Joint Modeling of Longitudinal and Time-to-Event Data: An Overview

Anastasios A. Tsiatis* **and Marie Davidian**

Department of Statistics, North Carolina State University

Box 8203, Raleigh, North Carolina 27695-8203, U.S.A.

tsiatis@stat.ncsu.edu davidian@stat.ncsu.edu

Abstract

A common objective in longitudinal studies is to characterize the relationship between a longitudinal response process and a time-to-event. Considerable recent interest has focused on so-called joint models, where models for the event time distribution and longitudinal data are taken to depend on a common set of latent random effects. In the literature, precise statement of the underlying assumptions typically made for these models has been rare. We review the rationale for and development of joint models, offer insight into the structure of the likelihood for model parameters that clarifies the nature of common assumptions, and describe and contrast some of our recent proposals for implementation and inference.

Key words and phrases: Conditional score; Likelihood; Random effects; Seminonparametric.

**Corresponding author's phone/fax:* +1 919-515-1928/+1 919-515-7591

Short title: Joint Modeling Overview

1 Introduction

Longitudinal studies where repeated measurements on a continuous response, an observation on a possibly censored time-to-event ("failure" or "survival"), and additional covariate information are collected on each participant are commonplace in medical research, and interest

often focuses on interrelationships between these variables. A familiar example is that of HIV clinical trials, where covariates, including treatment assignment, demographic information, and physiological characteristics, are recorded at baseline, and measures immunologic and virologic status such as CD4 count and viral RNA copy number are taken at subsequent clinic visits. Time to progression to AIDS or death is also recorded for each participant, although some subjects may withdraw early from the study or fail to experience the event by the time of study closure. The study may have been designed to address the primary question of effect of treatment on time-to-event, but subsequent objectives may be (i) to understand within-subject patterns of change of CD4 or viral load and/or (ii) to characterize the relationship between features of CD4 or viral load profiles and time to progression or death. Similarly, in studies of prostate cancer, repeated prostate specific antigen (PSA) measurements may be obtained for patients following treatment for prostate cancer, along with time to disease recurrence. Again, characterizing (i) within-subject patterns of PSA change or (ii) the association between features of the longitudinal PSA process and cancer recurrence may be of interest.

Statement of objectives (i) and (ii) may be made more precise by thinking of "idealized" data for each subject $i = 1, \ldots, n$ as $\{T_i, Z_i, X_i(u), u \geq 0\}$, where T_i is event time, Z_i is a vector of baseline (time 0) covariates, and $\{X_i(u),\ u \geq 0\}$ is the longitudinal response trajectory for all times $u \geq 0$. Objective (i), elucidating patterns of change of the longitudinal response (e.g., CD4 or PSA) may involve, for example, estimating aspects of average longitudinal behavior and its association with covariates, such as $E\{X_i(u)|Z_i\}$ and $\text{var}\{X_i(u)|Z_i\}$. A routine framework for addressing objective (ii), characterizing associations among the longitudinal and time-to-event processes and covariates, is to represent the relationship between T_i, $X_i(u)$, and Z_i by a proportional hazards model

$$\begin{aligned} \lambda_i(u) &= \lim_{du \to 0} du^{-1} \text{pr}\{u \leq T_i < u + du | T_i \geq u, X_i^H(u), Z_i\} \\ &= \lambda_0(u) \exp\{\gamma X_i(u) + \eta^T Z_i\}, \end{aligned} \tag{1}$$

where $X_i^H(u) = \{X_i(t), 0 \leq t < u\}$ is the history of the longitudinal process up to time u and, as in Kalbfleisch and Prentice ((2002), Sec. 6.3), $X_i(u)$ should be left-continuous; in the sequel, we consider only continuous $X_i(u)$ so do not dwell on this requirement. In (1), for definiteness, the hazard is taken to depend linearly on history through the current value $X_i(u)$, but other specifications and forms of relationship are possible. Here, $X_i(u)$ may be regarded as a time-dependent covariate in (1), and formal assessment of the association involves estimation of γ and η. The first author's interest in this problem arose from his work with the AIDS Clinical Trials Group (ACTG) in the 1990s, when the issue of whether longitudinal CD4 count is a "surrogate marker" for time to progression to AIDS or death was hotly debated; that is, can the survival endpoint, which may be lengthy to ascertain, be replaced by short-term, longitudinal CD4 measurements to assess treatment efficacy? Prentice (1989) set forth conditions for surrogacy: (I) treatment must have an effect on the time-to-event; (II) treatment must have effect on the marker; and (III) effect of treatment should manifest through the marker, i.e., the risk of the event given a specific marker trajectory should be independent of treatment. Tsiatis, DeGruttola, and Wulfsohn (1995) evaluated potential surrogacy by noting that, if (I) and (II) hold and Z_i is a treatment indicator, then in (1) $\gamma < 0$ and $\eta = 0$ would be consistent with (III) and hence surrogacy of $X_i(u)$.

Formalizing these objectives is straightforward in terms of the "idealized" data, but addressing them in practice is complicated by the nature of the data actually observed. Although the above formulation involves the longitudinal response at any time u, the response is collected on each i only intermittently at some set of times $t_{ij} \leq T_i$, $j = 1, \ldots, m_i$. Moreover, the observed values may not be the "true" values $X_i(t_{ij})$; rather, we may observe only $W_i(t_{ij}) = X_i(t_{ij}) + e_i(t_{ij})$, say, where $e_i(t_{ij})$ is an intra-subject "error." In fact, the event time T_i may not be observed for all subjects but may be censored for a variety of reasons.

A further difficulty for making inference on the longitudinal process is that occurrence of

the time-to-event may induce an informative censoring, as discussed by Wu and Carroll (1988), Hogan and Laird (1997ab), and many other authors. For example, subjects with more serious HIV disease may be more likely to experience the event or withdraw from the study earlier than healthier individuals, leading to fewer CD4 measurements, and to have sharper rates of CD4 decline. Failure to take appropriate account of this phenomenon, e.g., by using ordinary longitudinal data techniques, can lead to biased estimation of average quantities of interest. Valid inference requires a framework in which potential underlying relationships between the event and longitudinal process are explicitly acknowledged. A philosophical issue of whether the form of trajectories that might have occurred after death is a meaningful scientific focus is sometimes raised; we do not discuss this here.

For objective (ii), if, ideally, $X_i(u)$ were observable at all $u \leq T_i$, then the main difficulty for fitting (1) is censoring of the time-to-event. Letting C_i denote underlying potential censoring time for subject i in the usual way, then we observe only $V_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$, where $I(\cdot)$ is the indicator function. Following Cox (1972,1975) and under conditions discussed by Kalbfleisch and Prentice ((2002), Secs. 6.3, 6.4), inference on γ and η is made by maximizing the partial likelihood

$$\prod_{i=1}^{n} \left[\frac{\exp\{\gamma X_i(V_i) + \eta^T Z_i\}}{\sum_{j=1}^{n} \exp\{\gamma X_j(V_i) + \eta^T Z_j\} I(V_j \geq V_i)} \right]^{\Delta_i}. \tag{2}$$

It is clear from (2) that this approach is predicated on availability of $X_i(u)$ for all $i = 1, \ldots, n$ at each observed failure time. In practice, implementation is complicated by the fact that $X_i(u)$ is available only intermittently for each subject and is possibly subject to error. Early attempts to circumvent this difficulty involved some sort of "naive" imputation of required $X_i(u)$. For example, in the method of "Last Value Carried Forward (LVCF)," the unavailable, true $X_i(u)$ values in (2) are replaced by the most recent, observed value for i. Prentice (1982) showed that such substitution leads to biased estimation of model parameters. A more sophisticated framework in which these features may be incorporated is required.

The complications posed by the realities of the observed data and the potential for biased inferences for both objectives (i) and (ii) if ordinary or naive techniques are applied have led to considerable recent interest in so-called "joint models" for the longitudinal data and time-to-event processes. In this paper, we review the development of joint models and offer insights into the underlying assumptions that commonly accompany their use. As we describe in more detail in Section 2, models for the (possibly error-prone) longitudinal process and the hazard for the (possibly censored) time-to-event are taken to depend jointly on shared, underlying random effects. Under this assumption, it has been demonstrated by numerous authors, whose work is cited herein, that use of these models leads to correction of biases and the potential for enhanced efficiency. Our observation has been that, in much of this literature, precise, transparent statements of the assumptions on the interrelationships among model components have been lacking; thus, in Section 3, we attempt to offer clarification by stating a simplified version of a joint-model likelihood for model parameters and making explicit how common assumptions formally arise. Section 4 describes briefly two approaches for inference we have recently proposed, which are evaluated and contrasted in Section 5.

Throughout, we focus mainly on the data-analytic objective (ii), and thus on inference for γ and η in (1); however, the models we discuss in the sequel have been used for both objectives.

2 Joint modeling

2.1 Modeling considerations

As in Section 1, for subject i, $i = 1, \ldots, n$, let T_i and C_i denote the event and censoring times, respectively; let Z_i be a q-dimensional vector of baseline covariates and let $X_i(u)$ be the longitudinal process at time $u \geq 0$. Components of Z_i might also be time-dependent covariates

whose values are known exactly and that are "external" in the sense described by Kalbfleisch and Prentice (2002), Sec. 6.3.1); our consideration of baseline covariates for simplicity does not alter the general insights we highlight in the next section. Rather than observe T_i for all i, we observe only $V_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$. Values of $X_i(u)$ are measured intermittently at times $t_{ij} \leq V_i$, $j = 1, \ldots, m_i$, for subject i, which may be different for each i; often, target values for the observations times are specified by a study protocol, although deviations from protocol are common. The observed longitudinal data on subject i may be subject to "error" as described below; thus, we observe only $W_i = \{W_i(t_{i1}), \ldots, W_i(t_{im_i})\}^T$, whose elements may not exactly equal the corresponding $X_i(t_{ij})$.

A joint model is comprised of two linked submodels, one for the "true" longitudinal process $X_i(u)$ and one for the failure time T_i, along with additional specifications and assumptions that allow ultimately a full representation of the joint distribution of the observed data $O_i = \{V_i, \Delta_i, W_i, t_i\}$, where $t_i = (t_{i1}, \ldots, t_{im_i})^T$. The O_i are taken to be independent across i, reflecting the belief that the disease process evolves independently for each subject.

For the longitudinal response process, a standard approach is to characterize $X_i(u)$ in terms of a vector of subject-specific random effects α_i. For example, a simple linear model

$$X_i(u) = \alpha_{0i} + \alpha_{1i}u, \quad \alpha_i = (\alpha_{0i}, \alpha_{1i})^T \tag{3}$$

has been used to represent "true" (log) CD4 trajectories (e.g., Self and Pawitan (1992); Tsiatis, DeGruttola and Wulfsohn (1995); DeGruttola and Tu (1994); Faucett and Thomas (1996); Wulfsohn and Tsiatis (1997); Bycott and Taylor (1998); Dafni and Tsiatis (1998)). Here, α_{0i} and α_{1i} are subject-specific intercept and slope, respectively. More flexible models may also be considered; e.g., a polynomial specification $X_i(u) = \alpha_{0i} + \alpha_{1i}u + \alpha_{2i}u^2 + \cdots + \alpha_{pi}u^p$, or, more generally,

$$X_i(u) = f(u)^T \alpha_i, \tag{4}$$

where $f(u)$ is a vector of functions of time u (so including splines), may be posited, as may,

in principle, nonlinear functions of α_i, although we restrict attention to linear functions here.

Models for $X_i(u)$ of form (4) specify that the "true" longitudinal process follows a "smooth" trajectory dictated by a small number of time-invariant, subject-specific effects, so that evolution of the process throughout time is determined by these effects. Alternatively, Taylor, Cumberland, and Sy (1994), Lavalley and DeGruttola (1996), Henderson, Diggle, and Dobson (2000), Wang and Taylor (2001), and Xu and Zeger (2001a) consider models of the form

$$X_i(u) = f(u)^T \alpha_i + U_i(u), \tag{5}$$

where $U_i(u)$ is a mean-zero stochastic process, usually taken to be independent of α_i and Z_i. Taylor, Cumberland and Sy (1994), Lavalley and DeGruttola (1996), and Wang and Taylor (2001) specify $U_i(u)$ to be an integrated Ornstein-Uhlenbeck (IOU) process. Henderson, Diggle and Dobson (2000) and Xu and Zeger (2001) discuss taking $U_i(u)$ to be a stationary Gaussian process. As noted by these authors, (5) allows the trend to vary with time and induces a within-subject autocorrelation structure that may be thought of as arising from evolving biological fluctuations in the process about a smooth trend. For example, from a biological perspective, one may believe that a response like CD4 does not decline mostly smoothly, but rather that patients go through "good" and "bad" periods as their disease progresses. Typically, $f(u)^T \alpha_i$ is taken to be simple (e.g., random intercept and slope or random intercept only). We contrast the specifications (4) and (5) further below.

In (4) and (5), the α_i (given Z_i) are usually assumed to be normally distributed to represent inter-subject variation in the features of the "true" longitudinal trajectories. Depending on the situation, the mean and covariance matrix of α_i may be taken to depend on components of Z_i; e.g., if Z_i contains a treatment indicator, mean intercept and slope in (3) may have a different mean and covariance matrix for each treatment group. A common specification is that the α_i are independent of Z_i so that the same normal model holds for all i.

For the event time process, a model of the form (1) is posited, where dependence on the

covariate history $X_i^H(u)$ implies dependence on α_i (and $U_i(u)$ for models of the form (5)). Modifications are possible to reflect different specifications for the underlying mechanism; for example, under (3) or similar model of form (5), taking $\lambda_i(u) = \lim_{du \to 0} du^{-1} \text{pr}\{u \leq T_i < u + du | T_i \geq u, X_i^H(u), Z_i\} = \lambda_0(u) \exp\{\gamma \alpha_{1i} + \eta^T Z_i\}$ reflects the belief that the hazard, conditional on the longitudinal history and covariates, is associated mainly with the assumed constant rate of change of the underlying smooth trend; the rationale is discussed further below. See Song, Davidian, and Tsiatis (2002a) for consideration of such hazard models in an HIV case study.

Given a specification for $X_i(u)$, the observed longitudinal data W_i are taken to follow

$$W_i(t_{ij}) = X_i(t_{ij}) + e_i(t_{ij}), \tag{6}$$

where $e_i(t_{ij}) \sim \mathcal{N}(0, \sigma^2)$ are independent of α_i (and $U_i(u)$ for all $u \geq 0$ if present). Under the perspective of (5), the $e_i(t_{ij})$ represent deviations due to measurement error and "local" biological variation that is on a sufficiently short time scale that $e_i(t_{ij})$ may be taken independent across j; e.g., if the scale of time is on the order of months or years, diurnal variation may be regarded as "local." From the point of view of models like (4), the $e_i(t_{ij})$ may be thought of as representing measurement error and biological variation due both to local and longer-term within-individual autocorrelation processes; thus, one way to view (4) is that $U_i(u)$ in (5) has been "absorbed" into $e_i(t_{ij})$. Under this model, it is often argued (e.g., Tsiatis, DeGruttola and Wulfsohn (1995); Song, Davidian, and Tsiatis (2002ab)) that if the t_{ij} are at sufficiently long intervals so that within-subject autocorrelation among observed values is practically negligible, or if measurement error is large relative to biological fluctuations, the assumption that $e_i(t_{ij})$ are independent across j is still reasonable. Note that (6), along with (4), specifies a standard linear mixed effects model (e.g., Laird and Ware (1982)). Alternatively, under this model, if observations are sufficiently close that this autocorrelation cannot be disregarded, the $e_i(t_{ij})$ admit a covariance structure that must be taken into account.

2.2 Philosophical considerations

Whether one takes the model for the "true" longitudinal process to be of form (4) or (5) is to some extent a philosophical issue dictated in part by one's belief about the underlying biological mechanisms and the degree to which the proposed model is meant to be an empirical or "realistic" representation of the result of these mechanisms. In particular, taken literally, a model of form (4) for the longitudinal process takes the trajectory followed by a subject throughout time to be dictated by time-independent random effects α_i, thus implying that the smooth trend followed by the subject's trajectory is an "inherent" characteristic of the subject that is fixed throughout time. This may or may not be realistic from a biological point of view; e.g., it may be more biologically plausible to expect the trend to vary with time as disease evolves. However, often, the goal of investigators is to assess the degree of association between the dominant, smooth trend and event time, rather than to characterize accurately the underlying biological process. Here a model of form (4), although admittedly a gross simplification as is the case for many popular statistical models, may provide an adequate empirical approximation for this purpose; under this perspective, the α_i are empirical representations of the dominant part of the trajectory. On the other hand, a model of form (5) may be viewed as an attempt to capture more precisely the features of the trajectory, allowing the trend to vary over time, so in some sense getting "closer" to the true biological mechanism dictating the association with the event. Then again, from a purely empirical, statistical point of view, (5) can be viewed as a way to incorporate within-subject autocorrelation and allow a "wiggly" empirical representation of within-subject trends.

These considerations hence have implications for how one chooses to specify the event time model. Consider (1), where dependence of the hazard at u on longitudinal response history is through the current value $X_i(u)$. If a model of the form (4) is postulated, this reflects the view that the value of the smooth trend at u is the predominant feature associated with

prognosis; i.e., the association with event time depends on where, in the context of an HIV study, a subject's CD4 is in general at u, irrespective of local fluctuations from a dominant trend. In contrast, a model of the form (5) emphasizes the belief that fluctuations about an overall trend are themselves related to the event. For instance, from an approximate biological point of view, it may be thought that the occurrence of "good" and "bad" periods of CD4 are related to prognosis; alternatively, such a model may be postulated from the purely empirical perspective that "wiggles" are an important aspect of the association. Of course, along the lines of the usual "signal or noise?" debate, it may be argued that either type of model may be employed if the main purpose is to represent empirically the hazard (1) in terms of what are believed to be relevant features of the longitudinal process. A complicated model of the form (4) involving higher-order polynomial terms or splines, may capture "wiggles" in a way comparable to a model of form (5).

Alternatively, consider a model for the hazard at u depending on a longitudinal process of either form (4) or (5) through, say, a slope α_{1i} only, as discussed above. Such a model reflects the belief that whether a patient is on an overall downward CD4 trajectory, which presumably indicates increasingly serious disease, is the main factor thought to be associated with progression to AIDS or death, irrespective of "wiggles" or local "good" or "bad" periods.

A drawback of models of form (5) is their relatively more difficult implementation. Perhaps in part due to this feature and in part due to interest in the relationship between prognosis and an underlying, smooth trend (e.g., from an empirical modeling perspective), most accounts of application of joint models in the literature have focused on models of the form (4).

2.3 Joint models in the literature

Much of the early literature on joint modeling focuses on models of the form (4) for the longitudinal data process; we recount some highlights of this literature here. Schluchter (1992)

and Pawitan and Self (1993) considered joint models in which times-to-event (and truncation and censoring for the latter authors) are modeled parametrically, which facilitates in principle straightforward likelihood inference. Later papers adopted the proportional hazards model, mostly of the form (1), for the time-to-event. Raboud, Reid, Coates and Farewell (1993) did not consider a full joint model *per se* in which $X_i(u)$ is explicitly modeled; they focused mainly on potential bias due to use of LVCF and failure to account for measurement error, and also showed that simple smoothing methods to impute the $X_i(u)$ in (1) yield reduced bias relative to naive approaches.

Other work directly considered inference in a joint model defined by submodels for the longitudinal and event time processes. Self and Pawitan (1992) proposed a longitudinal model of form (4) and hazard model similar to (1) with the term $\exp\{\gamma X_i(u)\}$ replaced by $\{1 + \gamma X_i(u)\}$, so that the hazard is linear in α_i. They developed a "two-step" inferential strategy in which the individual least squares estimators for the α_i are used to impute appropriate values for $X_i(u)$ that are then substituted into a partial likelihood based on the hazard model. Tsiatis, DeGruttola and Wulfsohn (1995) proposed a model exactly of form (4) and (1) as a framework for inference on the relationship between log CD4 $X_i(u)$ and mortality T_i in an early HIV study. Due to the nonlinearity of (1) in the α_i, these authors developed an approximation to the hazard for T_i given the *observed* covariate history $W_i^H(u)$, say, that suggests implementing the joint model by maximizing the usual partial likelihood (2) with $X_i(u)$ for each failure time replaced by an "estimate" of $E\{X_i(u)|W_i^H(u), T_i \geq u\}$. In particular, for event time u at which $X_i(u)$ is required, they advocate using the empirical Bayes "estimator" or "EBLUP" for $X_i(u)$ based on a standard fit of the mixed effects model defined by (4) and (6) to the data up to time u from all subjects still in the risk set at u (so with $T_i \geq u$). This approach thus requires fitting as many linear mixed effects models as there are event times in the data set; an advantage over more computationally-intensive methods discussed below is that it may

be implemented using standard software for mixed effects and proportional hazards models. Such a "two-stage" approach to inference has also been considered by Bycott and Taylor (1998), who compared several procedures, and Dafni and Tsiatis (1998) who investigated the performance of the Tsiatis, DeGruttola and Wulfsohn (1995) procedure via simulation. These authors found that this approximate method yields estimators for γ and η in (1) that reduce but do not completely eliminate bias relative to naive methods (see also Tsiatis and Davidian (2001)). This is in part due to the approximation and in part a result of the fact that, under normality of $e_i(t_{ij})$ and α_i, at any event time $u > 0$, the $W_i(t_{ij})$, $t_{ij} \leq u$ for subjects still in the risk set, on which fitting the mixed model and hence the imputation of $X_i(u)$ for each u depends, are a biased sample from a normal distribution, and thus are no longer marginally normally distributed. Hence, as the form of the empirical Bayes estimator/EBLUP for $X_i(u)$ depends critically on validity of the normality assumption, which is successively less relevant as u increases, the resulting imputed $X_i(u)$ are not entirely appropriate.

Rather than rely on approximations, other authors have taken a likelihood approach, based on specification of a likelihood function for the parameters in (4), (6), and a model for the time-to-event. DeGruttola and Tu (1994) considered a longitudinal data model of the form (4) along with a parametric (lognormal) model for T_i and developed an EM algorithm to maximize the resulting loglikelihood, which involves intractable integrals over the distribution of α_i, taken to be normally distributed. For the proportional hazards model (1), (4), and (6), Wulfsohn and Tsiatis (1997) derived an EM algorithm based on a loglikelihood specification for γ, η, σ^2, the infinite-dimensional parameter $\lambda_0(u)$, and parameters in a normal model for α_i. More recently, Henderson, Diggle and Dobson (2000) have discussed likelihood inference for models of the form (1) and (5). Lin, Turnbull, McCulloch and Slate (2002) consider a related model in which dependence of longitudinal and event-time models on a common random effect is replaced by a shared dependence on a latent class variable that accommodates underlying population

heterogeneity. Song, Davidian and Tsiatis (2002b) considered the model of Wulfsohn and Tsiatis (1997) but relaxed the assumption of normality of the α_i for the random effects to one of requiring only that α_i have distribution with a "smooth" density; this method is described further in Section 4.2.

Faucett and Thomas (1996) took a Bayesian approach to joint models of the form (1) and (4), and developed and demonstrated implementation via Markov chain Monte Carlo (MCMC) techniques. Xu and Zeger (2001ab) considered generalizations of this approach, and allowed models of form (5). Wang and Taylor (2001) incorporated a longitudinal model of the form (5) into a Bayesian framework, again using MCMC to fit a joint model to data from a study of HIV disease, while Brown and Ibrahim (2003a) considered a semiparametric Bayesian joint model of form (1) and (4) that furthermore makes no parametric assumption on the random effects. Brown and Ibrahim (2003b) and Law, Taylor, and Sandler (2002) developed joint models in the presence of a fraction cured. Ibrahim, Chen and Sinha ((2001), Chap. 7) provide a detailed discussion of joint modeling, particularly from a Bayesian perspective.

Both Bayesian and likelihood procedures rely on specification of an appropriate likelihood for the joint model parameters. In the next section, we offer some insight into the form of specifications that are routinely used and the assumptions they embody.

Tsiatis and Davidian (2001) took an entirely different approach, focusing on estimation of γ and η in a joint model of the form (1) and (4). To minimize reliance on parametric modeling assumptions, these authors developed a set of unbiased estimating equations that yield consistent and asymptotically normal estimators with no assumptions on the α_i. The rationale for and derivation of this procedure are given in Section 4.1.

The focus of most authors in the preceding discussion was on characterizing the association between the longitudinal and event time processes. Alternatively, Faucett, Schenker, and Taylor (2002) and Xu and Zeger (2001a) used joint models as a framework in which to make

make more efficient inference on the marginal (given, say, baseline covariates such as treatment) event-time distribution by incorporating the longitudinal data as auxiliary information; see also Hogan and Laird (1997a).

3 Likelihood formulation and assumptions

For definiteness, we consider first the joint model defined by (1), (4), and (6) along with the usual assumption of normal, independent $e_i(t_{ij})$. Letting δ denote parameters in the density $p(\alpha_i|Z_i;\delta)$ for α_i given Z_i, which is ordinarily assumed multivariate normal, the usual form of a likelihood for the full set of parameters of interest, $\Omega = \{\lambda_0(\,\cdot\,), \gamma, \eta, \sigma^2, \delta\}$, conditional on Z_i, is

$$\prod_{i=1}^{n} \int \Big[\lambda_0(V_i)\exp\{\gamma X_i(V_i) + \eta^T Z_i\}\Big]^{\Delta_i} \exp\left[-\int_0^{V_i} \lambda_0(u)\exp\{\gamma X_i(u) + \eta^T Z_i\}\,du\right] \times \frac{1}{(2\pi\sigma^2)^{m_i/2}} \exp\left[-\sum_{j=1}^{m_i} \frac{\{W_i(t_{ij}) - X_i(t_{ij})\}^2}{2\sigma^2}\right] p(\alpha_i|Z_i;\delta)\,d\alpha_i. \qquad (7)$$

In most of the literature, (7) is advocated as the basis for inference with little comment regarding its derivation; generally, a statement that (7) follows from the assumption that censoring and timing of longitudinal measurements are "uninformative" is made, with no attempt to elucidate this assumption more formally. DeGruttola and Tu (1994) give a likelihood specification that explicitly includes components due to timing of measurements and censoring in making such an assumption, but the underlying formal considerations are not discussed.

For a longitudinal model of form (1), (5), and (6), the details in the literature are similarly sketchy. To obtain a likelihood function analogous to (7), one would be required to integrate the integrand in (7) with respect to the joint distribution of α_i and the infinite-dimensional process $U_i(\,\cdot\,)$ (given Z_i); even if $U_i(u)$ and α_i are independent for all u, this is still a formidable technical problem. From a practical perspective, some authors (e.g., Wang and Taylor (2001)) circumvent such difficulties by approximating continuous $U_i(u)$ by

a piecewise constant function on a fine time grid containing all time points where longitudinal measurements are available for subject i; the $U_i(u)$ on the grid then are regarded as a finite-dimensional vector in a similar vein as the α_i. Alternatively, Henderson, Diggle and Dobson (2000) argue in their formulation that, because the nonparametric estimator of the baseline hazard $\lambda_0(u)$ is zero except at observed event times, one need only consider finite such integration with respect to the joint distribution of $\{U_i(V_1), \ldots, U_i(V_{n_d})\}$, where $n_d = \sum_i \Delta_i$ is the number of (assumed distinct) event times.

To shed light on the nature of assumptions that may lead to the likelihood specification (7) and its counterpart for (5), we consider a discretized version of the problem in a spirit similar to arguments of Kalbfleisch and Prentice ((2002), Sec. 6.3.2). In particular, suppose that subjects are followed on the interval $[0, L)$, and consider time discretized over a fine grid $t_0, t_1, t_2, \ldots, t_M$, where $t_0 = 0$ is baseline, $t_M = L$, and ultimately M will be taken to be large. We may then conceptualize the data-generating process for subject i as we now describe. For simplicity, in the case of (4), conditional on α_i and Z_i, assume all subjects have a baseline longitudinal measurement, $W_i(t_0)$ for i, that, given (α_i, Z_i), is independent of T_i; this could be relaxed without altering the upcoming results. Similarly for a model of form (5), letting $U_{ik} = U_i(t_k)$ for the kth grid point, we may also condition on $\mathcal{U}_i = (U_{i0}, \ldots, U_{iM})$ here and in the developments below; for brevity, we consider (4) in what follows and comment briefly on this case. Additional data then arise according to the following scheme.

- Does death occur at t_1? Let $\mathcal{D}_i(t_1) = I(V_i = t_1, \Delta_i = 1)$, which corresponds to $T_i = t_1$, $C_i > t_1$. If $\mathcal{D}_i(t_1) = 1$, stop collecting data.

- Otherwise, if $\mathcal{D}_i(t_1) = 0$, does censoring occur at t_1? Let $\mathcal{C}_i(t_1) = I(V_i = t_1, \Delta_i = 0)$, corresponding to $T_i > t_1$, $C_i = t_1$. If $\mathcal{C}_i(t_1) = 1$, stop collecting data.

- Otherwise, if $\mathcal{C}_i(t_1) = 0$, then is a longitudinal measurement taken at t_1? If so, define

$\mathcal{R}_i(t_1) = 1$, with $\mathcal{R}_i(t_1) = 0$ otherwise.

- If $\mathcal{R}_i(t_1) = 1$, then $W_i(t_1)$ is the longitudinal observation taken at t_1.
- Does death occur at t_2? Let $\mathcal{D}_i(t_2) = I(V_i = t_2, \Delta_i = 1)$ $(T_i = t_2)$. If $\mathcal{D}_i(t_2) = 1$, stop collecting data.

This pattern continues until death or censoring terminates data generation, at which point all subsequent values for $\mathcal{D}_i(\,\cdot\,)$, $\mathcal{C}_i(\,\cdot\,)$, and $\mathcal{R}_i(\,\cdot\,)$ are set equal to zero. We may then summarize the observed data as $\{W_i(t_0), \mathcal{D}_i(t_j), \mathcal{C}_i(t_j), \mathcal{R}_i(t_j), \mathcal{R}_i(t_j)W_i(t_j),\ j = 1, \ldots, M\}$.

Now define $Y_i(t_j) = 1 - \sum_{\ell=1}^{j-1}\{\mathcal{D}_i(t_\ell) + \mathcal{C}_i(t_\ell)\} = I(V_i \geq t_j)$,

$$\lambda_i(t_j) = P\{T_i = t_j | T_i \geq t_j, C_i \geq t_j, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \leq \ell < j, \alpha_i, Z_i\} \tag{8}$$

$$\mu_i(t_j) = P\{C_i = t_j | T_i > t_j, C_i \geq t_j, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \leq \ell < j, \alpha_i, Z_i\} \tag{9}$$

$$\pi_i(t_j) = P\{\mathcal{R}_i(t_j) = 1 | T_i > t_j, C_i > t_j, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \leq \ell < j, \alpha_i, Z_i\} \tag{10}$$

$$\varphi_i(t_j, w_j) = P\{W_i(t_j) = w_j | T_i > t_j, C_i > t_j, \mathcal{R}_i(t_\ell), \mathcal{R}(t_\ell)W_i(t_\ell), 0 \leq \ell < j, \mathcal{R}_i(t_j) = 1, \alpha_i, Z_i\}, \tag{11}$$

where $\mathcal{R}_i(t_0) \equiv 1$. For (4), we may then specify the joint density for the observed data, conditional on (α_i, Z_i), to obtain a conditional likelihood contribution for i as

$$\begin{aligned} f_i(\,\cdot\,|\alpha_i, Z_i) = \varphi_i\{t_0, W_i(t_0)\} \prod_{j=1}^{M} \Big(& \{\lambda_i(t_j)Y_i(t_j)\}^{\mathcal{D}_i(t_j)}\{1 - \lambda_i(t_j)Y_i(t_j)\}^{1-\mathcal{D}_i(t_j)} \\ & \times [\mu_i(t_j)\{Y_i(t_j) - \mathcal{D}_i(t_j)\}]^{C_i(t_j)}[1 - \mu_i(t_j)\{Y_i(t_j) - \mathcal{D}_i(t_j)\}]^{1-\mathcal{C}_i(t_j)} \\ & \times [\pi_i(t_j)\{Y_i(t_j) - \mathcal{D}_i(t_j) - \mathcal{C}_i(t_j)\}]^{\mathcal{R}_i(t_j)}[1 - \pi_i(t_j)\{Y_i(t_j) - \mathcal{D}_i(t_j) - \mathcal{C}_i(t_j)\}]^{1-\mathcal{R}_i(t_j)} \\ & \times [\mathcal{R}_i(t_j)\varphi_i\{t_j, W_i(t_j)\}]^{\mathcal{R}_i(t_j)} \Big), \end{aligned} \tag{12}$$

where $0^0 = 1$. These developments lead to a likelihood for Ω (conditional on Z_i), of the form

$$\prod_{i=1}^{n} L_i(\,\cdot\,) = \prod_{i=1}^{n} \int f_i(\,\cdot\,|\alpha_i, Z_i) p(\alpha_i | Z_i; \delta)\, d\alpha_i. \tag{13}$$

In the case of (5), with $U_i(\,\cdot\,)$ independent of α_i, f_i would also involve conditioning on $\mathcal{U}_i$ and integration would also be with respect to the $M+1$-variate density of $\mathcal{U}_i$.

We may now make a correspondence between (13) and (7), where M is large, and highlight the usual assumptions that lead to (7). In particular, $\lambda_i(t_j)$ in (8) is taken to satisfy

$$\begin{aligned}\lambda_i(t_j) &= P\{T_i = t_j | T_i \geq t_j, C_i \geq t_j, \mathcal{R}_i(t_\ell), \mathcal{R}(t_\ell)W_i(t_\ell), 0 \leq \ell < j, \alpha_i, Z_i\} \\ &= P(T_i = t_j | T_i \geq t_j, \alpha_i, Z_i) = \lambda_0(t_j)dt_j \exp\{\gamma X_i(t_j) + \eta^T Z_i\}; \end{aligned} \tag{14}$$

here, $\exp[\int_0^{V_i} \lambda_0(u) \exp\{\gamma X_i(u) + \eta^T Z_i\}\, du]$ is approximated in discrete time by $\prod_{j=1}^{M}\{1 - \lambda_i(t_j)Y_i(t_j)\}^{1-\mathcal{D}_i(t_j)}$. Inspection of (7) shows that (11) is assumed to have the form

$$\varphi_i(t_j, w_j) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{\{W_i(t_j) - X_i(t_j)\}^2}{2\sigma^2}\right], \tag{15}$$

reflecting the usual specification that the "errors" in (6) are mutually independent and independent of all other variables, conditional on (α_i, Z_i). The likelihood (7) does not include terms corresponding to $\mu_i(t_j)$ and $\pi_i(t_j)$ in (9) and (10). If $\mu_i(t_j)$ and $\pi_i(t_j)$ do not depend on α_i, these terms "factor out" of the integral in (13) and may thus be disregarded. The term $\mu_i(t_j)$ has to do with the censoring mechanism, while $\pi_i(t_j)$ involves the timing of longitudinal measurements. Thus, the usual assumption that "censoring and timing of longitudinal measurements are uninformative" may be seen to correspond formally to the assumption that the conditional probabilities in this factorization do not depend on α_i. Practically speaking, from (9) and (10), this assumption implies the belief that decisions on whether a subject withdraws from the study or appears at the clinic for a longitudinal measurement depend on observed past history (longitudinal measurements and baseline covariates), but there is no additional dependence on underlying, latent subject characteristics associated with prognosis.

In the case of (5), failure of $\mu_i(t_j)$ and $\pi_i(t_j)$ to depend on $\mathcal{U}_i$ would have a similar interpretation. Note then that the integration in (13) with respect to $\mathcal{U}_i$ would only involve the elements up to the last grid time t_k for which $\mathcal{R}_i(t_k) = 1$,so that the $M+1$-dimensional

integration is in reality only k-dimensional. A rigorous argument to elucidate the behavior of (13) as M gets large in this case would be challenging; informally, mass will be placed only at the event times, so that we expect that (13) will involve only the multivariate distribution of the $U_i(u)$'s at the event times in the limit, as discussed by Henderson, Diggle and Dobson (2000).

It is natural to be concerned whether the usual assumption given in (14), that $P\{T_i = t_j | T_i \geq t_j, C_i \geq t_j, \mathcal{R}_i(t_\ell), \mathcal{R}(t_\ell)W_i(t_\ell), 0 \leq \ell < j, \alpha_i, Z_i\} = P\{T_i = t_j | T_i \geq t_j, \alpha_i, Z_i\}$, is relevant in practice; we consider this in the context of (4). Ideally, interest focuses on the relationship between the event time and the "true" longitudinal process, as in (1). However, the observable relationship is that in the first line of (14), that between event time and all observable quantities up to the current time in addition to the latent "true" process (represented through the random effects). In order to identify the relationship of interest from the observable information then, it is critical to identify reasonable conditions that would be expected to hold in practice that would lead to the equality in the second line of (14). We now elucidate such conditions; specifically, we demonstrate that (14) follows from assumptions that are similar in spirit to "missing at random," conditional on α_i. In particular, we sketch an argument showing that, if, for all $j = 1, \ldots, M$ and $k < j$,

$$P\{C_i = t_k | T_i > t_k, C_i \geq t_k, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \leq \ell < k, T_i = t_j, \alpha_i, Z_i\} \tag{16}$$
$$= P\{C_i = t_k | T_i > t_k, C_i \geq t_k, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \leq \ell < k, \alpha_i, Z_i\} = \mu_i(t_k),$$

$$P\{\mathcal{R}_i(t_k) = 1 | T_i > t_k, C_i > t_k, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \leq \ell < k, T_i = t_j, \alpha_i, Z_i\} \tag{17}$$
$$= P\{\mathcal{R}_i(t_k) = 1 | T_i > t_k, C_i > t_k, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \leq \ell < k, \alpha_i, Z_i\} = \pi_i(t_k).$$

$$P\{W_i(t_j) = w_k | T_i > t_k, C_i > t_k, \mathcal{R}_i(t_\ell), \mathcal{R}(t_\ell)W_i(t_\ell), 0 \leq \ell < k, \mathcal{R}_i(t_k) = 1, T_i = t_j, \alpha_i, Z_i\} \tag{18}$$
$$= P\{W_i(t_k) = w_k | T_i > t_k, C_i > t_k, \mathcal{R}_i(t_\ell), \mathcal{R}(t_\ell)W_i(t_\ell), 0 \leq \ell < k, \mathcal{R}_i(t_k) = 1, \alpha_i, Z_i\} = \varphi_i(t_k, w_k),$$

then (14) holds. In words, (16)–(18) state that probabilities of censoring, timing, and mea-

surement depend only on observed history and latent random effects but not on the future event time itself. To see this, note that

$$P\{T_i = t_j | T_i \geq t_j, C_i \geq t_j, \mathcal{R}_i(t_\ell), \mathcal{R}(t_\ell)W_i(t_\ell), 0 \leq \ell < j, \alpha_i, Z_i\}$$
$$= \frac{P\{T_i = t_j, C_i > t_1, \ldots, C_i > t_{j-1}, \mathcal{R}_i(t_\ell), \mathcal{R}(t_\ell)W_i(t_\ell), 0 \leq \ell < j, \alpha_i, Z_i\}}{P\{T_i \geq t_j, C_i > t_1, \ldots, C_i > t_{j-1}, \mathcal{R}_i(t_\ell), \mathcal{R}(t_\ell)W_i(t_\ell), 0 \leq \ell < j, \alpha_i, Z_i\}}, \quad (19)$$

where we have written the event $\{C_i \geq t_j\}$ equivalently as $\{C_i > t_1, \ldots, C_i > t_{j-1}\}$. By the assumption that $W_i(t_0)$ is independent of T_i, we may factor the numerator in (19) as

$$\begin{aligned}
&P(T_i = t_j | \alpha_i, Z_i) P\{W_i(t_0) = w_0 | \alpha_i, Z_i\} \\
&\times P\{C_i > t_1 | W_i(t_0), T_i = t_j, \alpha_i, Z_i\} P\{\mathcal{R}_i(t_1) = r_1 | C_i > t_1, W_i(t_0), T_i = t_j, \alpha_i, Z_i\} \\
&\times P\{W_i(t_1) = w_1 | C_i > t_1, \mathcal{R}_i(t_1) = 1, W_i(t_0), T_i = t_j, \alpha_i, Z_i\}^{\mathcal{R}_i(t_1)} \\
&\times \prod_{k=2}^{j-1} \Big[P\{C_i > t_k | C_i > t_{k-1}, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \leq \ell < k, T_i = t_j, \alpha_i, Z_i\}. \quad (20) \\
&\qquad \times P\{\mathcal{R}_i(t_k) = r_k | C_i > t_k, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \leq \ell < k, T_i = t_j, \alpha_i, Z_i\} \\
&\qquad \times P\{W_i(t_k) = w_k | C_i > t_k, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \leq \ell < k, \\
&\qquad\qquad \mathcal{R}_i(t_k) = 1, T_i = t_j, \alpha_i, Z_i\}^{\mathcal{R}_i(t_k)} \Big]
\end{aligned}$$

Now conditional probabilities for C_i in (20) may be reexpressed in the form of the right hand side of (16); i.e., for $k = 1, \ldots, j-1$,

$$P\{C_i > t_k | C_i > t_{k-1}, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \leq \ell < k, T_i = t_j, \alpha_i, Z_i\}$$
$$= 1 - P\{C_i = t_k | T_i > t_k, C_i \geq t_k, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \leq \ell < k, T_i = t_j, \alpha_i, Z_i\},$$

where we have used the fact that the events $\{T_i = t_j\}$ and $\{T_i > t_k, T_i = t_j\}$ are the same. This reduces by assumption (16) to $1 - \mu_i(t_k)$. Similarly, conditional probabilities involving $\mathcal{R}_i(t_k)$ and $W_i(t_k)$ reduce to $\pi_i(t_k)$ and $\varphi_i(t_k)$ by assumptions (17) and (18), respectively. The denominator of (19) may be factored in a manner identical to (20), but where $T_i \geq t_j$ replaces $T_i = t_j$ in the leading term and all conditioning sets. It is straightforward

to show that these conditional probabilities involving C_i, $\mathcal{R}_i(t_k)$, and $W_i(t_k)$ are equivalent to $1-\mu_i(t_k)$, $\pi_i(t_k)$ and $\varphi_i(t_k)$ for $k = 1, \ldots, j-1$, respectively. For example, consider $P\{C_i > t_k | C_i > t_{k-1}, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \le \ell < k, T_i \ge t_j, \alpha_i, Z_i\} = 1 - P\{C_i = t_k | C_i \ge t_k, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \le \ell < k, T_i \ge t_j, \alpha_i, Z_i\}$. We have

$$\begin{aligned} &P\{C_i = t_k | C_i \ge t_k, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \le \ell < k, T_i \ge t_j, \alpha_i, Z_i\} \\ &\quad = \frac{P\{C_i = t_k, T_i \ge t_j | C_i \ge t_k, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \le \ell < k, \alpha_i, Z_i\}}{P\{T_i \ge t_j | C_i \ge t_k, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \le \ell < k, \alpha_i, Z_i\}}. \end{aligned} \tag{21}$$

Defining $\mathcal{A}_{ik} = \{\mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \le \ell < k\}$, the numerator in (21) may be written as

$$\sum_{\ell=j}^{M} P(C_i = t_k | C_i \ge t_k, T_i = t_\ell, \mathcal{A}_{ik}, \alpha_i, Z_i) P(T_i = t_\ell | C_i \ge t_k, \mathcal{A}_{ik}, \alpha_i, Z_i) \tag{22}$$

Because $P(C_i = t_k | C_i \ge t_k, T_i = t_\ell, \mathcal{A}_{ik}, \alpha_i, Z_i) = P(C_i = t_k | T_i > t_k, C_i \ge t_k, T_i = t_\ell, \mathcal{A}_{ik}, \alpha_i, Z_i) = P(C_i = t_k | T_i > t_k, C_i \ge t_k, \mathcal{A}_{ik}, \alpha_i, Z_i)$ for $\ell > k$ by assumption (16), it follows that (22) becomes

$$\begin{aligned} &P(C_i = t_k | T_i > t_k, C_i \ge t_k, \mathcal{A}_{ik}, \alpha_i, Z_i) \sum_{\ell=j}^{M} P(T_i = t_\ell | C_i \ge t_k, \mathcal{A}_{ik}, \alpha_i, Z_i) \\ &\quad = P(C_i = t_k | T_i > t_k, C_i \ge t_k, \mathcal{A}_{ik}, \alpha_i, Z_i) P(T_i \ge j | T_i > t_k, C_i \ge t_k, \mathcal{A}_{ik}, \alpha_i, Z_i); \end{aligned}$$

substitution in the right hand side of (21) reduces it to

$$P\{C_i = t_k | T_i > t_k, C_i \ge t_k, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \le \ell < k, \alpha_i, Z_i\} = \mu_i(t_k),$$

as required. Applying these results to the quotient in (19), note that $P\{W_i(t_0) = w_0 | \alpha_i, Z_i\}$ and all other terms cancel, and (19) reduces to

$$\frac{P(T_i = t_j | \alpha_i, Z_i)}{P(T_i \ge t_j | \alpha_i, Z_i)} = P(T_i = t_j | T_i \ge t_j, \alpha_i, Z_i),$$

which yields (14), as claimed.

This argument shows that the assumption (14), usually stated directly without comment, follows naturally from assumptions that the censoring, timing, and measurement processes

depend only on the observable history and latent random effects and not additionally on the unobserved future event time itself, as formalized in (16)–(18). Note that this formulation allows for otherwise complicated dependencies, so that (14) may hold under rather general conditions. Of course, to obtain (7), one must make the additional assumption that the probabilities $\mu_i(t_j)$ and $\pi_i(t_j)$ do not depend on α_i, as noted above.

4 Recent approaches

In this section we review two of our recent proposals for inference in joint models; details may be found in the cited references. As in any model involving unobservable random effects, concern over whether a parametric (normality) assumption on the α_i may be too restrictive and may potentially compromise inference if incorrect leads to interest in models and methods that relax this assumption. The proposals discussed here take different approaches to this issue. The conditional score approach, discussed next, also involves potentially different assumptions from those discussed for the usual likelihood specification, as we demonstrate in Section 4.1.

4.1 Conditional score

A drawback of the approaches based on a likelihood formulation is the ensuing computational complexity involved in maximizing an expression like (7) in Ω. Tsiatis and Davidian (2001) considered a joint model defined by (1), (4) and (6) and proposed an alternative approach to inference on γ and η in (1) that is relatively simple to implement, yields consistent and asymptotically normal estimators, and moreover makes no distributional assumption on the underlying random effects α_i. The approach exploits the conditional score idea of Stefanski and Carroll (1987) for generalized linear models with measurement error, which suggests, in our context, unbiased estimating equations for γ and η based on treating the α_i as "nuisance

parameters" and conditioning on an appropriate "sufficient statistic," as we now outline.

The assumptions under which the approach is valid were not specified carefully by Tsiatis and Davidian (2001); here, we are more precise. Define $t_i(u) = \{t_{ij}, t_{ij} < u\}$, let $\tilde{W}_i(u)$ and $\tilde{e}_i(u)$ be the corresponding vectors of $W_i(t_{ij})$ and $e_i(t_{ij})$ in (6) at times in $t_i(u)$, and let $m_i(u)$ be the length of $e_i(u)$. Tsiatis and Davidian (2001) took $t_i(u)$ to involve $t_{ij} \leq u$; however, to facilitate the discrete-time argument given below, we must be more careful to use a strict inequality; this distinction is of little consequence in continuous time (or in practice). Assume that the event time hazard satisfies

$$\begin{aligned}
\lambda_i(u) &= \lim_{du \to 0} P\{u \leq T_i < u + du | T_i \geq u, C_i \geq u, t_i(u), \tilde{W}_i(u), \alpha_i, Z_i) \\
&= \lim_{du \to 0} P\{u \leq T_i < u + du | T_i \geq u, C_i \geq u, t_i(u), \tilde{e}_i(u), \alpha_i, Z_i) \\
&= \lim_{du \to 0} P\{u \leq T_i < u + du | T_i \geq u, \alpha_i, Z_i) = \lambda_0(u) \exp\{\gamma X_i(u) + \eta^T Z_i\}. \qquad (23)
\end{aligned}$$

This is the continuous-time analog to (14), so holds under (16)–(18); moreover, the conditioning set may be expressed equivalently in terms of $\tilde{W}_i(u)$ or $\tilde{e}_i(u)$, as it also contains α_i.

We now describe the conditional score approach. For definiteness, consider the simple linear model (3), and assume for the moment that σ^2 is known. Let $\hat{X}_i(u)$ be the ordinary least squares estimator for $X_i(u)$ based on the data $\{\tilde{W}_i(u), t_i(u)\}$, that is, based on the longitudinal measurements for i up to time u. Thus $\hat{X}_i(u) = (1, u)\hat{\alpha}_i(u)$, where $\hat{\alpha}_i(u) = \{\mathcal{T}_i^T(u)\mathcal{T}_i(u)\}^{-1}\mathcal{T}_i^T(u)\tilde{W}_i(u)$, and $\mathcal{T}_i(u)$ is the usual $\{m_i(u) \times 2\}$ design matrix with first column all ones and second column $t_i(u)$. Note that $\hat{\alpha}_i(u)$ and hence $\hat{X}_i(u)$ are defined only if there are at least two measurements prior to u. Define $Y_i(u) = I(V_i \geq u, t_{i2} \leq u)$. Assume that the distribution of "error" $e_i(t_{ij})$ at time t_{ij}, given that a measurement is taken at t_{ij}, i is at risk at t_{ij}, the measurement history prior to t_{ij}, α_i, and Z_i, is $\mathcal{N}(0, \sigma^2)$. In the discrete-time representation of Section 3, this corresponds to assumption (15). This is in contrast to (1) of Tsiatis and Davidian (2001). As we demonstrate below, it may be shown that, under these

conditions and certain assumptions regarding the censoring and timing processes,

$$(\hat{X}_i(u)|Y_i(u) = 1, t_i(u), \alpha_i, Z_i) \sim \mathcal{N}\{X_i(u), \sigma^2\theta_i(u)\}, \tag{24}$$

where $\sigma^2\theta_i(u)$ is the usual variance of the predicted value $\hat{X}_i(u)$, which depends on $t_i(u)$ (given in Tsiatis and Davidian (2001), p. 449).

Define $dN_i(u) = I(u \leq V_i < u + du, \Delta_i = 1, t_{i2} \leq u)$, which puts point mass at time u for an observed event time after the second longitudinal measurement on i. The motivation for the conditional score estimating equations relies on identifying a "sufficient statistic" for α_i. At time u, given i is at risk, the conditional density for $\{dN_i(u) = r, \hat{X}_i(u) = x\}$ is

$$\begin{aligned} &\text{pr}\{dN_i(u) = r, \hat{X}_i(u) = x|Y_i(u) = 1, t_i(u), \alpha_i, Z_i\} \qquad (25)\\ &\quad = \text{pr}\{dN_i(u) = r|\hat{X}_i(u) = x, Y_i(u) = 1, t_i(u), \alpha_i, Z_i\}\text{pr}\{\hat{X}_i(u) = x|Y_i(u) = 1, t_i(u), \alpha_i, Z_i\}. \end{aligned}$$

From (23), the first term on the right hand side of (25) is a Bernoulli density with probability $\lambda_0(u)\, du \exp\{\gamma X_i(u) + \eta^T Z_i\}$, and the second term is, from (24), the $\mathcal{N}\{X_i(u), \sigma^2\theta_i(u)\}$ density. Substituting these expressions into (25) and simplifying shows that, to order du, (25) is

$$\exp\left[X_i(u)\left\{\frac{\gamma\sigma^2\theta_i(u)dN_i(u) + \hat{X}_i(u)}{\sigma^2\theta_i(u)}\right\}\right] \frac{\{\lambda_0(u)\exp(\eta^T Z_i)du\}^{dN_i(u)}}{\{2\pi\sigma^2\theta_i(u)\}^{1/2}} \exp\left\{-\frac{\hat{X}_i^2(u) + X_i^2(u)}{2\sigma^2\theta_i(u)}\right\},$$

from which it follows that a "sufficient statistic" for α_i is $S_i(u, \gamma, \sigma^2) = \gamma\sigma^2\theta_i(u)dN_i(u) + \hat{X}_i(u)$.

Tsiatis and Davidian (2001) suggested that conditioning on $S_i(u, \gamma, \sigma^2)$ would remove the dependence of the conditional distribution on (the "nuisance parameter") α_i. In particular, they observed that the conditional intensity process

$$\begin{aligned} &\lim_{du\to 0} du^{-1}\text{pr}\{dN_i(u) = 1|S_i(u, \gamma, \sigma^2), Z_i, t_i(u), Y_i(u)\} \qquad (26)\\ &\quad = \lambda_0(u)\exp\{\gamma S_i(u, \gamma, \sigma^2) - \gamma^2\sigma^2\theta_i(u)/2 + \eta^T Z_i\}Y_i(u) = \lambda_0(u)E^*_{0i}(u, \gamma, \eta, \sigma^2). \qquad (27) \end{aligned}$$

Based on (26), they thus proposed the following estimating equations for γ and η based on equating "observed" and "expected" quantities in a spirit similar to such a derivation for the

usual partial likelihood score equations:

$$\sum_{i=1}^{n} \int \{S_i(u,\gamma,\sigma^2), Z_i^T\}^T \{dN_i(u) - E_{0i}^*(u,\gamma,\eta,\sigma^2)\lambda_0(u)du\} = 0$$
$$\sum_{i=1}^{n} \{dN_i(u) - E_{0i}^*(u,\gamma,\eta,\sigma^2)\lambda_0(u)du\} = 0. \tag{28}$$

With $dN(u) = \sum_{j=1}^n dN_j(u)$ and $E_0^*(u,\gamma,\eta,\sigma^2) = \sum_{j=1}^n E_{0j}^*(u,\gamma,\eta,\sigma^2)$, the second equation yields $\hat{\lambda}_0(u)du = dN(u)/E_0^*(u,\gamma,\eta,\sigma^2)$. By substitution of $\hat{\lambda}_0(u)du$ in the first equation in (28), one obtains the conditional score estimating equation for γ and η given by

$$\sum_{i=1}^{n} \int \left[\{S_i(u,\gamma,\sigma^2), Z_i^T\}^T - \frac{E_1^*(u,\gamma,\eta,\sigma^2)}{E_0^*(u,\gamma,\eta,\sigma^2)} \right] dN_i(u) = 0, \tag{29}$$

where $E_{1j}^*(u,\gamma,\eta,\sigma^2) = \{S_j(u,\gamma,\sigma^2), Z_j^T\}^T E_{0j}^*(u,\gamma,\eta,\sigma^2)$, and $E_1^*(u,\gamma,\eta,\sigma^2) = \sum_{j=1}^n E_{1j}^*(u,\gamma,\eta,\sigma^2)$.

As σ^2 is unknown, Tsiatis and Davidian (2001) proposed an additional estimating equation for σ^2 based on residuals from individual least squares fits to the m_i measurements for each i and gave arguments that indicate that the resulting estimators for γ and η are consistent and asymptotically normal under assumptions (23) and (24), with standard errors that may be derived based on the usual sandwich approach. An advantage of the procedure is that $\hat{\gamma}$ and $\hat{\eta}$ solving (29) are relatively easy to compute; although technically (29) may have multiple solutions, identifying a consistent solution is not problem in practice. The equation (29) reduces to the partial likelihood score equations when $\sigma^2 = 0$ (so $\hat{X}_i(u) = X_i(u)$).

The foregoing developments depend critically on assumption (24). We now elucidate conditions under which (24) holds using a discrete time representation as in Section 3. Partition time up to u as $0 = t_0 < t_1 < \cdots < t_{M_u} = u$. For an individual at risk at time u, i.e., with $T_i \geq u, C_i \geq u$ and at least two measurements, denote the j_i times at which measurements are taken as $t_{i1} < \cdots < t_{ij_i} < u$. Then, to demonstrate (24), it suffices to show that

$$\{e_i(t_{i1}), \ldots, e_i(t_{ij_i}) | T_i \geq u, C_i \geq u, \mathcal{R}_i(t_{i1}) = \cdots = \mathcal{R}_i(t_{ij_i}) = 1, \mathcal{R}_i(t_r) = 0,$$
$$\text{for } t_r, r = 1, \ldots, M \text{ with } t_r \neq t_{i1}, \ldots, t_{ij_i}, \alpha_i, Z_i\} \sim \mathcal{N}(0, \sigma^2 I_{j_i}), \tag{30}$$

where I_j is a $(j \times j)$ identity matrix. As before, let $\mathcal{A}_{ik} = \{\mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \leq \ell < k\}$. In the following arguments, $\mathcal{A}_{ik}$ always appears in conditioning sets with α_i, so that, under these conditions, $\mathcal{A}_{ik}$ may be written equivalently as $\{\mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)e_i(t_\ell), 0 \leq \ell < k\}$. Define $\mathcal{A}^*_{ik} = \{\mathcal{R}_i(t_\ell), 0 \leq \ell < k\}$. Then the conditional density (30) may be written as a quotient with numerator

$$\prod_{k=1}^{M-1} P(C_i > t_k | T_i \geq u, C_i \geq t_k, \mathcal{A}_{ik}, \alpha_i, Z_i) P\{\mathcal{R}_i(t_k) = r_k | T_i \geq u, C_i > t_k, \mathcal{A}_{ik}, \alpha_i, Z_i\}$$
$$\times \prod_{q=1}^{j_i} P\{e_i(t_{iq}) = e_{iq} | T_i \geq u, C_i > t_{iq}, \mathcal{A}_{iq}, \mathcal{R}_i(t_{iq}) = 1, \alpha_i, Z_i\} \qquad (31)$$

and denominator

$$\prod_{k=1}^{M-1} P(C_i > t_k | T_i \geq u, C_i \geq t_k, \mathcal{A}^*_{ik}, \alpha_i, Z_i) P\{\mathcal{R}_i(t_k) = r_k | T_i \geq u, C_i > t_k, \mathcal{A}^*_{ik}, \alpha_i, Z_i\} \qquad (32)$$

From (15), we have $\prod_{q=1}^{j_i} P\{e_i(t_{iq}) = e_{iq} | T_i \geq u, C_i > t_{iq}, \mathcal{A}_{iq}, \mathcal{R}_i(t_{iq}) = 1, \alpha_i, Z_i\} = \prod_{q=1}^{j_i} (2\pi\sigma^2)^{-1/2} \exp\{-e_{iq}^2)/(2\sigma^2)\}$, which is the distribution in (30); thus, the result follows if the remaining terms in the numerator (31) and (32) cancel. This is the case if

$$P(C_i = t_k | T_i \geq u, C_i \geq t_k, \mathcal{A}_{ik}, \alpha_i, Z_i) = P(C_i = t_k | T_i \geq u, C_i \geq t_k, \mathcal{A}^*_{ik}, \alpha_i, Z_i) \qquad (33)$$

$$P\{\mathcal{R}_i(t_k) = r_k | T_i \geq u, C_i > t_k, \mathcal{A}_{ik}, \alpha_i, Z_i\} = P\{\mathcal{R}_i(t_k) = r_k | T_i \geq u, C_i > t_k, \mathcal{A}^*_{ik}, \alpha_i, Z_i\}. \qquad (34)$$

In words, (33) and (34) state that the probabilities of censoring and measurement at time t_k, conditional on the underlying α_i and past measurements, cannot depend on past "errors."

These sufficient conditions may be contrasted with those that underlie the usual likelihood formulation (7), i.e, that $\mu_i(t_j) = P\{C_i = t_j | T_i > t_j, C_i \geq t_j, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \leq \ell < j, \alpha_i, Z_i\}$ and $\pi_i(t_j) = P\{\mathcal{R}_i(t_j) = 1 | T_i > t_j, C_i > t_j, \mathcal{R}_i(t_\ell), \mathcal{R}_i(t_\ell)W_i(t_\ell), 0 \leq \ell < j, \alpha_i, Z_i\}$ may not depend on α_i. For the conditional score, dependence of the timing and censoring processes on the α_i that dictate the "inherent" smooth trend is permitted, but these processes may not depend further on past errors (or, equivalently, past observations). For the likelihood approach, the conditions are "reversed." Thus, belief about the nature of these phenomena

may dictate the choice of approach. Note that neither assumption subsumes the other in general, and, as these are only sufficient conditions, it may be possible to identify assumptions that validate both formulations. Of course, if censoring and timing are thought not to depend on α_i or the past measurement history, as might be the case in a study where the censoring is administrative and timing of measurements adheres to protocol, then the conditions for either approach are satisfied.

4.2 Semiparametric likelihood

Song, Davidian, and Tsiatis (2002b) proposed an alternative, likelihood-based approach for the situation where the usual normality assumption on the α_i is relaxed. These authors considered the joint model given by (1), (4), and (6), but assumed only that α_i have conditional density $p(\alpha_i|Z_i;\delta)$ in a class of "smooth" densities $\mathcal{H}$ studied by Gallant and Nychka (1987), where densities in $\mathcal{H}$ do not have jumps, kinks or oscillations, but may be skewed, multi-modal, fat- or thin-tailed relative to the normal (and the normal is $\in \mathcal{H}$). In particular, they represent α_i as $\alpha_i = g(\mu, Z_i) + Ru_i$, where g is a regression function with parameter μ, R is a lower triangular matrix, and u_i has density $h \in \mathcal{H}$. Gallant and Nychka (1987) showed that densities in $\mathcal{H}$ may be approximated by a truncated Hermite series expansion of the form

$$h_K(z) = P_K^2(z)\varphi_q(z), \tag{35}$$

where $P_K(z)$ is a Kth order polynomial, e.g., for $K = 2$, $P_K(z) = a_{00} + a_{10}z_1 + a_{01}z_2 + a_{20}z_1^2 + a_{02}z_2^2 + a_{11}z_1z_2$; the vector of coefficients a must satisfy $\int h_K(z)\,dz = 1$ and an identifiability constraint; and $\varphi_q(z)$ is the q-variate standard normal density. Under these conditions, $K = 0$ yields the standard normal density; larger values of K allow departures from normality, so that K acts as a tuning parameter that controls the degree of flexibility in representing the true density. Gallant and Nychka (1987) termed this approach to representation and estimation

of a density "SemiNonParametric" (SNP) to emphasize that (35) acts like a nonparametric estimator but may be characterized by a finite-dimensional set of parameters for fixed K.

Song, Davidian and Tsiatis (2002b) proposed inference on $\Omega = \{\lambda_0(\cdot), \gamma, \eta, \sigma^2, \delta\}$ based on the likelihood (7), where now δ includes μ, the elements of R, and a; thus, the arguments of Section 3 clarify the assumptions underlying the approach. Following Davidian and Gallant (1993) and Zhang and Davidian (2001), they advocated choosing K by inspection of measures such as Akaike's information criterion (AIC). Implementation for fixed K is via an EM algorithm and is considerably more complex than that for the conditional score; see Song, Davidian and Tsiatis (2002b).

5 Simulation evidence

We offer a brief empirical comparison of the conditional score and semiparametric methods under a scenario for which the sufficient assumptions on censoring and timing of measurements are satisfied for both. The situation was based on an HIV clinical trial. For each of 200 Monte Carlo data sets, a sample of size $n = 200$ was generated assuming the joint model defined by (1) with $\eta = 0$, (3), and (6), with $\gamma = -1.0$, $\sigma^2 = 0.60$, and $\lambda_0(u) = 1$. Measurements were taken to follow a nominal time schedule of $t_{ij} = (0, 2, 4, 8, 16, 24, 32, 40, 48, 56, 64, 72, 80)$, where measurements at any of these times were missing with constant probability 0.10, and, for each subject i, C_i was generated independently of all other variables as exponential with mean 110. Thus, the censoring and timing processes satisfy the assumptions sufficient for both methods. With $E(\alpha_i) = (4.173, -0.0103)^T$ and $\{\text{var}(\alpha_{i0}), \text{cov}(\alpha_{0i}, \alpha_{i1}), \text{var}(\alpha_{i1})\} = \{4.96, -0.0456, 0.012\}$, we considered two true distributions for the α_i: (i) α_i distributed as a symmetric, bimodal, bivariate mixture of normals, and (ii) α_i distributed as bivariate normal. For each data set, γ was estimated five ways: using the "ideal" approach, where $X_i(u)$ is known for all u and γ is estimated by maximizing (2), denoted as I; maximizing (2)

with $X_i(u)$ imputed using LVCF; solving the conditional score equation (CS); maximizing (7) assuming normal α_i (so with $K = 0$ in the approach of Section 4.2); and maximizing (7) with $\alpha_i = \mu + Ru_i$ with u_i having SNP density representation (35) for $K = 1, 2, 3$ and K chosen objectively by inspection of the AIC. The "ideal" methods serves as an albeit unachievable benchmark against which the others may be compared. For each method, standard errors were obtained from the usual partial likelihood expression (I, LVCF), using the sandwich method (CS), or based on the maximized likelihood ($K = 0$, SNP) as described by Song, Davidian and Tsiatis (2002b). Nominal 95% Wald confidence intervals were constructed as the estimate ± 1.96 times standard error estimate in each case.

Some interesting features are notable from the results, given in Table 1. The LVCF estimator exhibits considerable bias in both cases, which leads to confidence intervals that fail seriously to achieve the nominal level. Confidence intervals for the other methods yield acceptable performance. Like the ideal procedure, the conditional score and likelihood methods yield approximately unbiased estimators in both cases; note that assuming α_i to be normal under the mixture scenario still results in unbiased inference. The conditional score estimator is inefficient relative to the likelihood approach in both scenarios; this is not unexpected, as the conditional score approach does not exploit the full information in the longitudinal data. Interestingly, the estimation of γ assuming normal α_i under the mixture scenario does not compromise efficiency relative to relaxing this assumption. This phenomenon, along with the unbiasedness, which we have noted in other situations not reported here, suggests a sort of "robustness" of the likelihood approach with α_i taken to be normal to departures from this assumption, see Section 6. Allowing more flexibility for this distribution than is required when the α_i are normal (case (ii)) does not result in efficiency loss; here, the information criterion chose $K = 0$ correctly for most of the 200 data sets. We have carried out extensive simulations of the likelihood approach under different scenarios, including taking the true random effects

distribution to be discrete with mass at a few points but assuming normality, regardless of scenario, and the "robustness" phenomenon persists. We believe that this feature deserves further investigation.

Overall, these simulations, and others we have carried out (Tsiatis and Davidian (2001); Song, Davidian and Tsiatis (2002ab)) indicate that both the conditional score and likelihood approaches using (7) lead to sound inferences on hazard parameters when the assumptions under which the methods are valid do indeed hold.

6 Discussion

Some critical issues for practice are that naive approaches to inference on relationships between longitudinal and time-to-event data are inappropriate, and approximate methods based on a joint model framework may not completely eliminate bias and may be inefficient. Methods based on a joint model likelihood formulation may yield most precise inferences, but can be computationally demanding. The conditional score approach is by comparison easier to compute, but at the expense of loss of efficiency. An interesting finding that deserves further, formal investigation is the apparent robustness of the likelihood approach for estimation of hazard parameters assuming normal random effects to deviations from this assumption. Another critical issue, mentioned only briefly, is the assumption on the distribution of α_i given Z_i. Although it is routine to assume this is independent of Z_i, if this assumption is incorrect, biased inference will result (e.g., Heagerty and Kurland (2001)). An advantage of the conditional score approach is that it is valid without requiring any assumption on this distribution.

There is considerable recent work (e.g., Zhang, Lin, Raz, and Sowers (1998)) on expressing splines as random effects models, connected with the work of Wahba (1990) relating smoothing splines to a stochastic process. It may be possible to exploit this ability to write the stochastic

process in a model like (5) as a "random effects model" and thereby develop new joint model implementations; this is a subject for future research.

An important consideration that we believe has been inadequately highlighted in the literature is the precise nature of assumptions that validate specification of a likelihood or estimating equations in these approaches. Assumptions on the processes leading to censoring and timing of longitudinal measurements play a central role. Research on the consequences for inference of violations of the usual assumptions on these processes would be valuable.

Throughout this presentation, we have regarded $X_i(u)$ as a scalar. In some settings, several such longitudinal responses are collected (e.g., CD4, CD8, and viral load in HIV studies), and interest focuses on the relationship between the multivariate longitudinal process and the time-to-event. The framework discussed herein may be extended to this situation (e.g., Xu and Zeger (2001b); Song, Davidian and Tsiatis (2002a), who adapt the conditional score to this setting); however, the modeling considerations become more complex, involving the joint distribution of several longitudinal processes. Similarly, rather than a single time-to-event, some studies may involve a sequence of event times (e.g., recurrent events); Henderson, Diggle, and Dobson (2000) discuss modifications for this setting. In both of these extensions, issues regarding assumptions and implementation are the same, albeit more complex.

Acknowledgements

This work was supported by National Institutes of Health grants R01-AI31789, R01-CA51962, and R01-CA085848.

References

Bycott, P. and Taylor, J. (1998). A comparison of smoothing techniques for CD4 data measured with error in a time-dependent Cox proportional hazards model. *Statistics in Medicine* **17**, 2061–77

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

Brown, E.R. and Ibrahim, J.G. (2003b). Bayesian approaches to joint cure rate and longitudinal models with applications to cancer vaccine trials. *Biometrics* **59**, in press.

Cox, D.R. (1972). Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–200.

Cox, D.R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Dafni, U.G. and Tsiatis, A.A. (1998). Evaluating surrogate markers of clinical outcome measured with error. *Biometrics* **54**, 1445–62.

Davidian, M. and Gallant, A.R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika* **80**, 475–88.

DeGruttola, V. and Tu, X.M. (1994). Modeling progression of CD-4 lymphocyte count and its relationship to survival time. *Biometrics* **50**, 1003–14.

Faucett, C.J. and Thomas, D.C. (1996). Simultaneously modeling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Statistics in Medicine* **15**, 1663–85.

Faucett, C.L., Schenker, N., and Taylor, J.M.G. (2002). Survival analysis using auxiliary

variables via multiple imputation, with application to AIDS clinical trials data. *Biometrics* **58**, 37–47.

Gallant, A. R. and Nychka, D. W. (1987). Seminonparametric maximum likelihood estimation. *Econometrica* **55**, 363–390

Heagerty, P.J. and Kurland, B.F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* **88**, 973–985.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modeling of longitudinal measurements and event time data. *Biostatistics* **4**, 465–480.

Hogan, J.W. and Laird, N.M. (1997a). Mixture models for the joint distributions of repeated measures and event times. *Statistics in Medicine* **16**, 239–257.

Hogan, J.W. and Laird, N.M. (1997b). Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine* **16**, 259–272.

Ibrahim, J.G., Chen, M.H., and Sinha, D. (2001). *Bayesian Survival Analysis.* Springer, New York.

Kalbfleisch, J.D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data, Second Edition.* John Wiley, New York.

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

Lavalley, M.P. and DeGruttola, V. (1996). Model for empirical Bayes estimators of longitudinal CD4 counts. *Statistics in Medicine* **15**, 2289–2305.

Law, N.J., Taylor, J.M.G., and Sander, H. (2002). The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics*

3, 547–563.

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.

Pawitan, Y. and Self, S. (1993). Modeling disease marker processes in AIDS. *Journal of the American Statistical Association* **83**, 719–26.

Prentice, R. (1982). Covariate measurement errors and parameter estimates in a failure time regression model. *Biometrika* **69**, 331–42.

Prentice, R. (1989). Surrogate endpoints in clinical trials: Definition and operation criteria. *Statistics in Medicine* **8**, 431–440.

Raboud, J., Reid, N., Coates, R.A., and Farewell, V.T. (1993). Estimating risks of progressing to AIDS when covariates are measured with error. *Journal of the Royal Statistical Society, Series A* **156**, 396–406.

Schluchter, M.D. (1992). Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine* **11**, 1861–1870.

Self, S. and Pawitan, Y. (1992). Modeling a marker of disease progression and onset of disease. In *AIDS Epidemiology: Methodological Issues* (N.P. Jewell, K. Dietz, and V.T. Farewell, eds.). Birkhäuser, Boston.

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

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

Stefanski, L.A. and Carroll, R.J. (1987). Conditional scores and optimal scores in generalized linear measurement error models. *Biometrika* **74**, 703–16.

Taylor, J.M.G., Cumberland, W.G., and Sy, J.P. (1994). A stochastic model for analysis of longitudinal data. *Journal of the American Statistical Association* **89**, 727–76.

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.

Tsiatis, A.A. and Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, **88**, 447–58.

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia.

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

Wu, M.C. and Carroll, R.J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* **44**, 175–188.

Wulfsohn, M.S. and Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330–9.

Xu, J. and Zeger, S.L. (2001a). Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics* **50**, 375–87.

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

Zhang, D. and Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics* **57**, 795–802.

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association* **93**, 710–719.

Table 1: Simulation results for 200 data sets for estimation of γ by the "ideal" approach where $X_i(u)$ is known (I), last value carried forward (LVCF), conditional score (CS), and the semiparametric likelihood approach with normal random effects ($K = 0$) and with K chosen using AIC (SNP). Entries are Monte Carlo average of estimates (Mean), Monte Carlo standard deviation (SD), Monte Carlo average of estimated standard errors (SE), and Monte Carlo coverage probability of nominal 95% Wald confidence interval. (True value is $\gamma = 1.0$).

	I	LVCF	CS	$K = 0$	SNP
			Case (i): Mixture Scenario		
Mean	-1.02	-0.82	-1.05	-1.01	-1.01
SD	0.09	0.07	0.17	0.10	0.10
SE	0.09	0.07	0.14	0.11	0.11
CP	0.96	0.31	0.94	0.97	0.95
			Case (i): Normal Scenario		
Mean	-1.01	-0.81	-1.04	-1.00	-1.00
SD	0.08	0.07	0.15	0.10	0.10
SE	0.08	0.07	0.13	0.10	0.10
CP	0.96	0.25	0.93	0.96	0.95