even further. It is also anticipated that, if censoring is associated with the history of the covariate level, the parameter estimates should be unbiased. Further work is required to investigate the degree to which the joint modeling approach results in unbiased estimates.

We have used a particular specification of a joint model that is appropriate to the dataset that we analyzed. The approach is generalizable to many different modeling situations. Simpler applications would include fixed covariates at baseline. More complex models can also be fitted. For example, we may be interested in more than the current level of a single covariate as a predictor of survival. We may also want to stratify the survival on some aspect of the covariate trajectory if the proportional hazards assumption is not satisfied. Improved growth curve models for the longitudinal covariate may result when polynomial models are explored and where aspects of the measurement times are incorporated as covariates for the model. Also, different subsets in the sample may require

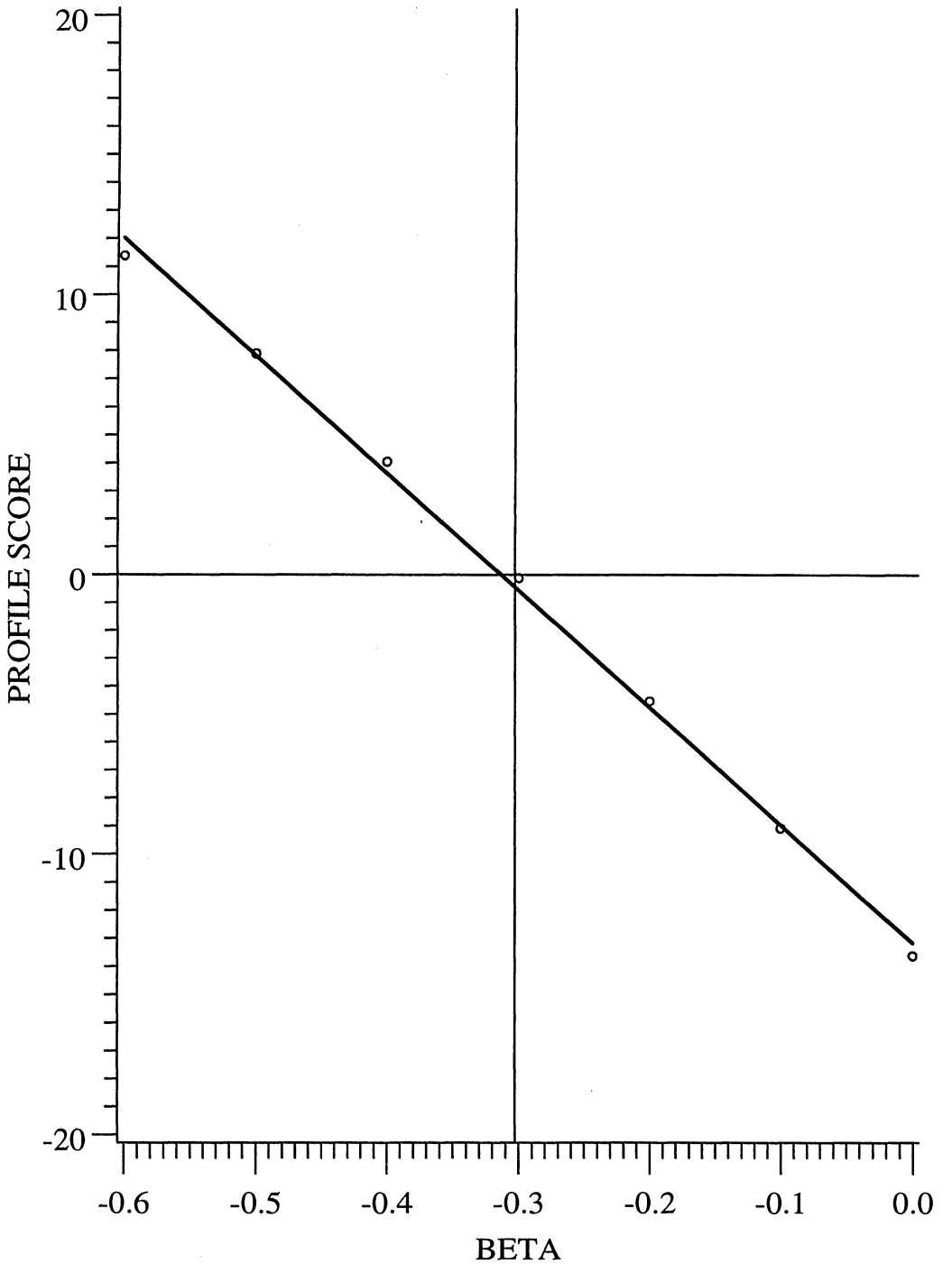

Figure 1. Profile score as a function of β.

different polynomial growth curve models. Other avenues to explore include model selection using likelihood techniques and model checking. Future work will be directed toward developing the relevant simplifications and generalizations of this joint modeling approach and toward developing software for public use.

The EM algorithm that we employ to estimate our parameters of interest is reliable and, for this dataset, converges to valid parameter estimates even when these estimates are exceptionally close to their boundary values. One problem with the algorithm is that it is slow. In this particular example, with 137 individuals with 24 failures and approximately 6 covariate measurements

per individual, it takes about 2 hours to find parameter estimates. This is even after using reasonable starting values for the parameters. We are currently exploring the possibility of using a Newton–Raphson approach to replace the EM algorithm for estimating our parameters of interest. A Newton–Raphson approach, similar to that used by Lindstrom and Bates (1988), would also facilitate the evaluation of the standard errors of our parameter estimates by inverting the observed information matrix.

Acknowledgements

This research was supported by grants from the National Institutes of Health and was done while both authors were in the Department of Biostatistics at the Harvard School of Public Health.

Résumé

La relation entre une covariable longitudinale et un processus des délais de survie peut être évalué par le modèle à risques proportionnels de Cox. Ici on considère le problème rencontré dans l'estimation des paramètres dans le modèle de Cox quand la covariable longitudinale n'est pas observé fréquemment et peut contenir des erreurs de mesures. On suppose que le processus de la covariable peut être modélisé par un modèle à mesures répétées à effets aléatoires. L'estimation des paramètres est obtenue en maximisant la vraisemblance jointe du processus de la covariable et du processus des délais de survie. Cette approche utilise l'information disponible de façon optimale, puisque la covariable et les délais de survie sont utilisés en même temps. Les paramètres sont maximisés par l'algorithme espérance-maximisation. Notre argument consiste à dire qu'une telle méthode est supérieure aux méthodes naïves où la vraisemblance partielle du modèle de Cox est maximisée en utilisant les valeurs des covariables observées. Aussi, cette méthode est une amélioration aux méthodes à deux étapes où on calcule les estimations Bayesiennes empiriques du processus des covariables dans la première étape. On utilise ensuite ces estimations comme covariables dépendant du temps dans la deuxième étape pour trouver les paramètres dans le modèle de Cox qui maximise la vraisemblance partielle.

References

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

Dafni, U. (1993). Evaluating surrogate markers of clinical outcome when measured with error. Ph.D. dissertation, Biostatistics Library, Harvard School of Public Health.

DeGruttola, V. and Tu, X. M. (1994). Modelling progression of CD-4 lymphocyte count and its relationship to survival time. *Biometrics* **50,** 1003–1014.

DeGruttola, V., Wulfsohn, M., Fischl, M., and Tsiatis, A. A. (1993). Modeling the relationship between survival and CD4 lymphocytes in patients with AIDS and AIDS-related complex. *Journal of AIDS* **6,** 359–365.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39,** 1–38.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38,** 963–974.

Lindstrom, M. J. and Bates, D. M. (1988). Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association* **83,** 1004–1022.

Pawitan, Y. and Self, S. (1993). Modeling disease marker processes in AIDS. *Journal of the American Statistical Association* **88,** 719–726.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69,** 331–342.

Press, W. H., Teutolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in FORTRAN. The Art of Scientific Computing,* 2nd edition. New York: Cambridge University Press.

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.

Received May 1995; *revised June* 1996; *accepted September* 1996.

APPENDIX A

Score and Information

In the maximization step of the EM algorithm, we evaluate the parameter estimates by setting the derivatives of the expectation of the complete log likelihood conditional on the observed data equal to zero. Call the observed data D_o.

In the Newton–Raphson algorithm, we require the observed score and information at the current value of the parameters in order to iterate to the next set of parameter estimates. If $\boldsymbol{\theta}$ is our current estimate of the parameter vector, an updated estimate is $\hat{\boldsymbol{\theta}}$, which is estimated from

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \mathbf{I}_{\boldsymbol{\theta}}^{-1}\mathbf{S}_{\boldsymbol{\theta}}, \tag{A.1}$$

where $\mathbf{S}_{\boldsymbol{\theta}}$ is the score and $\mathbf{I}_{\theta}$ is the observed information.

So the score is required by both the EM and Newton–Raphson algorithms, and is calculated as follows:

The conditional expectation of the complete log likelihood can be considered in three pieces. The first piece contains the parameter σ_e^2 and is

$$\mathrm{E}\left\{\log \prod_{i=1}^{n}\prod_{j=1}^{m_i} f(z_{ij} \mid \boldsymbol{\theta}_i, \sigma_e^2) \mid D_o\right\} = -\frac{1}{2}\sum_{i=1}^{n} m_i \log \sigma_e^2 - \frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i} \mathrm{E}_i(z_{ij} - \theta_{0i} - \theta_{1i}t_{ij})^2,$$

and the derivative with respect to σ_e^2 is equal to zero, i.e.,

$$-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n} m_i + \frac{1}{2\sigma_e^4}\sum_{i=1}^{n}\sum_{j=1}^{m_i} \mathrm{E}_i(z_{ij} - \theta_{0i} - \theta_{1i}t_{ij})^2 = 0,$$

and the maximum likelihood estimate of σ_e^2 is given by equation (3.3).

The second piece contains the parameters $\boldsymbol{\theta}$ and $\mathbf{V}$. The conditional expectation of this piece of the complete log likelihood is

$$\mathrm{E}\left\{\log \prod_{i=1}^{n} f(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}, \boldsymbol{\sigma}) \mid D_o\right\} = -\frac{n}{2}\log|\mathbf{V}| - \frac{1}{2}\sum \mathrm{E}_i\{(\boldsymbol{\theta}_i - \boldsymbol{\theta})'\mathbf{V}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\theta})\}.$$

The derivative of this term with respect to $\boldsymbol{\theta}$ is $\mathbf{V}\Sigma\{\mathrm{E}_i(\boldsymbol{\theta}_i - \boldsymbol{\theta})\}$ and the maximum likelihood estimate of $\boldsymbol{\theta}_i$ is given by equation 3.1.

Using the transformation $\mathbf{U} = \mathbf{V}^{-1}$, we can differentiate with respect to $\mathbf{U}$. Note that $\mathbf{V}$ and $\mathbf{U}$ are symmetric matrices, so the derivative with respect to $\mathbf{U}$ is

$$\begin{aligned}\frac{d}{d\mathbf{U}}&\left[\frac{n}{2}\log|\mathbf{U}| - \frac{1}{2}\mathrm{tr}\left\{\mathbf{U}\sum \mathrm{E}_i\left(\boldsymbol{\theta}_i - \frac{\sum \boldsymbol{\theta}_i}{n}\right)\left(\boldsymbol{\theta}_i - \frac{\sum_{i=1}^{n}\boldsymbol{\theta}_i}{n}\right)'\right\}\right] \\ &= \frac{n}{2}\left\{2\mathbf{U}^{-1} - \mathrm{diag}(\mathbf{U}^{-1})\right\} - \frac{1}{2}\left[2\sum \mathrm{E}_i\left(\boldsymbol{\theta}_i - \frac{\sum \boldsymbol{\theta}_i}{n}\right)\left(\boldsymbol{\theta}_i - \frac{\sum \boldsymbol{\theta}_i}{n}\right)'\right. \\ &\qquad \left. -\mathrm{diag}\left\{\sum \mathrm{E}_i\left(\boldsymbol{\theta}_i - \frac{\sum \boldsymbol{\theta}_i}{n}\right)\left(\boldsymbol{\theta}_i - \frac{\sum \boldsymbol{\theta}_i}{n}\right)'\right\}\right].\end{aligned}$$

This leads to the maximum likelihood estimate of $\mathbf{V}$ given by equation (3.2).

The third piece contains the parameters $\boldsymbol{\lambda}_0$ and β. The conditional expectation of this piece of the complete log likelihood is

$$\sum \Delta_i \log \lambda_0(X_i) + \sum \Delta_i \beta\{\mathrm{E}_i(\theta_{0i}) + \mathrm{E}_i(\theta_{1i})X_i\} - \sum \int_0^{X_i} \lambda_0(u)\mathrm{E}_i[\exp\{\beta(\theta_{0i} + \theta_{1i}u)\}]\,du.$$

Differentiating with respect to $\lambda_0(u)$, we get

$$\sum\left[\frac{\Delta_i I(X_i = u)}{\lambda_0(u)} - \mathrm{E}_i[\exp\{\beta(\theta_{0i} + \theta_{1i}u)\}]Y_i(u)\right],$$

where $Y_i(u) = I(X_i \geq u)$. This leads to the maximum likelihood estimate of $\lambda_0(u)$ given by equation (3.4). The baseline hazard is evaluated at each of the ordered failure times, denoted by $u_1, \ldots, u_k$.

Differentiating the third piece with respect to β gives

$$\sum_{i=1}^{n} \left[\Delta_i \{ \mathrm{E}_i(\theta_{0i}) + \mathrm{E}_i(\theta_{1i}) X_i \} - \sum_{j=1}^{k} \lambda_0(u_j) \mathrm{E}_i[(\theta_{0i} + \theta_{1i} u_j) \exp\{\beta(\theta_{0i} + \theta_{1i} u_j)\}] Y_i(u_j) \right]. \quad \text{(A.2)}$$

Note that $\lambda_0(u_j)$ is a function of β, so this is not a closed-form solution. At the maximum likelihood estimate of $\lambda_0(u)$ given by equation (3.4), the score S_β is

$$\sum_{i=1}^{n} \Delta_i \left[\mathrm{E}_i(\theta_{0i}) + \mathrm{E}_i(\theta_{1i}) X_i - \frac{\sum_{j=1}^{k} \mathrm{E}_i[(\theta_{0i} + \theta_{1i} u_j) \exp\{\beta(\theta_{0i} + \theta_{1i} u_j)\}] Y_i(u_j)}{\sum_{j=1}^{k} \mathrm{E}_i[\exp\{\beta(\theta_{0i} + \theta_{1i} u_j)\}] Y_i(u_j)} \right].$$

Appendix B

Gauss–Hermite Quadrature

Estimation of parameters in the maximization step of the EM algorithm involves the estimation of many quantities of the form $\mathrm{E}_i\{h(\theta_i)\}$. These can be evaluated from equation (3.5) using Gauss–Hermite quadrature. Note that constants are identical in numerator and denominator and thus will cancel. For example, $\mathrm{E}_i[\exp\{\beta(\theta_{0i} + \theta_{1i} t)\}]$ can be evaluated as follows. The nonconstant terms in the numerator are

$$\int \int_{-\infty}^{\infty} \exp \Big[\beta(\theta_{0i} + \theta_{1i} t) + \Delta_i \beta(\theta_{0i} + \theta_{1i} X_i) - \int_0^{X_i} \lambda_0(u) \exp\{\beta(\theta_{0i} + \theta_{1i} u)\} \, du - (\theta_i - \theta_{\mathbf{Z}_i})' \mathbf{W}_{\mathbf{Z}_i}^{-1} (\theta_i - \theta_{\mathbf{Z}_i})/2 \Big] d\theta_i.$$

In order to evaluate this, we first transform θ_i to $\gamma_i = (2\mathbf{W}_{\mathbf{Z}_i})^{-1/2}(\theta_i - \theta_{\mathbf{Z}_i})$. This new vector has an $\mathrm{N}(0, (1/2)I)$ distribution. The components γ_{0i} and γ_{1i} are thus independent and normally distributed. We can use Gauss–Hermite quadrature to integrate over γ_i, which is a function of θ_i. The transformation can be inverted to get θ_i as a function of γ_i. Let's call this $\theta_i' = (2\mathbf{W}_{\mathbf{Z}_i})^{1/2}\gamma_i + \theta_{\mathbf{Z}_i}$. Using this inverse transformation and ignoring constants, which will cancel, the numerator becomes

$$\int \int_{-\infty}^{\infty} \exp \Big[\beta(\theta_{0i}' + \theta_{1i}' t) + \Delta_i \beta(\theta_{0i}' + \theta_{1i}' X_i) - \int_0^{X_i} \lambda_0(u) \exp\{\beta(\theta_{0i}' + \theta_{1i}' u)\} \, du - ({\gamma_{0i}}^2 + {\gamma_{1i}}^2) \Big] d\gamma_i.$$

Evaluating this expression using an m-point Gauss–Hermite quadrature formula, the above expression becomes

$$\sum_{j=1}^{m} \sum_{k=1}^{m} \exp \left[\beta(\theta_{0i}' + \theta_{1i}' t) + \Delta_i \beta(\theta_{0i}' + \theta_{1i}' X_i) - \int_0^{X_i} \lambda_0(u) \exp\{\beta(\theta_{0i}' + \theta_{1i}' u)\} \, du \right] w_j w_k,$$

where γ_{0i} takes on m abscissa values x_j $(j = 1, \ldots, m)$ and γ_{1i} takes on m abscissa values x_k $(k = 1, \ldots, m)$. The weights associated with these abscissa values are w_j and w_k. Note that θ_{0i}' is a function of x_j and x_k, and so is θ_{1i}'.