Biometrics 61, 379–391
June 2005

DOI: 10.1111/j.1541-0420.2005.00323.x

Partly Conditional Survival Models for Longitudinal Data

Yingye Zheng

Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., M2-B230, P.O. Box 19024, Seattle, Washington 98109-1024, U.S.A.
email: yzheng@fhcrc.org

and

Patrick J. Heagerty

Department of Biostatistics, University of Washington, P.O. Box 357232, Seattle, Washington 98195-7232, U.S.A.

Summary. It is common in longitudinal studies to collect information on the time until a key clinical event, such as death, and to measure markers of patient health at multiple follow-up times. One approach to the joint analysis of survival and repeated measures data adopts a time-varying covariate regression model for the event time hazard. Using this standard approach, the instantaneous risk of death at time t is specified as a possibly semiparametric function of covariate information that has accrued through time t. In this manuscript, we decouple the time scale for modeling the hazard from the time scale for accrual of available longitudinal covariate information. Specifically, we propose a class of models that condition on the covariate information through time s and then specifies the conditional hazard for times t, where $t > s$. Our approach parallels the "partly conditional" models proposed by Pepe and Couper (1997, *Journal of the American Statistical Association* **92,** 991–998) for pure repeated measures applications. Estimation is based on the use of estimating equations applied to clusters of data formed through the creation of derived survival times that measure the time from measurement of covariates to the end of follow-up. Patient follow-up may be terminated either by the occurrence of the event or by censoring. The proposed methods allow a flexible characterization of the association between a longitudinal covariate process and a survival time, and facilitate the direct prediction of survival probabilities in the time-varying covariate setting.

Key words: Estimating equation; Joint model; Multivariate survival.

1. Introduction

Methods that can quantify the risk of death or disease as a function of current or past covariate measures can provide medical predictions, which are used to guide patient care. Altman and De Stavola (1994, p. 321) comment: "In clinical practice individual data are routinely collected at frequent time points after entry to a study . . . but are rarely examined in relation to survival. Yet a key clinical question is that of prognosis, and a means of updating prognosis on the basis of the latest observations on a patient would be valuable to many clinicians." For example, the likelihood of HIV infection advancing to an acquired immunodeficiency syndrome (AIDS) diagnosis may depend on a patient's observed history of CD4 T lymphocyte cell counts (Fusaro, Nielsen, and Scheike, 1993). Similarly, in cancer screening the advent of new molecular assays has led to a number of measurements that may have the ability to signal cancer onset. For prostate cancer the serum measurement prostate-specific antigen (PSA) has been studied (Etzioni et al., 1999; Slate and Turnbull, 2000) and for ovarian cancer the marker CA-125 has been used as a possible early indicator of disease (Skates, Pauler, and Jacobs, 2001). These examples illustrate the common biomedical data structure where a longitudinal measurement is taken at select follow-up times, and the scientific question focuses on the correlation between the longitudinal series and the time until a key clinical event.

One approach that links the time until an event to time-dependent covariates is the proportional hazards model of Cox (1972). Details of the time-varying covariate proportional hazards model are given in monographs by Kalbfleisch and Prentice (1980), Cox and Oakes (1984), and Andersen et al. (1995). Briefly, in a time-varying proportional hazards model the instantaneous risk of death is modeled as a function of the current value of the measured covariate. Let $Z_i(t)$ denote the value of the covariate for subject i at time t, and let T_i denote the time until a major clinical endpoint (i.e., disease or death). Classical survival analysis methods for time-varying covariates can be used to model the instantaneous risk of death, or hazard defined as $\lambda\{t \mid H_i^Z(t)\} = \lim_{\delta \to 0} \frac{1}{\delta} P\{T_i \in [t, t+\delta) \mid H_i^Z(t); T_i \geq t\}$ where we condition on the entire covariate history $H_i^Z(t) = \{Z_i(u) : u \leq t\}$. The hazard may depend on additional aspects of the measured history beyond

the current measurement. However, for simplicity of presentation, we assume that the hazard depends only on the present value of the marker. The proportional hazards model assumes $\lambda\{t \mid Z_i(t)\} = \lambda_0(t) \cdot \exp\{\beta \cdot Z_i(t)\}$, where $\lambda_0(t)$ represents a baseline hazard function.

The use of the time-varying covariate model typically assumes that $Z_i(t)$ is available for all possible times. However, in practice, we almost never observe $Z_i(t)$ continuously in time. Rather, we commonly measure the covariate process at discrete times $s_{i1}, s_{i2}, \ldots, s_{in_i}$. Altman and De Stavola (1994) discuss some of the practical issues associated with discrete covariate measurement. Prentice (1982) discusses issues of bias that result from mismeasurement of covariates, and the recent literature has developed models that require a joint model for covariate process and failure time (Pawitan and Self, 1993; Tsiatis, DeGruttola, and Wulfsohn, 1995; Faucett and Thomas, 1996; Henderson, Diggle, and Dobson, 1997; Wulfsohn and Tsiatis, 1997; Xu and Zeger, 2001; Lin et al., 2002).

The time-varying covariate hazard model is particularly useful for estimating regression parameters. However, unlike Cox models without time-varying covariates, estimates of survival probabilities are generally difficult to obtain, and require a model for the covariate process. In practice, we may monitor a patient's vital status through time s and have available a discretely measured covariate history $H_i^Z(s) = \{Z_i(s_{ik}) : s_{ik} \leq s\}$ and seek to estimate patient prognosis on the basis of observed data. For example, when using a model for $\lambda\{t \mid H_i^Z(t)\}$ estimating $P\{T_i > t \mid H_i^Z(s),\ 0 \leq s < t\}$ would require either knowledge of the future values of $Z_i(t)$, or integration over the conditional distribution of the future covariate process given the history: $\{Z_i(u) : u > s \mid H_i^Z(s)\}$. Therefore, although a joint model can characterize the distribution of longitudinal measurements and a survival time, use of a joint model for predictions requires both correct specification of the likelihood, and necessitates numerical integration to obtain survival probability estimates. A general method is desired that can directly structure and estimate $P\{T_i > t \mid Z_i(s), 0 \leq s < t\}$, for any pair of survival and measurement times, (s, t), where $s < t$.

The prediction of future failure, or "residual lifetime," based on a measured past marker process has been an area of interest in AIDS research (Taylor et al., 1990; Jewell and Nielsen, 1993; Jewell and Kalbfleisch, 1996). In particular, Shi et al. (1996) considered parametric models of residual time to AIDS as a function of months since infection and CD4 measurements. In this manuscript we introduce a semiparametric method, which we call a "partly conditional survival model," for estimating the prognostic effect of longitudinal measurements on survival without relying on multivariate assumptions regarding the longitudinal marker process. Here, we address the question concerning how well a longitudinal covariate measured at, or up to time s, $Z_i(s)$, predicts the risk of the occurrence of an important clinical event, T_i such as diagnosis or death, by any future time t. Specifically, we are interested in the conditional probabilities $P\{T_i > t \mid Z_i(s), 0 \leq s < t\}$ that can be computed for any time $t > s$ but condition only on the marker value through time s. In order to characterize a general conditional survival distribution, we focus our model on hazard functions of the form

$$\lambda\{t \mid Z_i(s), 0 \leq s < T_i\} = \lim_{\delta \to 0} \frac{1}{\delta} \cdot P\{T_i \in [t, t+\delta) \mid Z_i(s), 0 \leq s < t\}. \quad (1)$$

The basic idea is that we can consider the marker value at time s as a "frozen" or baseline measurement rather than explicitly as a realization of a continuous time stochastic process. Also, in a typical survival model, the event time T_i is usually the time from study entry to the occurrence of the event. To estimate how well a marker can predict subsequent survival, similar to the modeling strategy used by Shi et al. (1996), we suggest modeling the time scale in terms of years since measurement, $T_i - s$. As a result there can be multiple derived event times for each individual corresponding to her/his repeatedly measured marker values, $T_{ik} = T_i - s_{ik}$, where s_{ik} is the kth marker measurement time. Thus, we cast the problem of using longitudinal marker values for predicting survival within the general framework of multivariate survival models. Our partly conditional survival model is similar to marginal Cox regression models (Wei, Lin, and Weissfeld, 1989; Lee, Wei, and Amato, 1992) in the sense that we do not make any parametric assumptions about the dependence among the survival times from one individual. However, the proposed model is "partly conditional," because when modeling hazards we do not condition on the full and dynamic covariate history, $H_i^Z(t)$, but rather on a static covariate subset, $H_i^Z(s)$ for fixed $s < t$.

One issue that arises with the use of partly conditional models is the need to allow regression parameters to depend on both the time of measurement for the predictor and the time of measurement for the outcome. Pepe, Heagerty, and Whitaker (1999) give examples where the linear predictor for the mean at time t conditional on covariate information through time s takes the form $\beta_0(t, s) + \beta_1(t, s) \cdot Z_i(s)$. In specific examples, a varying-coefficient model of the form $\beta_1(t, s) = \beta_1(t - s)$ may be used which assumes that the association between the outcome and the covariate depends only on their time separation. In the survival setting, we also anticipate the need for time-varying coefficient models because a time-varying measure may not satisfy the standard proportional hazards assumption. A Cox model with coefficient functions such as $\beta_1(t - s)$ may be adopted allowing the association between the marker measured at time s, $Z_i(s)$, and the hazard of death at future times t, to change as the distance, $t - s$, increases.

A formal definition of our partly conditional survival model is given in Section 2. In Section 3, we describe the estimation procedures and characterize large sample properties. In Section 4, we discuss simulations that evaluate both coverage probabilities and efficiency of the proposed estimation methods. In Section 5, we illustrate the new methods by analyzing a well-known data set from HIV research, the Multicenter Aids Cohort Study (MACS) data.

2. Partly Conditional Models in Survival Analysis

2.1 *Notation*

Let T_i be the time to diagnosis (or failure) for subject i. We assume that T_i may be censored at time C_i, and therefore we

only observe $X_i = \min(T_i, C_i)$ and an associated censoring indicator Δ_i, where $\Delta_i = 1$ if $X_i = T_i$ and 0 otherwise. Also, assume that each subject in the study has a time-dependent covariate measured $K_i \leq K$ times during follow-up, where K is relatively small compared to the total number of subjects n. Let $\mathbf{Z}_{ik}^{\mathrm{T}} = (\mathbf{Z}_{i\cdot a}^{\mathrm{T}}, Z_{ikb}, s_{ik})$ denote a vector of covariates associated with subject i measured at time s_{ik}, where $\mathbf{Z}_{i\cdot a}$ denotes a vector of baseline covariates such as treatment or gender, while Z_{ikb}, or equivalently $Z_i(s_{ik})$, denotes a time-varying marker value measured at time s_{ik}.

For the longitudinal analysis setting, we need to explicitly state the model assumptions regarding both measurement and missingness. First, we assume that the censoring time, C_i, is independent of the survival time T_i. Second, we assume that the measurement times, s_{ik}, are independent of the longitudinal marker process, $Z_i(s_{ik})$, and the survival time T_i. Finally, we assume that subjects may have missing marker measurements, but we assume that any such missingness is completely at random (MCAR). In Section 6, we comment on approaches that can relax these measurement and missingness mechanism assumptions.

In the partly conditional survival approach, we focus analysis on derived survival times. Corresponding to $\mathbf{Z}_{ik}$, let T_{ik} denote the time from s_{ik} to T_i, and C_{ik} denote the time from s_{ik} to C_i for censored $T_i : T_{ik} = T_i - s_{ik}$ and $C_{ik} = C_i - s_{ik}$. For failure time T_{ik}, one observes a bivariate vector (X_{ik}, Δ_{ik}), where $X_{ik} = \min(T_{ik}, C_{ik})$ and $\Delta_{ik} = 1$ if $X_{ik} = T_{ik} > 0$ and 0 otherwise. To specify hazards we convert to the derived time scale, $t^* = t - s_{ik}$, which measures the follow-up time since measurement of the marker. We use the standard counting process notation, where $N_{ik}(t^*) = I(X_{ik} \leq t^*, \Delta_{ik} = 1)$, and write $dN_{ik}(t^*)$ for the increment $N_{ik}\{(t^* + dt)-\} - N_{ik}(t^*)$. The at-risk process is defined as $R_{ik}(t^*) = I(X_{ik} \geq t^*, T_i > s_{ik})$. In situations where the covariate at time s_{ik} is not measured due to an MCAR mechanism we modify the at-risk process definition: $R_{ik}(t^*) = I(X_{ik} \geq t^*, T_i > s_{ik}) \times O_{ik}$, where $O_{ik} = 1$ if the kth covariate measurement is available, and $O_{ik} = 0$ otherwise. We assume that the random vectors $(\mathbf{X}_i, \boldsymbol{\Delta}_i, \mathbf{R}_i, \mathbf{Z}_i)$ are independent and identically distributed with $\mathbf{Z}_i$ bounded. In addition, we assume the censoring time C_{ik} is independent of T_{ik} conditional on $\mathbf{Z}_{ik}$. Because the measurement time, s_{ik}, is part of $\mathbf{Z}_{ik}$, the conditional independence of $T_{ik} = T_i - s_{ik}$ and $C_{ik} = C_i - s_{ik}$ follows from the assumption of independence for T_i and C_i given the marker process and baseline covariates.

2.2 *Partly Conditional Survival Model*

We now propose a class of methods that may be used to estimate the survival probability conditional on a longitudinal marker value, $P(T_i > t \,|\, \mathbf{Z}_{ik}, T_i > s_{ik})$, or equivalently $P(T_{ik} > t^* \,|\, \mathbf{Z}_{ik}, T_i > s_{ik})$, which we call "partly conditional Cox regression models." We define the partly conditional hazard function $\lambda_{ik}(t^*)$ for the derived survival outcomes T_{ik} as

$$\lambda_{ik}\big(t^* \,\big|\, \mathbf{Z}_{ik}, 0 \leq s_{ik} \leq T_i\big) = \lim_{\delta \to 0} \frac{1}{\delta} \cdot P\big(t^* \leq T_{ik} < t^* + \delta \,|\, T_{ik} \geq t^*, \mathbf{Z}_{ik}, T_i \geq s_{ik}\big).$$

A regression model for the hazard can take the general form: $\lambda_{ik}(t^* \,|\, \mathbf{Z}_{ik},\, 0 \leq s_{ik} \leq T_i) = g\{\lambda_0(t^*,\, s),\, \boldsymbol{\beta}(t^*,\, s)^{\mathrm{T}}\, \mathbf{Z}_{ik}\}$, where $g(\lambda,\, \eta)$ is a link function. The baseline hazard, $\lambda_0(t^*,\, s)$, and the regression coefficient, $\boldsymbol{\beta}(t^*,\, s)$, may be functions of both the time since measurement, t^*, and measurement time, s (or equivalently t and s). The model is "partly conditional" because rather than conditioning on the entire covariate history through time t, the model conditions on the partial history measured through time s. Although we present a general form, we will focus on the proportional hazards model.

With the partly conditional model, we address several complications that frequently arise when using survival analysis with longitudinal measurements. First, we anticipate that the phenomenon of nonproportional hazards may be more frequently encountered when covariates are updated over time. To this end, we specify a time-varying coefficient Cox model which can be used without imposing any functional form on the coefficient functions. Second, the time at which the measurement is taken, s_{ik}, may be associated with the predictive capacity of $Z_i(s_{ik})$ on survival.

Here, we briefly describe some examples of partly conditional survival models that model both $Z_i(s_{ik})$ and the measurement time s_{ik}, in particular we show how the dependence of $\boldsymbol{\beta}(t^*,\, s)$ on s can be handled by different model specifications. The simplest approach would be to create separate regression models for survival beyond key measurement times. Such a stratified method would define G "measurement intervals" $\mathcal{I}_1, \ldots, \mathcal{I}_G$, that partition time $\{t_0,\, \max(T_i)\}$ and then adopt the G models:

$$\lambda_{ik}\big(t^* \,\big|\, \mathbf{Z}_{ik}, 0 \leq s_{ik} \leq T_i, s_{ik} \in \mathcal{I}_g\big) = \lambda_{0g}(t^*) \exp\big\{\boldsymbol{\alpha}_g^{\mathrm{T}} \mathbf{Z}_{i\cdot a} + \beta_g(t^*) Z_{ikb}\big\}, \quad t^* > 0. \tag{2}$$

Here, $\lambda_{0g}(t^*)$ is an unspecified baseline hazard function, whereas $\boldsymbol{\alpha}_g$ and $\beta_g(t^*)$ are unknown regression parameters that are unique for each measurement interval. If only at most one measurement per subject fell into $\mathcal{I}_g$, then separate baseline Cox models could be estimated with standard methods using the measurement Z_{ikb}, where $s_{ik} \in \mathcal{I}_g$, as a baseline measurement for interval g, and the residual lifetime, $T_i - s_{ik}$ as the outcome, after restricting to those subjects who survive long enough to have a measurement recorded in $\mathcal{I}_g$. Such models are sometimes referred to as "landmark analyses" (Anderson, Cain, and Gelbber, 1983) when only a small number of measurement times are chosen to identify the time origins for analysis. The major limitation to a fully stratified approach is the lack of parsimony with completely unstructured baseline hazard and relative risks as a function of the measurement time s. The partly conditional models that we propose allow adoption of smooth variation in the coefficient functions $\beta_g(t^*)$ and the baseline hazards by explicitly modeling the measurement time s.

A second model assumes that the prognostic capacity of the longitudinal marker is the same regardless of the time at which it is measured. However, the model may still allow different baseline hazard functions for different measurement intervals $\mathcal{I}_g$:

$$\lambda_{ik}\big(t^* \,\big|\, \mathbf{Z}_{ik}, 0 \leq s_{ik} \leq T_i\big) = \lambda_{0g}(t^*) \exp\big\{\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Z}_{i\cdot a} + \beta(t^*) Z_{ikb}\big\}, \quad t^* > 0. \tag{3}$$

Alternatively, rather than assuming completely separate baseline hazards, $\lambda_{0g}(t^*)$, we may use the time of measurement as

a covariate in a model with a common unspecified baseline hazard. A general form for this model is

$$\lambda_{ik}\left(t^* \mid \mathbf{Z}_{ik}, 0 \le s_{ik} \le T_i\right) = \lambda_0(t^*)\exp\left[\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{Z}_{i\cdot a} + \boldsymbol{\alpha}_2^{\mathrm{T}}\{f_j(s_{ik})\}_{j=1}^{p} + \beta(t^*)Z_{ikb}\right], \quad t^* > 0, \quad (4)$$

where $\{f_j(s)\}_{j=1}^{p} = \{f_1(s), f_2(s), \ldots, f_p(s)\}$, and $f_j(s_{ik})$ represent basis functions for some parametric but flexible function of s_{ik}, such as a cubic spline with fixed knots. The modeling procedure thus does not rely on the partition of measurement stratum anymore. In the proposed partly conditional approach, we are using regression methods to specify a general hazard function $\lambda(t^*, s, z) = \lambda_i(t^* \mid 0 \le s \le T_i, Z_i = z)$ which can be used to obtain predictions as a function of residual lifetime (t^*), measurement time (s), and marker value (z). Jewell and Nielsen (1993) also investigate such predictive functions and provide "global compatibility criteria" (their equation (2.4)), which ensures that the function $\lambda(t^*, s, z)$ provides mutually compatible conditional survival probabilities. In particular, Jewell and Nielsen (1993) show that $\lambda(t^*, s, z)$ must satisfy certain restrictions when averaged over the survival process and the marker process from time s to time $s + t^*$. A direct verification that a given partly conditional model satisfies the global condition of Jewell and Nielsen (1993) is not possible because we do not specify the marker distribution. However, models of the form $\lambda(t^*, s, z) = \lambda_0(t^*)\exp\{\alpha f_1(s) + \beta(t^*, s)\, f_2(z)\}$ only impose smoothness assumptions through the choice of how the measurement time is incorporated, $f_1(s)$, how the marker is incorporated, $f_2(z)$, and how the relative hazard changes with residual and measurement time, $\beta(t^*, s)$. Therefore, these models impose weak assumptions and should be a good approximation to a range of assumptions on the marker and survival processes.

Estimation procedures for partly conditional models depend on whether we allow the influence of covariates to vary with time, $\beta(t^*)$, or assume a constant covariate effect $\beta(t^*) = \beta$. We distinguish among three classes of partly conditional models: proportional hazards; varying-coefficient hazards; and combined proportional and varying-coefficient hazards models. For a proportional hazards model, the effects of all covariates, including the longitudinal marker, are assumed constant. For a varying-coefficient hazards model, the effects of all covariates vary with time, and are estimated nonparametrically. Finally, for a combined proportional and varying-coefficient hazards model, the influence of only a few covariates varies nonparametrically over time, while the remaining covariates have time-invariant effects.

3. Estimation

To estimate the regression parameters and the baseline hazard function, we propose use of "working independence" estimating equations applied to the derived failure time data $(X_{ik}, \Delta_{ik}, R_{ik}, \mathbf{Z}_{ik})$. Because we have chosen to directly model the partly conditional hazard function for these multiple correlated failure times, a likelihood-based estimation approach would be analytically and computationally difficult for two reasons. First, in a likelihood approach a joint model would be required for the event time and the repeated measures process. Parameterization of the joint model in terms of the partly conditional hazards would be analytically difficult as the regression structure we adopt is for pairwise marginal distributions induced by the joint model. Second, a likelihood-based approach would generally require proper parametric specification of the longitudinal covariate distribution, and the validity of the induced partly conditional regression estimates would depend on correct marker model specification. As an alternative, we develop a direct estimating equation approach that proves computationally simple and yields consistent estimators under correct specification of the partly conditional regression structure without reliance on any distributional assumptions for the marker process. Sandwich variance estimators permit valid asymptotic inference.

In the subsections below, we discuss the estimation of regression parameters for three specific model classes. We first discuss estimation under a standard proportional hazards assumption, and then discuss relaxation to allow a nonparametric varying-coefficient specification. Finally, we discuss a model that allows both parametric and nonparametric components.

3.1 *Proportional Hazards Model*

For the situation where the proportional hazard assumption holds, estimation procedures for the partly conditional models are similar to those for marginal survival models (Wei et al., 1989; Lee et al., 1992). We assume that the baseline hazard is a function of the time since measurement, $t^* = t - s$, (hereafter referred to as the identical baseline model), or more generally, a function of both t^* and s (hereafter referred to as the stratified baseline model), and the effect of the marker on the failure time is constant over time. Specifically, the hazard function for the ith subject and the failure corresponding to the measurement at s_{ik} is

$$\lambda_{ik}\left(t^* \,\middle|\, \mathbf{Z}_{ik}, 0 \le s_{ik} \le T_i\right) = \lambda_0(t^*)\exp\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_{ik}\right)$$

or in the stratified model

$$\lambda_{ik}\left(t^* \,\middle|\, \mathbf{Z}_{ik}, 0 \le s_{ik} \le T_i\right) = \lambda_{0g}(t^*)\exp\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_{ik}\right)$$

where $\lambda_{0g}(t^*)$, $g = 1, \ldots, G$ denote the unspecified baseline hazard function that corresponds to the gth measurement interval $\mathcal{I}_g$ for $s_{ik} \in \mathcal{I}_g$. The unknown parameter $\boldsymbol{\beta}$ can be obtained by solving the "working-independence" estimating equation:

$$\sum_{i}^{n}\sum_{k}^{K}\int_0^{\tau}\{\mathbf{Z}_{ik} - \bar{\mathbf{Z}}(u)\}\, dN_{ik}(u) = 0$$

where $\tau = \inf[t^* : E\{R_{ik}(t^*)\} = 0]$, $\bar{\mathbf{Z}}(u) = \sum_{g=1}^{G}\mathcal{S}_g^{(1)}(\boldsymbol{\beta}, u)/\sum_{g=1}^{G}\mathcal{S}_g^{(0)}(\boldsymbol{\beta}, u)$ under the identical baseline model and $\bar{\mathbf{Z}}(u) = \sum_{g=1}^{G}\mathcal{S}_g^{(1)}(\boldsymbol{\beta}, u)1(s_{ik} \in \mathcal{I}_g)/\sum_{g=1}^{G}\mathcal{S}_g^{(0)}(\boldsymbol{\beta}, u)1(s_{ik} \in \mathcal{I}_g)$ under the stratified baseline model, with

$$\mathcal{S}_g^{(j)}(\beta, u) = \frac{1}{n}\sum_{l=1}^{n}\sum_{m=1}^{K}R_{lm}(u)\exp\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_{lm}\right)\mathbf{Z}_{lm}^{\otimes j}\cdot 1(s_{lm} \in \mathcal{I}_g).$$

For a column vector a, $a^{\otimes 0}$ refers to a scalar 1, $a^{\otimes 1}$ refers to the vector a, and $a^{\otimes 2}$ refers to the matrix aa^{T}.

Large sample distributional theory and robust variance–covariance estimation can be developed along the lines of the

marginal survival model of Lee et al. (1992) for the identical baseline models, or Wei et al. (1989) for the stratified baseline models.

3.2 *Varying-Coefficient Hazards Model*

The partly conditional regression model decouples the marker measurement time, s_{ik}, from the time scale for the hazard model, $t^* = t - s_{ik}$ for $t > s_{ik}$. Therefore, the hazard ratio corresponding to the longitudinal marker, $\lambda_{ik}(t^* \mid \mathbf{Z}_{i\cdot a},\, Z_{ikb} = (z+1),\, s_{ik})/\lambda_{ik}(t^* \mid \mathbf{Z}_{i\cdot a},\, Z_{ikb} = z,\, s_{ik}) = \mathrm{HR}(t^*)$ may not be constant over time, and methods that allow relaxation of the standard proportional hazards assumption will be important in practice. For example, Hastie and Tibshirani (1993) studied a varying-coefficient model of the form:

$$\lambda_{ik}\big(t^* \mid \mathbf{Z}_{ik}\big) = \lambda_{0g}(t^*) \exp\big\{\boldsymbol{\beta}(t^*)^{\mathrm{T}} \mathbf{Z}_{ik}\big\}. \tag{5}$$

A parametric spline basis can be adopted to characterize $\boldsymbol{\beta}(t^*)$, or nonparametric smoothing methods can be used. Here, we modify local linear estimation described by Cai and Sun (2003) for use in the partly conditional setting. The idea of local linear estimation is that for a neighborhood around each time point $t^*, u \in \mathcal{N}(t^*, \epsilon)$, by Taylor series approximation, we have

$$\boldsymbol{\beta}(u) \approx \boldsymbol{\beta}(t^*) + \boldsymbol{\beta}'(t^*)(u - t^*).$$

Based on a local "working-independence" partial likelihood function, we can estimate $\beta(t^*)$ using a weighted estimating equation:

$$\sum_i^n \sum_k^K \int_0^\tau K_h(u - t^*)\big\{\tilde{\mathbf{Z}}_{ik}(1, u - t^*) - \bar{\mathbf{Z}}(u)\big\}\, dN_{ik}(u) = 0 \tag{6}$$

where $K(\cdot)$ is a kernel function with bounded support on $[-1,\, 1]$, h is the bandwidth, $K_h(x) = K(x/h)/h$, and $\tilde{\mathbf{Z}}_{ik}(1, u - t^*) = \mathbf{Z}_{ik} \otimes (1,\, u - t^*)$ with $\otimes$ denoting the Kronecker product. Under the stratified baseline model, $\bar{\mathbf{Z}}(u) = \sum_{g=1}^G \mathcal{S}_g^{(1)}\{\boldsymbol{\beta}(t^*), u\} 1(s_{ik} \in \mathcal{I}_g) / \sum_{g=1}^G \mathcal{S}_g^{(0)}\{\boldsymbol{\beta}(t^*), u\} 1(s_{ik} \in \mathcal{I}_g)$, with

$$\mathcal{S}_g^{(j)}\{\beta(t^*), u\} = \frac{1}{n} \sum_{l=1}^n \sum_{m=1}^K R_{lm}(u) \exp\big\{\boldsymbol{b}(t^*)^{\mathrm{T}} \tilde{Z}_{lm}(1, u - t^*)\big\} \times \tilde{Z}_{lm}(1, u - t^*)^{\otimes j} \cdot 1(s_{lm} \in \mathcal{I}_g),$$

where $\boldsymbol{b}(t^*) = \{\boldsymbol{b}_0(t^*),\, \boldsymbol{b}_1(t^*)\} = \{\boldsymbol{\beta}(t^*),\, \boldsymbol{\beta}'(t^*)\}$. The coefficient function $\boldsymbol{\beta}(t^*)$ is then estimated for each t^* using $\hat{\boldsymbol{\beta}}(t^*) = \hat{\boldsymbol{b}}_0(t^*)$.

Cai and Sun (2003) show the pointwise consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}(t^*)$ in the univariate case. We modify their results for the partly conditional setting. The consistency of $\hat{\boldsymbol{\beta}}(t^*)$ in the multivariate setting can be established in the same way as in the partly conditional proportional hazards model discussed in the previous section and thus the proof is omitted here. By imposing a stronger condition, we can establish the uniform consistency of $\boldsymbol{\beta}(t^*)$. The result is useful for the derivation of the large sample distribution for the survival functions and for estimators in the partly conditional survival models presented in the next section. Furthermore, let $\Delta(t^*, h, n) = \frac{h^2 \mu_2}{2} \boldsymbol{\beta}''(t^*) + o_p(h^2)$ denote the finite sample bias, and $\mu_2 = \int u^2 K_1(u) du$, it can be shown that $(nh)^{1/2}\{\hat{\boldsymbol{\beta}}(t^*) - \boldsymbol{\beta}(t^*) - \Delta(t^*, h, n)\}$ is asymptotically normal, with covariance structures which can be calculated explicitly using a sandwich-type estimator. Formal statements and proofs of these properties are given in an unpublished technical report that can be accessed at `http://www.bepress.com/uwbiostat/paper221/`. These asymptotic properties allow computation of confidence intervals and permit data-driven bandwidth selection.

Following Cai and Sun (2003), the theoretical optimal bandwidth can be obtained by minimizing the asymptotic weighted mean integrated squared error. In the multivariate situation, this quantity depends on the robust variance estimator and the second derivative of the coefficient function at point t^*, which are unknown in advance. Data-dependent procedures for selecting the optimal bandwidth have been suggested in the literature for nonparametric function estimation (Hall and Carroll, 1989; Eubank and Speckman, 1993; Ducharme et al., 1995). Further research on adapting data-driven automatic procedures to the multivariate survival setting is warranted, and in practice a sensitivity analysis is suggested.

3.3 *Combined Proportional and Varying-Coefficient Hazards Model*

Finally, we consider a class of models that accommodates the time-varying effect of a longitudinal marker in addition to other covariates whose effects are assumed independent of time. For presentation in this section, we only consider identical baseline hazard models. These models are essentially partly parametric hazard models as described by McKeague and Sasieni (1994) and have the following form for the hazard function:

$$\lambda_{ik}\big(t^* \,\big|\, \mathbf{Z}_{ik}, 0 \le s_{ik} \le T_i\big) = \lambda_0(t^*) \exp\big[\boldsymbol{\alpha}_1^{\mathrm{T}} \mathbf{Z}_{i\cdot a} + \boldsymbol{\alpha}_2^{\mathrm{T}} \{f_j(s_{ik})\}_{j=1}^p + \beta(t^*) Z_{ikb}\big], \quad t^* > 0. \tag{7}$$

To estimate the parameters under this model, we use two sets of estimating equations. Let $\boldsymbol{\theta} = \{\boldsymbol{\alpha},\, \beta(t^*)\}$, where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1,\, \boldsymbol{\alpha}_2)$ is a vector of time-invariant coefficients for $\mathbf{Z}_{ika} = \{\mathbf{Z}_{i\cdot a},\, f_1(s_{ik}), \ldots, f_p(s_{ik})\}$, and $\beta(t^*)$ is the time-varying coefficient for Z_{ikb}. In general, one can obtain $\hat{\boldsymbol{\theta}}$ by simultaneously solving the pair of estimating equations for $\boldsymbol{\alpha}$ and $\beta(t^*)$. A standard approach to solving these equations would involve backfitting (Hastie and Tibshirani, 1995) and therefore require iterative solution to the equations. A computationally simple "one-step" alternative can yield an asymptotically equivalent estimator. The alternative estimator is defined by the following steps:

(i) Fit a nonparametric Cox model with all the covariates $\mathbf{Z}_{ik}$. Here, the model is of a purely time-varying form: $\lambda_{ik}(t^* \mid \mathbf{Z}_{ik},\, 0 \le s_{ik} \le T_i) = \lambda_0(t^*) \exp\{\boldsymbol{\alpha}^{\mathrm{T}}(t^*) \mathbf{Z}_{ika} + \beta(t^*) Z_{ikb}\}$. The resulting estimator $\tilde{\beta}(t^*)$ is consistent for $\beta(t^*)$ (see Zheng, 2002 for details).

(ii) Fit a parametric Cox model with $\mathbf{Z}_{ika}$ as covariate, using a time-dependent offset $\tilde{\beta}(t^*) Z_{ikb}$. The estimator $\hat{\boldsymbol{\alpha}}$ obtained from this model is consistent for $\boldsymbol{\alpha}$ (see Zheng, 2002 for details).

(iii) Fit a time-varying coefficient Cox model with Z_{ikb}, using offset $\hat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{Z}_{ika}$. The new estimate $\hat{\beta}(t^*)$ is the final estimate for $\beta(t^*)$.

3.4 *Predictive Survival Functions with a Partly Conditional Survival Model*

Here, we present an estimation procedure for predicting the survival function for patients with a marker measurement z_b obtained at a certain specific measurement time s, in addition to a vector of time-invariant covariates $\mathbf{z}_a$. Let $P(T_i > t^* + s \,|\, \mathbf{Z}_i = \mathbf{z}_0,\ s) = S(t^* \,|\, \mathbf{Z}_i = \mathbf{z}_0,\ s) = \exp\{-\Lambda(\mathbf{z}_0,\ s)\}$. Here, $\Lambda_0(t^*)$ can be estimated by the natural Breslow-type estimator: $\hat{\Lambda}_0(t^*) = \int_0^{t^*} \frac{1}{\hat{J}^*_{n,0}(u)} d\bar{N}(u)$, where

$$\hat{J}^*_{n,0}(t^*) = \sum_{i=1}^{n}\sum_{k=1}^{K} R_{ik}(t^*) \times \exp\left[\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{Z}_{i\cdot a} + \boldsymbol{\alpha}_2^{\mathrm{T}}\{f_j(s_{ik})\}_{j=1}^{p} + \beta(t^*)Z_{ikb}\right]$$

then

$$\hat{S}\big(t^* \,|\, \mathbf{Z}_i = \mathbf{z}_0, s\big) = \exp\left[-\int_0^{t^*} \exp\left[\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{z}_a + \boldsymbol{\alpha}_2^{\mathrm{T}}\{f_j(s)\}_{j=1}^{p} + \beta(t^*)z_b\right] d\hat{\Lambda}_0(u)\right].$$

4. Simulations

4.1 *Simulations for Regression Function*

We first investigate the performance of the partly conditional regression estimates with respect to bias, coverage, and efficiency. Estimation using the derived survival data requires careful attention to mechanisms that lead to unbalanced cluster sizes (i.e., death and censoring). One goal of the study is to empirically demonstrate that the proposed methods do indeed lead to asymptotically unbiased estimation when the number of contributing observations per cluster is stochastic.

Our partly conditional model does not completely specify the joint distribution for the event time and the longitudinal marker process. In the related multivariate survival literature, it has been noted that it is surprisingly difficult to construct joint distributions that satisfy the marginal proportional hazards assumptions (Wei et al., 1989; Yang and Ying, 2001). We show that it is possible to construct a valid joint distribution where derived survival times simultaneously satisfy the partly conditional hazards assumption. Consider data with a single binary treatment group indicator $Z_{i\cdot a}$, and a single longitudinal marker $Z_i(s_k)$ measured at a common set of times s_1, $s_2, \ldots, s_K$. To generate data $[T_i,\ Z_{i\cdot a},\ \mathbf{Z}_{ib} = \mathrm{vec}\{Z_i(s_k)\}]$, we first create $Z_{ib0} = b_i + \sum_{j=1}^{K}\log(V_{ij})/\gamma_2$, where $b_i \sim \mathcal{N}(\mu, \sigma^2)$ and $V_{ij} \sim P(\rho)$, independent positive stable random variables with index ρ (Hougaard, 1986). A failure time T_i is then generated with a hazardship $\lambda_i(t) = \lambda_0(t)\exp\{(\gamma_1 Z_{i\cdot a} + \gamma_2 Z_{ib0})\}$, with $\lambda_0(t) = a$. Let $Z_i(s_k) = Z_{ib0} - \log(V_{ik})/\gamma_2$. This creates a form of exchangeable marker measurements. For a partly conditional model, we only include for analysis those $Z_i(s_k)$ with $s_k < T_i$. Based on properties of positive stable random variables the procedure leads to partly conditional hazards of the form $\lambda\{t^* \,|\, Z_{i\cdot a},\ Z_i(s_k),\ T_i > s_k\} = \lambda_0(t^* + s_k)\,\rho\{\Lambda_0(t^* + s_k)\}^{(\rho-1)}\exp\{\rho\gamma_1 Z_{i\cdot a} + \rho\gamma_2 Z_i(s_k)\}$. Using this construction the hazard for $T_{ik} = T_i - s_k$ will generally depend on s_K and therefore stratified models similar to those considered by Wei et al. (1989) would be appropriate. However, if we choose $\Lambda_0(t) = (t/a)^{1/\rho}$, then $\lambda_0(t + s_k)\times \rho\{\Lambda_0(t + s_k)\}^{(\rho-1)} = 1/a$, and thus a common baseline hazard is obtained. By varying ρ and σ we can create marker measurements with differing amounts of within- and between-person variation. For similarity to the MACS CD4 data analyzed in Section 5, we used $\mu = 600$ and $\sigma = 30$. For the induced partly conditional hazards model, the regression coefficients are $\alpha = \rho \cdot \gamma_1$ and $\beta = \rho \cdot \gamma_2$.

We evaluate samples with $n = 100$, 500, 1000 clusters using a uniform censoring distribution to obtain approximately 0%, 25%, 50%, and 75% censoring. We consider a study where the markers can be measured up to 10 times per subject. For all our simulated situations, 500 Monte Carlo data sets are used. We present results using $\rho = 0.6$, $\alpha = -2$, and $\beta = -0.02$ (note: We use this coefficient value [scale] because we create simulations to approximate analysis of CD4 where the marker ranges from less than 200 to greater than 1000).

With clusters of size 10, the average number of observations per subject ranges from 1.7 to 6.1, depending on the censoring level. Table 1 presents the simulation results for a partly conditional model using a single common baseline hazard. The relative bias is less than 4.3% for all situations, and tends to decrease with increasing sample size. Coverage probabilities using the robust variance estimator are very close to the nominal 95% level, whereas coverage probabilities from a naive independence model are generally below the nominal level.

We also use simulation studies to assess the performance of local linear estimation for a time-varying coefficient partly conditional Cox model. We simulated data according to the scenario described above, and use a local linear Cox model with Epanechnikov kernel. We investigate the performance at two different time points, namely, $t^* = 25$ and $t^* = 75$. At each time point, we choose bandwidth h such that 30% of the data points are included in the local linear estimation. As a result, at $t^* = 25$, the average bandwidth h is about 15, whereas at $t^* = 75$, the average bandwidth h is about 30. The actual bandwidth may vary slightly from sample to sample. The results presented in Table 2 indicate that estimates can be obtained with small bias using the partly conditional approach. We find -1.17% bias at $t^* = 25$ and -3.06% bias at $t^* = 75$. In addition, using robust standard errors to create 95% pointwise confidence intervals yielded empirical coverage rates of 93.2% and 95.2% at $t^* = 25$, 75, respectively.

Finally, we also compare the multivariate approach with a valid univariate procedure that randomly chooses a single marker measurement for analysis. In Table 2, we display the bias and the standard deviation for $\hat{\beta}(t^*)$ at $t^* = 25$ and $t^* = 75$. The estimates based on the partly conditional model have a variance that is 1/2.44 times the variance of the univariate estimator for $t^* = 25$ and 1/3.43 times smaller for $t^* = 75$, demonstrating the potential gain in efficiency through use of all marker measurements.

4.2 *Simulations for Survival Function*

We also conduct simulations to investigate the estimation of partly conditional survival functions.

Table 1

Simulation results for identical baseline models, $K = 10$. The partly conditional hazard model is $\lambda_0(t^*)\exp\{\alpha Z_{i\cdot a} + \beta Z_i(s_{ik})\}$.

			$\alpha = -2$				$\beta = -0.02$			
n	% censored	AN	RBias $\times 10^{-2}$	SE $\times 10^{-3}$	CP(N)	CP(R)	RBias $\times 10^{-2}$	SE $\times 10^{-5}$	CP(N)	CP(R)
100	0	4.2	1.63	11.31	0.656	0.906	0.13	3.55	0.838	0.942
100	25	3.7	2.18	13.95	0.690	0.924	0.87	4.22	0.888	0.958
100	50	1.5	1.47	14.59	0.940	0.932	2.07	5.95	0.952	0.940
100	75	1.0	2.08	22.39	0.958	0.914	4.21	9.45	0.954	0.924
200	0	4.2	0.63	7.17	0.678	0.930	0.05	2.67	0.794	0.936
200	25	3.7	0.26	9.16	0.724	0.944	0.54	2.97	0.858	0.940
200	50	1.5	0.26	10.58	0.918	0.926	0.68	3.99	0.936	0.930
200	75	1.0	0.61	15.13	0.940	0.934	2.01	5.78	0.956	0.930
500	0	4.2	0.33	4.54	0.652	0.928	0.09	1.53	0.818	0.966
500	25	3.7	0.20	6.27	0.698	0.922	0.07	1.92	0.836	0.932
500	50	1.5	0.47	6.19	0.930	0.948	0.33	2.48	0.934	0.932
500	75	1.0	0.18	9.30	0.940	0.934	0.84	3.48	0.946	0.924
1000	0	4.2	0.20	2.96	0.700	0.956	0.08	1.19	0.782	0.930
1000	25	3.7	0.04	3.83	0.712	0.946	0.01	1.42	0.838	0.944
1000	50	1.5	0.66	4.30	0.942	0.960	0.08	1.63	0.952	0.956
1000	75	1	0.36	6.27	0.960	0.962	0.23	2.33	0.948	0.942

Note: AN is the average number of measurements per subject that is used in the partly conditional model. RBias (relative bias) is the sampling mean of the ratio $|\hat{\beta} - \beta_0|/\beta_0$. SE is the sampling mean of the robust standard error estimator for $\hat{\beta}$. CP(N) and CP(R) are the coverage probabilities of the 95% confidence intervals corresponding to the naive and robust variance estimates.

In particular, we generate data using a standard joint model and then use partly conditional methods to directly provide estimates of conditional survival probabilities. In these simulations, we evaluate the ability of the proposed methods to approximate the conditional probabilities induced by a standard joint model. For simplicity, we consider only one marker $\mathbf{Z}_i$ that is measured longitudinally and now focus on estimation of the partly conditional survival probability $P\{T_i > t \,|\, Z_i(s), t > s\}$. The data generation follows the popular approach (e.g., Tsiatis and Davidian, 2001) that assumes longitudinal data follow a linear mixed effects model with measurement errors and that survival depends on the covariates through a proportional hazards relationship with the underlying random effects. The specific parameters studied are based on the MACS data. Specifically, $\{Z_i(s) \,|\, \alpha_i\}$ is generated as $Z_i(s) = W_i(s) + \epsilon_i(s)$, where $W_i(s) = \alpha_{0i} + \alpha_{1i} f(s)$ and $f(s) = \log(s/30)$. The random effects $(\alpha_{0i}, \alpha_{1i})$ are generated as a bivariate normal with mean $(\alpha_0, \alpha_1)^{\mathrm{T}} = (0.6, -0.1)^{\mathrm{T}}$ and covariance matrix $\Sigma_\alpha = \begin{bmatrix} 0.83^2 & -0.005 \\ -0.005 & 0.13^2 \end{bmatrix}$. The measurement error $\epsilon_i(s)$ is taken to be independent and identically distributed normally with mean 0 and standard deviation $\sigma_\epsilon = 0.1$. Given $W_i(s)$, the failure time T_i has a conditional hazard relationship with $\lambda_i(u) = \lambda_0(u)\exp\{\beta^* W_i(u)\}$, where $\beta^* = -1.5$. The baseline hazard is $\lambda_0(t) = 1/30$. Censoring was generated from a uniform distribution with mean 300, leading to approximately 75% censoring. For each subject, we assume the marker Z_i is measured 10 times and consider two scenarios for the measurement times s_{ik}, for $k = 1, \ldots, 10$. In the first scenario, the marker is measured at a fixed interval of 6 months, i.e., with $s_{ik} = 6, 12, 18, \ldots, 60$ months. In the second scenario, we consider a marker that is measured irregularly, with $s_{ik} = 6 \times k + 6 \times e_{ik}$, where e_{ik} is the standard uniform random variate. We carry out simulations for both scenarios with $n = 200$ and $n = 500$.

For a partly conditional model, we assume that only those $Z_i(s_{ik})$ with $s_{ik} < T_i$ are available for analysis. The average number of observations per subject is 3.5. For the first scenario, because the longitudinal marker $Z_i(s_{ik})$ is measured at a common set of times $s_1, s_2, \ldots, s_{10}$, we consider a partly conditional model for T_{ik} and $t^* = t - s$ of the form: $\lambda_i(t^* \,|\, Z_{ik}, 0 \le s_{ik} \le t) = \lambda_{0k}(t^*)\exp\{\beta Z_i(s_{ik})\}$, allowing the unspecified

Table 2

Simulation results for $\beta(t^*) = -0.02$ with identical baseline models. $n = 200$, $K = 10$. The partly conditional hazard model is $\lambda_0(t^*)\exp\{\alpha Z_{i\cdot a} + \beta(t^*) \cdot Z_i(s_{ik})\}$.

	Univariate model				Partly conditional model				
t	RBias $\times 10^{-2}$	$\mathrm{SE}_{\mathrm{emp}} \times 10^{-3}$	$\mathrm{SE}_{\mathrm{est}} \times 10^{-3}$	CP	RBias $\times 10^{-2}$	$\mathrm{SE}_{\mathrm{emp}} \times 10^{-3}$	$\mathrm{SE}_{\mathrm{est}} \times 10^{-3}$	CP(R)	EF
25	−3.525	2.247	2.547	0.938	−1.170	1.630	1.427	0.932	2.442
75	8.035	4.681	3.801	0.938	−3.055	2.203	2.052	0.952	3.432

Note: RBias (relative bias) is the sampling mean of the ratio $|\hat{\beta}(t^*) - \beta_0(t^*)|/\beta_0(t^*)$. $\mathrm{SE}_{\mathrm{est}}$ is the mean of the standard error estimates; $\mathrm{SE}_{\mathrm{emp}}$ is the standard error of the estimates of β; EF is the relative efficiency of multivariate model versus univariate model $(\hat{\sigma}^2_{\mathrm{univ}}/\hat{\sigma}^2_{\mathrm{mult}})$.

Table 3
Simulation results for $P\{T_i > t \mid Z_i(s),\ s < T\}$ from model with $\lambda_i(t^*) = \lambda_{0s}(t^*)\exp\{\beta Z_i(s_{ik})\}$. s measured at regular interval. $n = 200$.

	$s = 6$				$s = 24$			
	$Z = 1$		$Z = 2$		$Z = 1$		$Z = 2$	
	$t^* = 12$	$t^* = 36$	$t^* = 12$	$t^* = 36$	$t^* = 12$	$t^* = 36$	$t^* = 12$	$t^* = 36$
$S(t)$	0.754	0.431	0.937	0.824	0.796	0.506	0.949	0.854
$\hat{S}(t)$	0.753	0.435	0.932	0.814	0.797	0.511	0.945	0.847
SE_{est}	0.039	0.051	0.022	0.048	0.047	0.065	0.019	0.042
SE_{emp}	0.040	0.050	0.023	0.049	0.047	0.064	0.020	0.044
RBias	0.042	0.091	0.02	0.048	0.047	0.102	0.016	0.040

Note: $S(t)$ is the true partly conditional survival probability. $\hat{S}(t)$ is the estimated survival probability from the PCS model. SE_{est} is the sampling mean of the robust standard error estimator for $\hat{S}(t)$. SE_{emp} is the standard error of the estimates of $\hat{S}(t)$. RBias (relative bias) is the sampling mean of the ratio $|\hat{\beta} - \beta_0|/\beta_0$.

baseline hazard function to depend on s_k. We then estimate the partly conditional survival probability $P\{T_i > t \mid Z_i(s) = z,\ s < t\}$ with $\hat{S}(t \mid s, z) = \exp\{-\hat{\Lambda}_{0k}(t)\exp(\hat{\beta}z)\}$.

For the second scenario, because the measurement times do not necessarily occur at a common set of times, we consider a different form of the partly conditional model for T_{ik} and $t^* = t - s$: $\lambda_i(t^* \mid Z_{ik},\ 0 \le s_{ik} \le T_i) = \lambda_0(t^*)\times \exp\{\beta Z_i(s_{ik}) + \gamma B(s_{ik})\}$, where $B(s_{ik})$ denotes a spline function of s_{ik}. That is, rather than stratifying the baseline hazard function by the measurement time s, we now smoothly capture the effect of s in the regression part of the model. We estimate the partly conditional survival function with $\hat{S}(t \mid s, z) = \exp[-\hat{\Lambda}_0(t)\exp\{\hat{\beta}z + \hat{\gamma}B(s)\}]$. In both scenarios, the baseline cumulative hazard $\Lambda_{0k}(t)$ or $\Lambda_0(t)$ is estimated with the Breslow-type estimator. A robust variance is calculated for $\hat{S}(t \mid s, z)$.

With the joint distribution of T_i and $Z_i(s_{ik})$ specified using a hierarchical model, the analytic form of the true partly conditional survival function is not readily derived. Therefore, we used numerical methods to compute the target probabilities. This simulation study evaluates the ability of flexible directly specified models to estimate conditional survival probabilities.

Table 3 presents the results for simulations from the first scenario with a sample size of $n = 200$. We consider both a subject with a low marker value ($z = 1$), and a subject with a high value ($z = 2$). We estimate their residual survival probabilities at 12 months and 36 months post-measurement with measurement times of 6 and 24 months, respectively. In all cases, the estimated survival probabilities are quite close to their theoretical counterparts, with the average relative bias <5% in all except one of the cases. The bias tends to decrease as the sample size increases to $n = 500$ (results not shown). For the second scenario when the marker is simulated with an irregular measurement interval, we use a partly conditional model that specifies s parametrically using a natural cubic spline basis function $B(s)$ with a single knot at $s = 24$. Similar to the first scenario, we again observe very small bias in all cases (see Table 4). In summary, these results indicate that accurate point estimates can be obtained by directly using our partly conditional approach.

5. Example

Here, we apply the partly conditional survival model to data from the MACS, which was reported in detail by Kaslow et al. (1987). Of the 5622 homosexual/bisexual men enrolled, 3426 were seronegative at baseline and 479 of these became seropositive between 1984 and 1996. Because we focus on the relationship between T-cell levels and AIDS diagnosis, we adopt the 1987 CDC definition of AIDS, which relies on symptoms rather than CD4 lymphocyte counts to define AIDS. Under this definition, 211 seroconverters developed AIDS during the study period. The mean time from seroconversion to the onset of AIDS among these subjects with observed times is 72 months (SD = 28 months; median = 71 months). The present analysis uses data from the 438 seroconverters with dates of seroconversion known to within ±4.5 months. These

Table 4
Simulation results for $P\{T_i > t \mid Z_i(s),\ s < T\}$ in model with $\lambda_i(t) = \lambda_0(t)\exp\{\beta Z_i(s_{ik}) + \gamma B(s_{ik})\}$. s measured at irregular interval. $n = 200$.

	$s = 6$				$s = 24$			
	$Z = 1$		$Z = 2$		$Z = 1$		$Z = 2$	
	$t^* = 12$	$t^* = 36$	$t^* = 12$	$t^* = 36$	$t^* = 12$	$t^* = 36$	$t^* = 12$	$t^* = 36$
$S(t)$	0.754	0.431	0.937	0.824	0.796	0.506	0.949	0.854
$\hat{S}(t)$	0.762	0.446	0.937	0.825	0.792	0.498	0.946	0.848
SE_{est}	0.03	0.043	0.019	0.044	0.021	0.029	0.015	0.035
SE_{emp}	0.034	0.049	0.021	0.049	0.027	0.054	0.017	0.042
RBias	0.038	0.094	0.018	0.047	0.027	0.086	0.014	0.039

subjects have an average of 13 measurements per person ($N = 3807$ total observations).

The objective of the present analysis is to investigate the relationship between a biomarker such as CD4 count and the risk of AIDS. We define the measurement time (s) as the time from seroconversion to the time that the CD4 count is recorded, and survival time as the time from CD4 count measurement to the AIDS diagnosis time ($t^* = t - s$). We seek to quantify the predictive value of serially measured CD4 counts. In all analyses, we use $CD4(s)$ equal to the raw CD4 count divided by 300 (approximate standard deviation).

5.1 *Partly Conditional Regression Function Estimation*

We start by investigating the simplest model that assumes a common baseline hazard but allows a time-varying coefficient (model 1): $\lambda_{ik}(t^*) = \lambda_0(t^*)\exp\{\beta(t^*)CD4_i(s_{ik})\}$. Subsequent analysis will relax this model to allow dependence on the measurement time s. For estimation, we use all CD4 measurements after seroconversion and before AIDS diagnosis as the time-varying predictor, $CD4(s)$. Estimates of the function $\beta(t^*)$ are obtained by fitting a partly conditional Cox model using local linear estimation. We use the Epanechnikov kernel $K(u) = 0.75(1 - U^2)_+$ with a bandwidth of 30. The bandwidth is selected to ensure that we have substantial data available at each data point for stable estimation. We also consider other values of the bandwidth to assess the sensitivity of the results to this choice. Eubank and Speckman (1993) suggest use of an undersmoothed bandwidth h, i.e., one that satisfies $n^{1/3}h \to 0$, so that the inherent bias (given as $\Delta(t^*, h, n)$ in Section 3.2) is negligible asymptotically. The function $\beta(t^*)$ is estimated at the grid points $t^* = 4 \times j$ months, $j = 1, \ldots, 32$. In Figure 1, the estimate of $\beta(t^*)$ shows a strong time trend, with diminishing relative risk as t^* increases. For example, at any time $0 < s < T$, for two individuals whose CD4 differ by 300, the log relative hazard is the highest immediately after the measurement is taken ($t^* = 1$), with $\hat{\beta}(1) \approx -3$, and then attenuates steadily over the next 60 months to $\hat{\beta}(60) \approx -1$. Finally, the predictive capacity of CD4 wanes to nearly 0 when the measurement is more than 60 months old.

To explore the potential gain in efficiency that arises from using all of the longitudinal data, we compare results from estimation using only a single randomly selected CD4 measurement per individual. Figure 1 also shows the local linear estimates using these 438 independent observations. We find similar point estimates yet narrower confidence intervals when all 3807 longitudinal measurements are used to estimate $\beta(t^*)$. As suggested by our simulation studies, estimation based on all the available longitudinal data is apparently more efficient.

Next we investigate whether the effect of CD4 depends on the measurement time s by adopting a model where both the baseline hazard and the form of the coefficient function, $\beta(t^*)$ may depend on s (model 2). We create three groups of data $\{X_{ik}, \Delta_{ik}, Z_i(s_{ik}), s_{ik}\}$ based on the CD4 measurement time, s_{ik}: group 1 is comprised of CD4 measurements within the first year after seroconversion ($0 \le s \le 12$, $n = 392$, and $N = 686$); group 2 is comprised of measurements between the second and the third year ($24 < s \le 36$, $n = 321$, and $N = 577$); and group 3 contains data from between the fourth and fifth year post-seroconversion ($48 < s \le 60$, $n = 229$, and $N = 396$). Figure 2 shows the coefficient functions estimated separately for these three groups. It appears that CD4 measured earlier after seroconversion loses its predictive power more rapidly than CD4 measured at later times. For example, for CD4 observed within the first year the log relative hazard at $t^* = 4$ months is $\hat{\beta}(t^* = 4) \approx -3$, and then

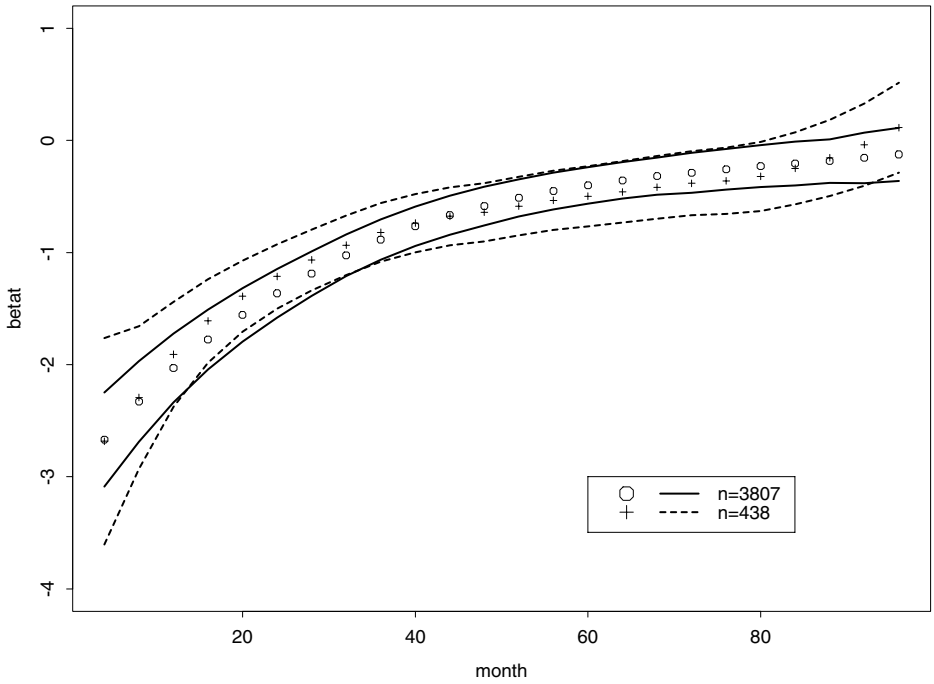
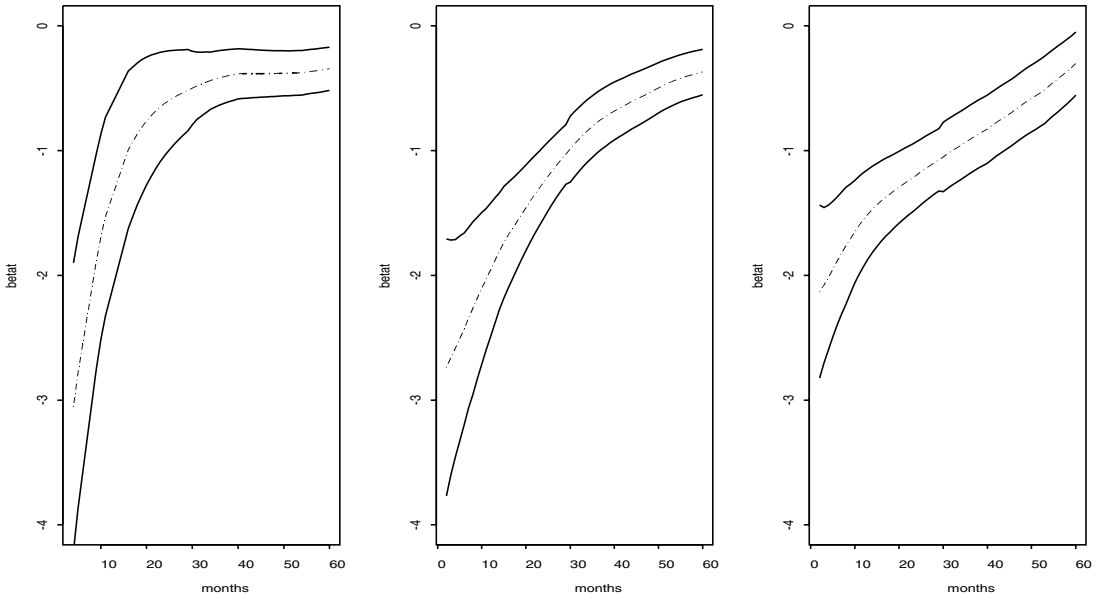
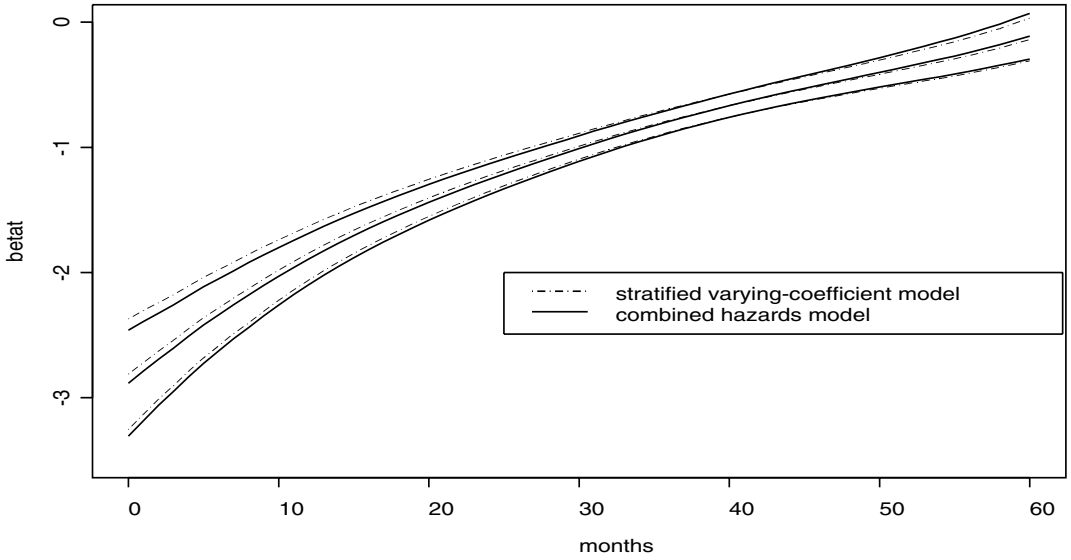


Figure 1. Coefficient functions and pointwise 95% confidence intervals for the time-varying coefficient of standardized CD4 cell counts.



(a)



(b)

Figure 2. Coefficient functions for standardized CD4 cell counts using different partly conditional survival models. (a) Coefficient functions and pointwise 95% confidence intervals for standardized CD4 cell counts measured for $s \leq 12$ months, $12 < s \leq 36$ months, and $36 < s \leq 60$ months following seroconversion based on separate partly conditional varying-coefficient hazards models. (b) Coefficient functions and pointwise 95% confidence intervals for standardized CD4 cell counts based on a varying-coefficient and a combined proportional and varying-coefficient hazards model.

quickly decays to $\hat{\beta}(t^*) = -1$ by 20 months. In contrast, for measurements taken in the fifth year after conversion we find somewhat weaker short-term association, $\hat{\beta}(t^* = 4) \approx -2$, but a longer follow-up time is required before the log relative hazard decays to -1, with the point estimate crossing -1 after 30 months of follow-up. Essentially, this analysis approach estimates the coefficient function $\beta(t^*, s)$ in the model $\lambda_0(t^*, s) \times \exp\{\beta(t^*, s) CD4(s)\}$. If a parametric form for $\beta(t^*, s)$ were adopted, the simplifying assumption $\beta(t^*, s) = \beta(t^*)$ used in model 1 could be formally tested.

We also consider two intermediate models that differ in the way they model the measurement time s while retaining a common estimate for $\beta(t^*)$. Model 3 assumes different, but unspecified, baseline functions for CD4 measured at different follow-up times. Because subjects in MACS are followed semiannually, we divide observations into $G = 11$ strata. An

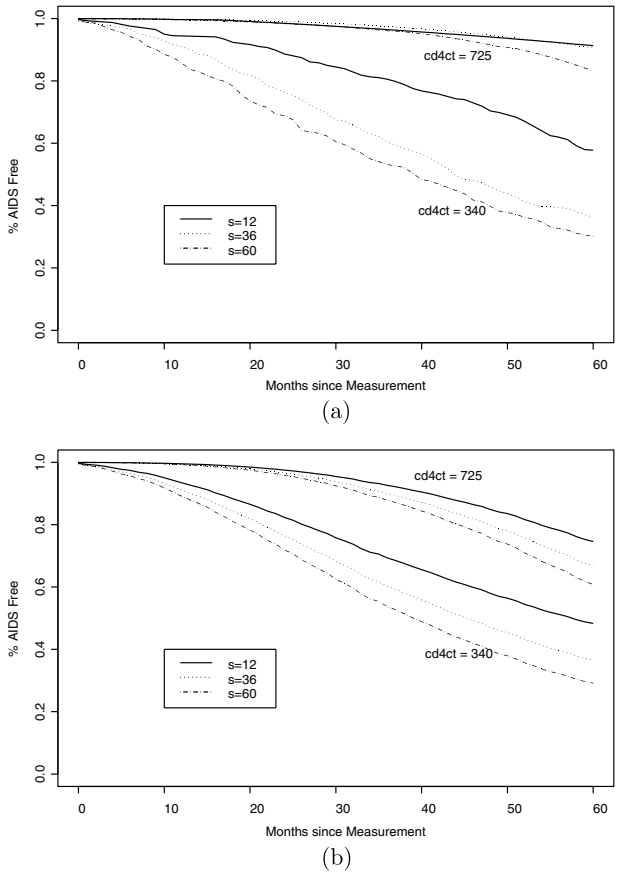


Figure 3. Partly conditional survival functions for hypothetical subjects with CD4 cell counts of 725 and 340 at the time of measurement. (a) Survival estimates based on separate varying-coefficient hazards models fit to $s \leq 12$ months, $12 < s \leq 36$ months, and $36 < s \leq 60$ months. (b) Survival estimates based on a combined proportional varying-coefficient hazards model.

observation belongs to the gth stratum if it is measured within the gth year since seroconversion. Specifically, the model takes the form $\lambda_{ik}(t^*) = \lambda_{0g}(t^*) \exp\{\beta(t^*) CD4_i(s_{ik})\}$. Alternatively, model 4 uses both the CD4 count and the measurement time as covariates in a partly parametric Cox model, namely, $\lambda_{ik}(t^*) = \lambda_0(t^*) \exp[\boldsymbol{\alpha}^{\mathrm{T}}\{f_j(s)\}_{j=1}^{p} + \beta(t^*) CD4_i(s_{ik})]$. We specify a flexible parametric model for measurement time, s, using natural cubic spline basis functions, $f_j(s)$, with a single knot at $s = 48$. We also conduct analyses with a single knot at $s = 36$ and with a pair of knots at $s = 16$, 36, respectively. These alternative choices for $f_j(s)$ result in very minor differences in the estimated coefficient function $\hat{\beta}(t^*)$. Because later follow-up times tend to have shorter survival time due to administrative censoring, in order to obtain stable estimates, we restrict the analysis to a time interval of $0 \leq t - s \leq 60$. Figure 2 shows the estimated coefficient functions for CD4 along with 95% pointwise confidence intervals from both the stratified and the partly parametric models. The two methods give very similar estimated coefficient functions.

5.2 *Partly Conditional Survival Function Estimation*

One important objective of our partly conditional survival model is to estimate the updated survival probability $P\{T_i > t \,|\, CD4_i(s),\ T_i > s\}$ for an arbitrary pair of survival and measurement times, $(s,\ t)$, where $s < t$. Figure 3 shows the predictive survival probabilities given in terms of months since measurement, $P\{T_i > t^* + s \,|\, CD4_i(s),\ T_i > s\}$, based on models 2 and 4 described above. Recall that in our application the survival probability is equivalent to the probability of being free of AIDS. Figure 3 illustrates the estimated probabilities for two hypothetical individuals: one with a high CD4 value (725) and one with a low CD4 value (340). We also consider measurement times of 1-year post-seroconversion, $s = 12$, and the third, $s = 36$, and fifth year, $s = 60$, after seroconversion. In Figure 3a, we use model 2 which allows the baseline hazard and the coefficient function to depend on the measurement time s. We see that for both individuals, the chance of being free of AIDS decreases steadily with time, but the individual with a higher CD4 value is less likely to develop AIDS during the follow-up period. Furthermore, the predictive survival functions appear to depend on the time at which CD4 count is measured. For example, an individual with a CD4 value of 340 measured at the first year after seroconversion has a chance of developing AIDS within 4 years of approximately 30%, whereas if the same value of CD4 is obtained at 3 years post-seroconversion, then his chance of getting AIDS within the next 4 years is approximately 60%.

To assess the sensitivity of survival predictions to the choice of model, we also display estimates based on a more structured model that assumes a common coefficient function, and uses the measurement time s as a covariate (model 4). In Figure 3b, we find estimated survival probabilities that are qualitatively similar to those from the less structured model, but specific estimates differ, particularly for longer follow-up times. For example, the effects of measurement time s on the high and the low CD4 curves become less evident compared with panel (a).

If the ultimate goal is to create accurate predictions then the tradeoff between a potentially less biased but more variable approach (model 2) and a less variable but potentially biased approach (model 4) may be empirically evaluated using cross-validation methods if a meaningful measure of the discrepancy between data and prediction can be adopted. Unfortunately, there is no well-accepted summary of predictive model accuracy for survival models, although alternatives have been proposed (Heagerty, Lumley, and Pepe, 2000; Schemper and Henderson, 2000). In addition, empirical evaluation of accuracy for varying-coefficient models would be computationally demanding.

6. Discussion

We have proposed a new approach that can quantify the risk of a key clinical event at time t as a function of the marker process accumulated through time s, for any pair of times $(s,\ t)$ with $s < t$. In contrast to the standard time-varying covariate regression model for the event time, our method decouples the time scale for modeling the hazard from the time scale for accrual of available longitudinal covariate information, and thus directly facilitates the calculation of quantities

such as $P\{T_i > t \,|\, Z_i(s),\ 0 \le s < t\}$ without assumptions regarding the longitudinal marker distribution.

One important feature of our partly conditional model is that we allow regression parameters to depend on both the time of measurement for the predictor and the time of measurement for the outcome. In specific applications, a varying-coefficient model of the form $\beta(t,\ s) = \beta(t - s)$ may be used which assumes that the association between the survival outcome and the covariate depends only on their time separation. For estimation, we extended local linear estimation for the univariate Cox model (Cai and Sun, 2003) to the partly conditional setting, and we provided detailed estimation procedures for three classes of partly conditional survival models.

One issue that arises with use of a partly conditional hazard model is the need to model the measurement time s. We have introduced alternative models that differ in the way the effect of the measurement time, s, is specified. In general, the choice of model may depend on the aim of the study and the specific features of the data structures such as the frequency and spacing of visits. Our example analysis explored the extent to which results were sensitive to model choice. Further research to develop methods for model checking and to define appropriate criteria for selecting models would be useful.

In this article, we make fairly strong assumptions regarding both measurement timing and marker missingness. In particular, we assume that each individual provides a sequence of measurements at either a set of fixed times, or at times that occur in a completely random fashion. However, in some observational studies, individuals are not necessarily followed at scheduled intervals. Furthermore, the timing at which individuals are measured may depend on the previous value of the marker measurement. In the repeated measures setting with outcome-dependent follow-up, it has been demonstrated that potential bias could be associated with the use of an estimating equation approach (Lipsitz et al., 2002). Further work is needed to evaluate the robustness of our estimation procedure to the timing assumptions, or to develop a more general method that can relax the measurement timing assumptions.

Instead of employing a likelihood-based estimation procedure, we develop nonparametric and semiparametric methods. One advantage of using a semiparametric approach is that it provides a computationally simple and robust solution. One potential weakness is that our methods are based on a working independence assumption, and as such, may be less efficient than a full-likelihood approach. Future work that compares our semiparametric approach with specific likelihood-based alternatives in terms of both efficiency and robustness would be valuable.

References

Altman, D. G. and De Stavola, B. L. (1994). Practical problems in fitting a proportional hazards model to data with updated measurements of the covariates. *Statistics in Medicine* **13,** 301–341.

Anderson, J., Cain, K., and Gelbber, R. (1983). Analysis of survival by tumor response. *Journal of Clinical Oncology* **1,** 710–719.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1995). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Cai, Z. and Sun, Y. (2003). Local linear estimation for time-dependent coefficients in Cox's regression models. *Scandinavian Journal of Statistics* **30,** 93–111.

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

Cox, D. R. and Oakes, D. O. (1984). *Analysis of Survival Data*. New York: Chapman and Hall.

Ducharme, G. R., Gannoun, A., Guertin, M.-C., and Jéquier, J.-C. (1995). Reference values obtained by kernel-based estimation of quantile regressions. *Biometrics* **51,** 1105–1116.

Etzioni, R., Pepe, M., Longton, G., Hu, C., and Goodman, G. (1999). Incorporating the time dimension in receiver operating characteristic curves: A case study of prostate cancer. *Medical Decision Making* **19,** 242–251.

Eubank, R. L. and Speckman, P. L. (1993). Confidence bands in nonparametric regression. *Journal of the American Statistical Association* **88,** 1287–1301.

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

Fusaro, R. E., Nielsen, J. P., and Scheike, T. H. (1993). Marker-dependent hazard estimation: An application to AIDS. *Statistics in Medicine* **12,** 843–865.

Hall, P. and Carroll, R. J. (1989). Variance function estimation in regression: The effect of estimating the mean. *Journal of the Royal Statistical Society, Series B* **51,** 3–14.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B* **55,** 757–796.

Hastie, T. and Tibshirani, R. (1995). Generalized additive models for medical research. *Statistical Methods in Medical Research* **4,** 187–196.

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56,** 337–344.

Henderson, R., Diggle, P., and Dobson, A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53,** 330–339.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions (Corr: V75 p395). *Biometrika* **73,** 387–396.

Jewell, N. P. and Kalbfleisch, J. D. (1996). Marker processes in survival analysis. *Lifetime Data Analysis* **2,** 15–29.

Jewell, N. P. and Nielsen, J. P. (1993). A framework for consistent prediction rules based on markers. *Biometrika* **80,** 153–164.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: John Wiley and Sons.

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., and Rinaldo, C. R. A. (1987). The multicentre AIDS cohort study: Rationale, organization, and selected

characteristics of the participants. *American Journal of Epidemiology* **1,** 310–318.

Lee, E. W., Wei, L. J., and Amato, D. A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations (Discussion: p247). In *Survival Analysis: State of the Art*, J. P. Klein and P. K. Goel (eds), 237–247. Dordrecht: Kluwer Academic Publishers.

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.

Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Richard, G., and Steven, L. (2002). Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics* **58,** 621–630.

McKeague, I. W. and Sasieni, P. D. (1994). A partly parametric additive risk model. *Biometrika* **81,** 501–514.

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

Pepe, M. S. and Couper, D. (1997). Modeling partly conditional means with longitudinal data. *Journal of the American Statistical Association* **92,** 991–998.

Pepe, M. S., Heagerty, P., and Whitaker, R. (1999). Prediction using partly conditional time-varying coefficients regression models. *Biometrics* **55,** 944–950.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model (Corr: V71 p219). *Biometrika* **69,** 331–342.

Schemper, M. and Henderson, R. (2000). Predictive accuracy and explained variation in Cox regression. *Biometrics* **56,** 249–255.

Shi, M., Taylor, J. M. G., Currier, R., Tang, H., Hoover, D., Chmiel, J., and Bryant, J. (1996). Replacing time since human immunodeficiency virus infection by marker values in predicting residual time to acquired immunodeficiency syndrome diagnosis. *Journal of Acquired Immune Deficiency Syndromes* **12,** 309–316.

Skates, S. J., Pauler, D. K., and Jacobs, I. J. (2001). Screening based on the risk of cancer calculation from Bayesian hierarchical changepoint and mixture models of longitudinal markers. *Journal of the American Statistical Association* **96,** 429–439.

Slate, E. H. and Turnbull, B. W. (2000). Statistical models for longitudinal biomarkers of disease onset. *Statistics in Medicine* **19,** 617–637.

Taylor, J. M. G., Munoz, A., Bass, S. M., Saah, A., Chmiel, J. S., and Kingsley, L. A. (1990). Estimating the distribution of times from HIV seroconversion to AIDS using multiple imputation. *Statistics in Medicine* **9,** 505–514.

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.

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.

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84,** 1065–1073.

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). Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics* **50,** 375–387.

Yang, Y. and Ying, Z. (2001). Marginal proportional hazards models for multiple event-time data. *Biometrika* **88,** 581–586.

Zheng, Y. (2002). Semiparametric methods for longitudinal diagnostic accuracy. Ph.D. Thesis, University of Washington, Seattle.

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