Published in final edited form as:

Lifetime Data Anal. 2019 April ; 25(2): 229–258. doi:10.1007/s10985-018-9434-7.

New Approaches for Censored Longitudinal Data in Joint Modelling of Longitudinal and Survival Data, with Application to HIV Vaccine Studies

Tingting Yu,

Department of Statistics, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada

Lang Wu, and

Department of Statistics, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada

Peter Gilbert

Department of Biostatistics, University of Washington, Seattle, WA 98195-7232, USA

Vaccine and Infectious Disease and Public Health Sciences Divisions, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

Abstract

In HIV vaccine studies, longitudinal immune response biomarker data are often left-censored due to lower limits of quantification of the employed immunological assays. The censoring information is important for predicting HIV infection, the failure event of interest. We propose two approaches to addressing left censoring in longitudinal data: one that makes no distributional assumptions for the censored data – treating left censored values as a “point mass” subgroup – and the other makes a distributional assumption for a subset of the censored data but not for the remaining subset. We develop these two approaches to handling censoring for joint modelling of longitudinal and survival data via a Cox proportional hazards model fit by h-likelihood. We evaluate the new methods via simulation and analyze an HIV vaccine trial data set, finding that longitudinal characteristics of the immune response biomarkers are highly associated with the risk of HIV infection.

Keywords

Lower limit of quantification; H-likelihood; Shared-parameter model; Mixed-effect model; Cox model

1 Introduction

In the analysis of complex longitudinal data, a common challenge in applications is that some longitudinal data are censored due to instrument detection or quantification limits. For example, in HIV vaccine studies biomarkers measuring immune responses to vaccination are sometimes left-censored below the lower limit of quantification (LLOQ) of the employed

immunological assays, and the censoring may have important implications for the efficacy of the vaccine. Moreover, in joint modelling of longitudinal and survival data, the longitudinal trajectories may be associated with events of interest such as HIV infection, and censoring is an important characteristic of the longitudinal trajectories. Similarly, for anti-HIV treatments, left-censored viral loads (viral suppression) may reflect the efficacy of the treatments and may be associated with the risk of AIDS or death. Therefore, appropriately addressing censored longitudinal data is an important research question.

There has been some literature on modelling longitudinal data with censoring, especially in AIDS studies (e.g., Hughes (1999); Jacqmin-Gadda et al. (2000); Wu (2002); Dagne and Huang (2013); Thiébaud et al. (2005); Yu et al. (2017)). A common approach for censored data is to assume that the censored (unobserved) longitudinal data follow the same distribution assumed for the uncensored (observed) data, such as a normal distribution. A likelihood method is then typically used for inference where the probability of censoring is incorporated into the likelihood based on the distributional assumption. There are several issues with this approach: (i) from a biological or clinical point of view, it may be unreasonable to assume that the censored values below the LLOQ follow the same distribution as the observed values that are above LLOQ since these two may reflect quite different biological or clinical mechanisms; (ii) from a statistical point of view, it may be unreasonable to assume that the left-censored values follow a normal distribution which is typically assumed for the observed data (or their transformed versions) since the censored values are very small and cannot be negative; and (iii) the distributional assumption for censored values cannot be verified. Therefore, the common approach in the literature may be convenient for statistical inference but may not be realistic in practice.

Moulton and Halsey (1995) and Lachenbruch (2002) considered a set of mixture (or two-part) models for continuous responses with excessive left-censored data, and Berk (2002) extended their approaches to mixed-effects models for repeated measurements. Their models contained two components: a continuous distribution for the observed data and a subset of the censored data and a point mass distribution for the rest of the censored data, and a logistic regression model is used to model the two components. However, their method has not been considered in the joint model setting. Yu et al. (2017) proposed an approach that does not assume any distribution for the censored values, by treating left censored values as a concentrated subgroup. This approach is advantageous for studying censoring status as a correlate of events of interest, but it may lose some information if a subset of the censored values follows the same distribution as the observed data.

In this article, we propose a new approach, called a “two-regime model”, to address censored longitudinal data in the context of joint models for longitudinal and survival data. The idea is similar to the method of Moulton and Halsey (1995) but is used in different context. For censored data, we assume that (a) a portion of values continue to follow the distribution assumed for the observed (uncensored) data, especially those close to the LLOQ, which may be reasonable since the same biological mechanism may still govern the censored values close to the LLOQ (moreover, the LLOQ may decrease over time due to technology advances); (b) for the remaining censored values, we do not make any distributional assumption, especially for those way below the LLOQ, which may also be

reasonable since these censored values may follow a different biological mechanism (e.g., HIV infected individuals with viral loads way below the LLOQ may be considered healthy).

This approach is particularly reasonable when there is a “biological detection limit”, that is, some data below the biological detection limit are actually observed but they are of no or little biological significance, e.g., with variability dominated by technical assay measurement errors or otherwise deemed to be implausibly associated with risk of the outcome. In preventive HIV vaccine studies, it may be reasonable to assume that the censored immune response data come from two regimes, such as *potentially associated with HIV infection risk* versus *implausibly associated with HIV infection risk*; for studies of HIV infected individuals there is a parallel application where viral load data come from two regimes *non-progressors* versus *progressors* as discussed in Dagne and Huang (2013). Also motivating (b) is the fact that in many applications the variability of biomarker values well below the LLOQ is predominantly due to technical assay measurement error rather than to biological heterogeneity, implying that a model assuming that inter-subject variability of biomarkers well below versus above the LLOQ affect risk of an outcome in a similar way would be incorrect. While in general membership of censored values in category (a) vs. category (b) is unobservable, sometimes other variables are available that can be used to predict these membership statuses based on biological knowledge, which is the approach we take with our newly proposed method.

In Section 2 we present a joint Cox proportional hazards model for longitudinal and survival data, and introduce our newly proposed method for addressing the left-censoring problem, as well as summarizing our previously developed method Yu et al. (2017) that we study as an alternative approach. In Section 4 we analyze an HIV vaccine efficacy trial data set using the different methods. Section 5 evaluates and compares the proposed methods via a comprehensive simulation study, which also compares these methods to two other methods: a *naïve method*, which replaces the left-censored data with the LLOQ, and a *Tobit mixed effects model*, which assumes that all the left censored data follow the same distribution as the uncensored data (Schnedler, 2005; Hughes, 1999). We conclude with some discussions in Section 6. For validity, all of the methods we consider assume that any missing longitudinal data are missing at random (MAR), such that the probability of missing data may depend on the observed data but not on unobserved data. Under the MAR assumption, all of the methods and models we study provide valid inferences in the presence of missing data since MAR is ignorable in likelihood inference and the statistical inference can be carried out based on the observed-data likelihood.

2 Methods for Censored Longitudinal Data in Joint Models

We consider a joint model for longitudinal and survival data, which models the association of the longitudinal data with a failure time of interest. Motivated by an HIV vaccine efficacy trial data set to be analyzed in Section 4, we consider two correlated longitudinal variables of mixed types: a continuous longitudinal variable Y that is subject to left censoring, and a binary longitudinal variable Z that may be associated with Y . Then, we consider a survival model linking the hazard of the failure time to the longitudinal covariates. Our focus will be on the different methods to address the left censoring in the longitudinal data Y .

Let z_{ik} and y_{ij} be the observed values of the longitudinal variables Z and Y for individual i at times t_{ik} and t_{ij} respectively, $k = 1, \dots, m_i$, $j = 1, \dots, n_i$, $i = 1, \dots, n$. The values of y_{ij} 's may be left censored due to a LLOQ of d , i.e., the values $y_{ij} < d$ cannot be observed. Note that we allow the measurement times t_{ik} of Z to be different from the measurement times t_{ij} of Y for each individual i . For the event of interest, let t_i^* be the true event time and $\mathcal{C}_i$ be the right-censoring time for individual i . The observed survival data are $\{(s_i, \delta_i), i = 1, 2, \dots, n\}$, where $s_i = \min(t_i^*, \mathcal{C}_i)$ is the observed event time and $\delta_i = I(t_i^* \leq \mathcal{C}_i)$ is the event indicator. In the following, we first introduce two approaches to handling left censoring in the longitudinal data, and then we describe a joint model for longitudinal and survival data.

2.1 Models for censored longitudinal data

When longitudinal data are left censored, a standard approach in the literature assumes that the censored values follow the same distribution as the observed values. As discussed in the previous section, this approach may be unreasonable from both biological and statistical points of view and can produce biased results if the distributional assumption is violated (Lorimer and Kiermeier, 2007). We introduce two approaches, called the one-regime method and the two-regime method. For the one-regime method previously proposed by Yu et al. (2017), we do not make any distributional assumption for the censored values and view the censored values as a “point mass” subgroup. For the newly proposed two-regime method, we assume that some of the censored values follow the same distribution as the observed data while the remaining values again can be viewed as the “point mass” subgroup. These two approaches are described below.

Approach 1: One-regime method—In this approach, we do not assume any distribution or model for the censored values. Instead we view the censored values as a point mass or a cluster/subgroup of unobserved values below the LLOQ without positing any distributional assumption. Let C_{ij} be a censoring indicator such that $c_{ij} = 1$ when $y_{ij} < d$ and $c_{ij} = 0$ otherwise. It is important to know whether Y_{ij} is censored or not over time in some studies such as HIV/AIDS studies. Note that the value of C_{ij} is determined by the value of Y_{ij} . However, in some studies the Y values may be measured infrequently and thus are unavailable at some time points of interest. In this case, we can use other observed variables to predict the value of C_{ij} at times when the values of Y_{ij} are unavailable but are of interest. Thus, we consider a generalized linear mixed effects (GLME) model for the censoring variable

$$\text{logit}(P(c_{ij} = 1)) = \mathbf{w}_{ij}^T \boldsymbol{\eta} + \mathbf{v}_{ij}^T \mathbf{b}_{1i} \equiv \zeta_{ij}, \quad \mathbf{b}_{1i} \stackrel{iid}{\sim} N(\mathbf{0}, \mathbf{D}_1), \quad (1)$$

where $\mathbf{w}_{ij}$ and $\mathbf{v}_{ij}$ are covariate vectors, $\boldsymbol{\eta}$ are fixed parameters, and $\mathbf{b}_{1i}$ are random effects. When $c_{ij} = 1$, i.e., when the y_{ij} value is below the LLOQ d , we view y_{ij} as a point mass with $f(y_{ij} | c_{ij} = 1) = 1_{(y_{ij} \leq d)}$. The censoring model (1) can be used to predict the censoring status of Y at time points when the Y data are important but not measured. For example, in preventive HIV vaccine efficacy trials, biomarkers measuring immune responses to

vaccination carry information about the risk of HIV infection throughout study follow-up yet these biomarkers may be measured only sparsely. Model (1) helps us to make reasonable predictions. Another advantage of this approach is that it can also be used to model non-ignorable or informative missing data in the longitudinal data. When longitudinal data have both left-censored data and non-ignorable missing data, we may consider two different models (1) for the censoring mechanism and the missingness mechanism, respectively. Here we do not consider the issue of non-ignorable missing data.

For the observed or uncensored y_{ij} values (i.e., when $c_{ij}=0$ or $y_{ij} > d$), we may consider a LME model with a truncated normal error

$$y_{ij} \mid y_{ij} \geq d = \boldsymbol{\psi}_{ij}^T \boldsymbol{\beta} + \boldsymbol{\phi}_{ij}^T \mathbf{b}_{2i} + e_{ij}, \quad \mathbf{b}_{2i} \stackrel{iid}{\sim} N(\mathbf{0}, \mathbf{D}_2), \quad e_{ij} \stackrel{iid}{\sim} N^*(0, \sigma^2), \quad (2)$$

where $\boldsymbol{\psi}_{ij}$ and $\boldsymbol{\phi}_{ij}$ are covariate vectors, $\boldsymbol{\beta}$ are fixed parameters, $\mathbf{b}_{2i}$ are random effects, and e_{ij} denotes within-individual error following a truncated normal distribution as described below. The conditional density function of y_{ij}/c_{ij} can be written as

$$f_{y|c} = f(y_{ij} \mid c_{ij} = 0, \mathbf{b}_{2i})^{1-c_{ij}} f(y_{ij} \mid c_{ij} = 1)^{c_{ij}} \quad (3)$$

where

$$f(y_{ij} \mid c_{ij} = 0, \mathbf{b}_{2i}) = \frac{f(y_{ij} \mid \mathbf{b}_{2i})}{1 - F(d \mid \mathbf{b}_{2i})}, \quad F(d \mid \mathbf{b}_{2i}) = 1 - \int_d^\infty f(y_{ij} \mid \mathbf{b}_{2i}) dy_{ij}, \quad f(y_{ij} \mid \mathbf{b}_{2i}) = \frac{1}{\sigma} \phi\left(\frac{y_{ij} - (\boldsymbol{\psi}_{ij}^T \boldsymbol{\beta} + \boldsymbol{\phi}_{ij}^T \mathbf{b}_{2i})}{\sigma}\right),$$

and $\phi(\cdot)$ is the probability density of the standard normal distribution. For the one-regime method, we choose $f(y_{ij}/c_{ij} = 1) = 1_{(y > d)}$. The contribution to the likelihood of the observed data (y_{ij}, c_{ij}) given the random effects is

$$f(y_{ij}, c_{ij} \mid \mathbf{b}_{1i}, \mathbf{b}_{2i}) = f_{y|c} \times f_c = \left[\frac{f(y_{ij} \mid \mathbf{b}_{2i})}{1 - F(d \mid \mathbf{b}_{2i})} P(c_{ij} = 0) \right]^{1-c_{ij}} P(c_{ij} = 1)^{c_{ij}}. \quad (4)$$

Approach 2: Two-regime method—In some applications, it may be reasonable to assume that some censored values follow the same distribution as the observed uncensored values, such as those close to the LLOQ. In other words, we may assume two regimes for the censored longitudinal data: (i) some censored longitudinal values are assumed to still follow the assumed models for the observed values y_{ij} (denoted by Regime $\mathcal{R}_1$); and (ii) the remaining censored longitudinal values are not assumed to follow any parametric

distribution (denoted by Regime $\mathcal{R}_2$). The proportions of censored values in the two regimes are unknown. However, if some censored values may be reasonably assumed to follow a particular distribution, then this knowledge can be leveraged to improve predictions of the censored values when necessary.

An important issue is which of the two regimes does a censored data point fall? Let τ_{ij} be the regime indicator such that $\tau_{ij} = 1$ when y_{ij} is in Regime $\mathcal{R}_1$ (may be censored) and $\tau_{ij} = 0$ when y_{ij} is in Regime $\mathcal{R}_2$ (always censored). In many applications, the values of τ_{ij} are not observed, and in this project we always treat them as missing data or unknown latent indicators and never assume they are known. When $\tau_{ij} = 0$ (i.e., $y_{ij} \in \mathcal{R}_2$), similar to Approach 1, we view the censored values as a point mass subgroup and do not make any distributional assumption. When $\tau_{ij} = 1$ (i.e., $y_{ij} \in \mathcal{R}_1$), we assume the censored values follow the same distribution as the observed y -data, i.e., model (2).

While the regime indicator τ_{ij} is not observed, sometimes it may be roughly predicted by observed covariates. For example, if viral load is left-censored but the corresponding CD4 cell count is observed, then if the observed CD4 value is very high, then the censored viral load is likely to be small due to the typical negative association between viral load and CD4. In such a case it is prudent to not assume the censored viral load follows a parametric distribution (i.e., the viral load may be in $\mathcal{R}_2$). Moreover, τ_{ij} may sometimes be predicted to be in $\mathcal{R}_1$, for example if the neutralizing antibody response is observed to be slightly above the LLOQ and the binding antibody response is left-censored, then the latter response is likely near the LLOQ given that binding antibodies are necessary for a neutralizing antibody response.

Thus, we assume the following logistic regression model with random effects

$$\text{logit}(P(\tau_{ij} = 1)) = \mathbf{w}_{ij}^T \boldsymbol{\eta} + \mathbf{v}_{ij}^T \mathbf{b}_{1i} \equiv \zeta_{ij}, \quad \mathbf{b}_{1i} \sim N(\mathbf{0}, \mathbf{D}_1), \quad (5)$$

where $\mathbf{w}_{ij}$ and $\mathbf{v}_{ij}$ are observed covariate vectors (including time), $\boldsymbol{\eta}$ are fixed parameters, and $\mathbf{b}_{1i}$ are random effects. Finally, the conditional distribution of (y_{ij}, c_{ij}) given the random effects may be written as (the parameters are suppressed for simplicity)

$$\begin{aligned} f(y_{ij}, c_{ij} | \mathbf{b}_{1i}, \mathbf{b}_{2i}) &= \mathbf{f}(\mathbf{y}_{ij}, \mathbf{c}_{ij} | \tau_{ij} = 0, \mathbf{b}_{1i}, \mathbf{b}_{2i}) P(\tau_{ij} = 0) + \mathbf{f}(\mathbf{y}_{ij}, \mathbf{c}_{ij} | \tau_{ij} = 1, \mathbf{b}_{1i}, \mathbf{b}_{2i}) \\ &\quad P(\tau_{ij} = 1) = \left[P(\tau_{ij} = 0) + F(d | \mathbf{b}_{2i}) P(\tau_{ij} = 1) \right]^{c_{ij}} \left[f(y_{ij} | \mathbf{b}_{2i}) P(\tau_{ij} = 1) \right]^{1 - c_{ij}}, \end{aligned} \quad (6)$$

where $F(d | \mathbf{b}_{2i})$ is the normal cumulative distribution function the same as in (4).

2.2 Models for uncensored longitudinal data and survival data

In addition to the longitudinal variable Y , which may be left censored, there are often other longitudinal variables without censoring. For simplicity, we consider one possibly discrete longitudinal variable Z not subject to censoring, such as a binary variable, which may be associated with Y and survival time. For the longitudinal variable Z , we consider a GLME model, which allows Z to be binary, count, or other types,

$$q(E(Z_{ik})) = \mathbf{x}_{ik}^T \boldsymbol{\alpha} + \mathbf{u}_{ik}^T \mathbf{b}_{3i}, \quad \mathbf{b}_{3i} \sim N(\mathbf{0}, \mathbf{D}_3) \quad (7)$$

where $q()$ is a known link function, $\mathbf{x}_{ik}$ and $\mathbf{u}_{ik}$ are vectors of covariates, $\boldsymbol{\alpha}$ are fixed parameters, $\mathbf{b}_{3i}$ is a vector of random effects, and Z is assumed to follow a distribution in the exponential family.

The two longitudinal variables Y and Z may be of different types, such as Y being continuous and Z being binary, and Y and Z may be highly associated, as in HIV vaccine studies. Moreover, the censoring indicator c_{ij} for the longitudinal variable Y may also be highly associated with Z . Because these variables are of different types, the nature of the association can be highly complicated. However, it seems reasonable to assume that they are all governed by common individual characteristics such as disease status, genetic, or innate immunity information. These individual characteristics may be reflected by the random effects in the three mixed effects models for these three variables. Therefore, to incorporate the associations between these longitudinal variables, we assume that the random effects from their models are correlated, i.e.,

$$\mathbf{b}_i = (\mathbf{b}_{1i}, \mathbf{b}_{2i}, \mathbf{b}_{3i}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ is an arbitrary covariance matrix.

Finally, a common objective in many longitudinal studies is to associate the longitudinal information with the risk of an event of interest, such as HIV infection. The risk of event may be associated with longitudinal information in various ways, such as current, cumulative, or past longitudinal information. Since some longitudinal data may be left censored (and thus not observed) and the random effects in the longitudinal model reflect individual-specific characteristics of the longitudinal processes, it seems reasonable to use the random effects in the longitudinal models as “covariates” for the survival model; such models are commonly used in the literature and are often called “shared-parameter models”. For the time-to-event model, we consider the following Cox proportional hazards model whose parameters have attractive practical interpretations as instantaneous event risks

$$h_i(t) = h_0(t) \exp(\mathbf{x}_{si}^T \boldsymbol{\gamma}_0 + \mathbf{b}_i^T \boldsymbol{\gamma}_1) \quad (8)$$

where $h_0(t)$ is an unspecified baseline hazard function, $\mathbf{x}_{si}$ are baseline covariates, $\mathbf{b}_i$ are random effects in the longitudinal and censoring models, and $\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1$ are fixed parameters.

We assume that the survival data and the longitudinal data are conditionally independent given the random effects.

3 Approximate Joint Likelihood Inference

For efficient inference, we consider estimating all model parameters *simultaneously* based on the joint likelihood for all of the observed data. Let $\theta = (\beta^T, \alpha^T, \eta^T, \gamma_0^T, \gamma_1^T)^T$ be the collection of all mean parameters, and let $\xi = (\sigma, \Sigma)$ be the collection of all variance-covariance (dispersion) parameters. The joint likelihood for all observed data $\{(\mathbf{y}_i, \mathbf{c}_i, \mathbf{z}_i, \mathbf{s}_i, \delta_i), i = 1, 2, \dots, n\}$ is given by

$$L(\theta, \xi) = \prod_{i=1}^n \int \{f(\mathbf{y}_i, \mathbf{c}_i | \mathbf{b}_{1i}, \mathbf{b}_{2i}) f(\mathbf{z}_i | \mathbf{b}_{3i}) f(\mathbf{s}_i, \delta_i | h_{0i}, \mathbf{b}_{1i}, \mathbf{b}_{2i}, \mathbf{b}_{3i}) f(\mathbf{b}_i | \Sigma)\} d(\mathbf{b}_{1i}, \mathbf{b}_{2i}, \mathbf{b}_{3i}).$$

As we can see, the above likelihood involves a high-dimensional intractable integration. Numerical methods such as Gaussian quadrature or Monte Carlo methods such as the Monte Carlo EM algorithm can be computationally very intensive and may even have convergence problems. Thus, in the following, we propose approximate joint likelihood inference based on the “h-likelihood” method (Lee et al., 2006; Commenges et al., 2011). The h-likelihood method is essentially a Laplace approximation to the integral, but it can provide approximate MLEs for the mean parameters and approximate restricted maximum likelihood estimates for dispersion parameters, with reasonable accuracy. Moreover, it is computationally much more efficient.

The log h-likelihood function of the joint model can be written as

$$\ell_h = \log \{f(\mathbf{y}, \mathbf{c} | \mathbf{b}_1, \mathbf{b}_2) f(\mathbf{z} | \mathbf{b}_3) f(\mathbf{s}, \delta | h_0, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3) f(\mathbf{b} | \Sigma)\}, \quad (9)$$

where $f(\mathbf{y}, \mathbf{c} | \mathbf{b}_1, \mathbf{b}_2)$ is given in (4) for Approach 1 and (6) for Approach 2, $f(\mathbf{z} | \mathbf{b}_3)$ is the density function of model (7), $f(\mathbf{s}, \delta | h_0, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$ is the density function of the survival model (8), and $f(\mathbf{b} | \Sigma)$ is the density function of the random effects. The estimation procedure is described below. Beginning with some starting values $(\xi^{(0)}, \theta^{(0)}, h_0^{(0)})$, we iterate the following steps until convergence: for $k = 1, 2, 3, \dots$,

- Step 1: At iteration k , given $(\hat{\xi}^{(k)}, \hat{\theta}^{(k)}, \hat{h}_0^{(k)})$, obtain updated estimates of the random effects $\mathbf{b}^{(k+1)}$ by maximizing ℓ_h in (9) with respect to $\mathbf{b}$;
- Step 2: Given $(\hat{\xi}^{(k)}, \hat{h}_0^{(k)}, \mathbf{b}^{(k+1)})$, obtain updated estimates of the mean parameters $\theta^{(k+1)}$ by maximizing the following adjusted profile h-likelihood as in Lee and Nelder (1996) with respect to θ .

$$p_{\theta} = \left(\ell_h - 0.5 \log \left| \frac{H(\ell_h, \mathbf{b})}{2\pi} \right| \right) \Big|_{\xi = \xi^{(k)}, h_0 = h_0^{(k)}, \mathbf{b} = \mathbf{b}^{(k+1)'}}$$

where $H = -\partial^2 \ell_h / \partial \mathbf{b}^T \partial \mathbf{b}$.

– Step 3: Given $(\hat{h}_0^{(k)}, \hat{\mathbf{b}}^{(k+1)}, \hat{\boldsymbol{\theta}}^{(k+1)})$, obtain updated estimates of the variance-covariance $\boldsymbol{\xi}^{(k+1)}$ by maximizing the following adjusted profile h-likelihood,

$$p_{\xi} = \left(\ell_h - 0.5 \log \left| \frac{H[\ell_h, (\boldsymbol{\theta}, \mathbf{b})]}{2\pi} \right| \right) \Big|_{h_0 = h_0^{(k)}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(k+1)}, \mathbf{b} = \mathbf{b}^{(k+1)'}}$$

where

$$H[\ell_h, (\boldsymbol{\theta}, \mathbf{b})] = - \begin{pmatrix} \frac{\partial^2 \ell_h}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} & \frac{\partial^2 \ell_h}{\partial \boldsymbol{\theta} \partial \mathbf{b}^T} \\ \frac{\partial^2 \ell_h}{\partial \mathbf{b} \partial \boldsymbol{\theta}^T} & \frac{\partial^2 \ell_h}{\partial \mathbf{b} \partial \mathbf{b}^T} \end{pmatrix}.$$

– Step 4: Given $(\hat{\mathbf{b}}^{(k+1)}, \hat{\boldsymbol{\theta}}^{(k+1)}, \hat{\xi}^{(k+1)})$, obtain an updated nonparametric estimate of the baseline hazard $\hat{h}_0^{(k+1)}$ as follows

$$h_0^{(k+1)}(t) = \sum_{i=1}^n \frac{\delta_i I(s_i = t)}{\sum_{i=1}^n \exp(\mathbf{x}_{si}^T \boldsymbol{\gamma}_0^{(k+1)} + (\mathbf{b}_i^{(k+1)})^T \boldsymbol{\gamma}_1^{(k+1)}) I(s_i \geq t)},$$

where $I(\cdot)$ is an indicator function.

We may set the starting values by first fitting the models separately and then choosing the resulting parameter estimates as the starting values for $\xi^{(0)}, \boldsymbol{\theta}^{(0)}, h_0^{(0)}$. **For η , we treat all the censored data as in Regime $\mathcal{R}_2$ with $\tau = 0$ and all the uncensored data as in Regime $\mathcal{R}_1$ with $\tau = 1$. Then we fit a logistic model of τ and use the resulting parameter estimates as the starting values of η .** After convergence, we can estimate the standard errors of the parameter estimates by numerically calculating the covariance matrix

$$\widehat{Cov}(\hat{\boldsymbol{\theta}}) = \left\{ - \sum_{i=1}^n \frac{\partial S_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\}^{-1} \Big|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}}, \text{ where } S_i(\boldsymbol{\theta}) \text{ is the score function of } \boldsymbol{\theta} \text{ for subject } i. \text{ We}$$

approximate $S_i(\boldsymbol{\theta})$ by using adaptive Gauss-Hermite method.

4 Application to an HIV vaccine trial data set

Our research is motivated by a 36-month follow-up efficacy study of a candidate vaccine to prevent HIV infection, the VAX004 trial. In this trial, HIV-uninfected participants received HIV vaccinations (with a recombinant glycoprotein 120 HIV vaccine containing the MN and GNE8 strains of HIV gp120) at months 0, 1, 6, 12, 18, 24, and 30. Plasma samples were collected at the immunization visits, two weeks after the immunization visits, and the final visit (i.e, months 0, 0.5, 1, 1.5, 6, 6.5, ..., 30, 30.5, 36), for measurement of several immune response biomarkers. The data set includes 8 such biomarkers, but most are highly correlated with each other. Here we focus on two immune response biomarkers: NAb (the titer of neutralizing antibodies to the MN strain of the HIV gp120 protein) and MNGNE8 (the average level of binding antibodies, measured by ELISA, to the MN and GNE8 HIV gp120 proteins) (Gilbert et al., 2005). Figure 1 shows the longitudinal patterns of NAb and MNGNE8 for some randomly selected vaccine recipients, where the solid lines represent participants who were HIV infected during the study, the dashed lines represent those who were not HIV infected by the end of the study, and the left censored values of the NAb variable are substituted by the LLOQ. Both NAb and MNGNE8 show periodic sinusoidal patterns, and, for MNGNE8, there seem to be some differences between HIV-infected and uninfected participants separated by the median value.

The data set we consider in this paper is a subset of the original data and has 194 vaccine recipients in total, among whom 21 acquired HIV infection during the trial with time from entry until infection diagnosis ranging from 43 to 954 days. The 194 vaccine recipients were selected such that each of them has received all vaccinations prior to infection diagnosis or dropout and has at least 3 measurement visits during the follow-up period. The ranges of observed NAb and MNGNE8 are [1.48, 4.82] and [0.01, 2.40] with medians 2.75 and 0.57, respectively. Moreover, NAb has a LLOQ of 1.48, and about 27% of NAb measurements are below the LLOQ (left-censored). See Yu et al. (2017) for more descriptions.

We denote by t_{ij} the measurement time (in months), $t_{d_{ij}}$ the time from the most recent immunization to the measurement time t_{ij} (in months), and τ_{ij} the time between two consecutive vaccine administrations (in months) such that t_{ij} falls between the two administrations. To avoid too small/large coefficients, we also scale some time variables and denote $t_{ij}^* = t_{ij}/12$ (in years) and $t_{d_{ij}}^* = t_{d_{ij}} * 30/7$ (in weeks).

In this section, we apply our proposed methods and joint models to the HIV vaccine data set. The joint models are described in Section 4.1 and the fitting results are given in Section 4.2. As comparisons, we also consider the other two methods for data analysis: a naive joint method, which substitutes the censored values by LLOQ, and a Tobit joint model, which assumes all the censored values follow the same normal distribution as the observed data.

4.1 The models

For the NAb longitudinal data, denoted by Y , which has about 27% censoring (below LLOQ d), we model the data using our proposed two approaches, respectively.

Approach 1: One-regime method—We model the observed (uncensored) data based on the following LME model

$$Y_{ij} \left| Y_{ij} \geq d = \beta_0 + \beta_1 t_{ij}^* + \beta_2 t_{ij}^{*2} + \beta_3 \sin\left(\frac{\pi}{\Delta_{ij}} \times t_{d_{ij}}\right) + d_1 b_{1i} + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma^2), b_{1i} \sim N(0, 1) \quad (10)$$

where b_{1i} is a random effect, ϵ_{ij} is an error term, $d_1 > 0$ is a variance parameter, and $Y_{ij}/Y_{ij} \geq d$ follows a *truncated normal distribution*. The $\sin()$ function is used to model the periodic pattern in the longitudinal response.

We also model the censoring indicator $C_{ij} = I_{(Y_{ij} \leq d)}$ associated with the NAb values to find possible associations of censoring with the failure time variables and other covariates, which can be used to predict the censoring status of NAb values at times when NAb values are unavailable. The observed data and standard variable selection methods (i.e. AIC and BIC criteria) lead to the following censoring model

$$\text{logit}(P(C_{ij} = 1)) = \eta_0 + \eta_1 t_{ij}^* + \eta_2 t_{ij}^{*2} + \eta_3 \sin\left(\frac{\pi}{\Delta_{ij}} \times t_{d_{ij}}\right) + d_2 b_{1i}, \quad (11)$$

where d_2 is a dispersion parameter and the model shares the same random effects as the NAb model since these two longitudinal processes exhibit very similar individual-specific patterns. For the one-regime method, we do not make any distributional assumption for the censored values.

Approach 2: Two-regime method—Since the censoring rate of NAb is high, it may be reasonable to assume that the censored data come from two regimes: $\mathcal{R}_1$ where the variability in the biomarker below the LLOQ may associate with HIV risk in a similar way as the variability in the biomarker above the LLOQ, and $\mathcal{R}_2$ where the variability below the LLOQ is a priori unlikely to be associated with HIV risk, which could stem from the variability being dominated by technical measurement error or from the low value marking a vaccine recipient receiving protection against HIV infection by the vaccine without additional information in the variation (see the discussion in Section 2.1). Let τ_{ij} be the regime indicator of the biomarkers at time t_{ij} . If $\tau_{ij} = 1$, the censored y_{ij} values in $\mathcal{R}_1$ follow the same distribution as the observed (uncensored) y_{ij} values, whereas if $\tau_{ij} = 0$ the censored y_{ij} values in $\mathcal{R}_2$ do not follow any distribution. Although the values of τ_{ij} are not observed, we assume they are closely related to the time from the most recent vaccination. Thus, we consider the following logistic model for the regime indicator τ_{ij}

$$\text{logit}(P(\tau_{ij} = 1)) = \lambda_0 + \lambda_1 t_{d_{ij}} + \lambda_2 \mathbf{I}_{t_{ij}}, \quad (12)$$

where $\mathbf{I}_{t_{ij}}$ is an indicator such that $\mathbf{I}_{t_{ij}} = 1$ when t_{ij} falls in the first immunization cycle (i.e. $t_{ij} \in [0, 1]$) and $\mathbf{I}_{t_{ij}} = 0$ otherwise. The reason why we consider these two predictors in the above model is two-fold. First, the impact of vaccination on HIV risk changes with time $t_{d_{ij}}$, with a delay between vaccination and an effective immune response induced by vaccination, such that when (say) $t_{d_{ij}} = 1/30$ (i.e. the first day of an immunization cycle; in months), y_{ij} likely belongs to $\mathcal{R}_2$, while when $t_{d_{ij}}$ is around (say) 0.5 (i.e. two weeks after immunization; in months), y_{ij} likely belongs to $\mathcal{R}_1$ as an effective/protective immune response could exist by then. Second, in general the first dose of vaccination response, given at month 0, does not produce protective immunity but only “primes” the immune system (Centers for Disease Control and Prevention et al., 2012), so that censored immune response biomarkers in the first immunization cycle (month 0–1) are more likely to belong to regime $\mathcal{R}_2$ where it is a priori unlikely for the variability in the immune response biomarker below the LLOQ to be associated with HIV risk. Note that we do not include other variables and random effects in the model in order to reduce the model complexity and avoid the problem of non-identifiability.

For the longitudinal data y_{ij} in regime $\mathcal{R}_1$, we assume that they follow the same distribution as the observed (uncensored) y_{ij} data, i.e., model (10) but without conditioning on $Y_{ij} = d$ and the y_{ij} data are left censored.

For the binary MNGNE8 data without censoring, denoted by Z_{ij} we consider a logistic GLME model below,

$$\text{logit}(P(Z_{ij} = 1)) = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 \sin\left(\frac{\pi}{\Delta_{ij}} \times t_{d_{ij}}\right) + \alpha_3 t_{d_{ij}}^* + d_3 b_{1i} + d_4 b_{2i} t_{ij}, \quad (13)$$

where the individual characteristics are incorporated via random effects b_{2i} and the random effects b_{1i} in the model for NAb, $b_{2i} \sim N(0, 1)$, d_3 is a dispersion parameter, and $d_4 > 0$ is the variance parameter.

The survival model for the time to HIV infection is given by the following “shared-parameter” Cox model

$$h_i(t|x_i, \mathbf{b}_i) = h_0(t) \exp\{\gamma_0 x_i + \gamma_1 b_{1i} + \gamma_2 b_{2i}\}, \quad (14)$$

where x_i is the standardized baseline measurement of the immune response GNE8_CD4 for individual i .

The association between the two different types of longitudinal variables Y_{ij} and Z_{ij} is incorporated through correlated random effects in the two models. That is, the random effects from different longitudinal models are associated with correlation matrix

$$\text{cor}(b_{1i}, b_{2i}) = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}$$

with $r_{12} \in (-1, 1)$. The correlation of Z and Y does not affect the classification of censored values of Y into regime $\mathcal{R}_1$ or regime $\mathcal{R}_2$ as the model (12) of regime indicator only depends on some time variables.

4.2 Data analysis results

Table 1 summarizes estimation results based on the four methods, where the standard errors of parameter estimates were estimated based on the adaptive Gauss-Hermite method with 4 quadrature points. The h-likelihood methods were considered to have converged when the relative change is less than 0.05 for all of the mean parameter estimates on average or less than 0.01 for the approximated log-likelihood, whichever occurs first. The h-likelihood method is computationally very efficient and converged without problems. For parameter estimates, we focus on γ_1 and γ_2 , since these two parameters link HIV-infection risk to longitudinal characteristics and our main objective is to investigate if individual-specific characteristics of the longitudinal immune responses (represented by the random effects) are associated with risk of HIV infection. The parameter γ_1 measures the relationship of the HIV-infection risk with the baseline values of NAb, the baseline probabilities of NAb below LLOQ, and the baseline probabilities of MNGNE8 being above the median, since b_{1i} is shared by the three mixed effects models. The parameter γ_2 measures the relationship between the HIV-infection risk and the changing rate of the log-odds of MNGNE8 above the median over time.

From Table 1, we see that the four methods lead to quite different results, especially the estimates of parameters in the longitudinal model (10) of Y , which is subject to left censoring, and the survival model (14). For the naive method and Tobit method, γ_1 seems not significantly different from zero, suggesting that individual-specific baseline immune response biomarkers may not be associated with the risk of HIV infection. However, the other two methods show that γ_1 may in fact be significantly different from zero, i.e., individual-specific baseline immune response biomarkers may in fact be associated with the risk of HIV infection. Therefore, our proposed methods lead to different and potentially important conclusions than existing methods. The negative sign of $\hat{\gamma}_1$ shows that participants with higher baseline values of NAb, lower baseline probabilities of NAb below LLOQ, and higher baseline probabilities of MNGNE8 above the median, seem to have lower risks of HIV infection. The positive sign of $\hat{\gamma}_2$ suggests that participants with larger increasing rate over time for the probability of MNGNE8 above the median seem to have larger risks of

HIV infection. Moreover, the estimates of γ_1 , γ_2 , and β 's are different based on different joint models, suggesting that the contributions of the longitudinal information to the risk of HIV infection may be over-estimated or under-estimated if inappropriate models are used.

In addition, in model (12) for the two-regime method, we have the parameter estimates $\hat{\lambda}_0 = 0.74$, $\hat{\lambda}_1 = 5.29$, $\hat{\lambda}_2 = -6.9$ with standard errors **0.11**, **0.71**, **0.62**, respectively. The positive sign of $\hat{\lambda}_1$ shows that when $t_{d_{ij}}$ is larger, y_{ij} is more likely in $\mathcal{R}_1$, that is, the reduced risk of HIV by vaccination likely arises two weeks after the vaccination time. The negative sign of $\hat{\lambda}_2$ suggests that, in the first vaccination cycle, y_{ij} has lower probability of belonging to $\mathcal{R}_1$. These results match with our discussion and expectation described in Section 4.1 earlier.

5 Simulation Studies

We conducted a comprehensive simulation study to evaluate the proposed methods. In particular, we focus on performances of the one-regime and two-regime methods for censored longitudinal data. The models are similar to those used in the real data analysis in Section 4, and the true parameter values used for data generation are chosen to be close to those parameter estimates obtained from the real data analysis based on the one-regime and two-regime methods.

5.1 Simulation Design

Given the values of random effects generated from some Gaussian distribution, we generate the binary longitudinal data Z_{ik} from a Bernoulli distribution with probability $p_{ij} = \exp(\xi_{ij}) / (1 + \exp(\xi_{ij}))^{-1}$, where $\xi_{ij} = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 \sin\left(\frac{\pi}{\Delta_{ij}} t_{d_{ij}}\right) + \alpha_3 t_{d_{ij}}^* + d_1 b_{1i} + d_2 b_{2i} t_{ij}$ is chosen to mimic the periodic patterns typically seen in HIV vaccine studies (see Section 4). For the continuous longitudinal variable Y_{ij} with some left censored values, we consider the following two scenarios to generate the data:

- *Scenario 1:* The longitudinal data Y are generated from two regimes: one that follows a normal distribution ($\mathcal{R}_1$) and the other a point mass subgroup below the LLOQ ($\mathcal{R}_2$) – see Section 2 for a detailed description.

We first randomly generate longitudinal data y_{ij} from the normal distribution $Y_{ij}/b_{1i} \sim \mathcal{N}(\mu_{ij}, \sigma^2)$, with $\mu_{ij} = \beta_0 + \beta_1 t_{ij}^* + \beta_2 t_{ij}^{*2} + \beta_3 \sin\left(\frac{\pi}{\Delta_{ij}} t_{d_{ij}}\right) + d_3 b_{1i}$ to mimic the periodic patterns typically seen in HIV vaccine studies (see Section 4). If $y_{ij} \leq \text{LLOQ}$, it is considered to be censored. This step produces left-censored data which are assumed to be in $\mathcal{R}_1$. Then, we create additional left censored values in the late period of each vaccination cycle to mimic the real data described in Section 4. Specifically, if $t_{d_{ij}} > \mathcal{T}_i$, where $\mathcal{T}_i$ is a time point whose values are to be determined by the desired censoring rates, we let y_{ij} be left-censored. This step produces left-

censored data which are assumed to be in $\mathcal{R}_2$. In the simulation study we consider three cases: (i) a total of roughly 20% left-censored longitudinal data, with 25% from $\mathcal{R}_1$ and 75% from $\mathcal{R}_2$, LLOQ $d=2$, $\mathcal{T}_i \sim N(1/3, 1/6)$ for the first vaccination, and $\mathcal{T}_i \sim N(4.83, 1/6)$ for the other vaccinations; (ii) a total of roughly 40% left-censored longitudinal data, with 13% from $\mathcal{R}_1$ and 87% from $\mathcal{R}_2$ for the first vaccination, and $\mathcal{T}_i \sim N(1/3, 0.1)$ for the other vaccinations; (iii) a total of roughly 40% left-censored longitudinal data, with 37% from $\mathcal{R}_1$ and 63% from $\mathcal{R}_2$ for the first vaccination, and $\mathcal{T}_i \sim N(1/3, 0.1)$ for the other vaccinations. The three cases are designed to study how different proportions of left-censoring and different proportions of the two regimes affect the performances of the proposed methods. Note that this censoring generating process is highly associated with $t_{d_{ij}}$ and whether the data are generated during the first vaccination cycle or not, because the censoring is determined by $t_{d_{ij}} > \mathcal{T}_i$ and $\mathcal{T}_i$ is simulated from normal distributions with different means for the first vaccination and the rest vaccinations. This is consistent with the model (12) discussed earlier.

– *Scenario 2:* We assume the same normal distribution as that in Scenario 1, but assume that the censored data and the observed uncensored data follow the same distribution. This scenario may not be reasonable in practice, as discussed in Section 1, but it is typically assumed in the literature for censored longitudinal data, so it is considered here for comparison purposes to see how our proposed methods perform if this typical assumption indeed holds. We consider two cases, 20% censored data with $d=2.7$ and 40% censored data with $d=3.3$. The longitudinal data y_{ij} d are left censored.

In the longitudinal models, the random effects are generated from a multivariate normal distribution $\mathbf{b}_i = (b_{1i}, b_{2i})' \sim N(0, \Sigma)$, with $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$. The parameter values are given as $\boldsymbol{\alpha} = (-1.65, 0.15, 1.8, -0.05)'$, $\boldsymbol{\beta} = (2, 1, -0.3, 1.5)'$, $\boldsymbol{\gamma} = (-0.75, -1.5, 2)$, $d_1 = 0.4$, $d_2 = 0.5$, $d_3 = 0.15$, $\sigma = 0.5$. The survival times T_i^* are generated from a Weibull distribution with shape parameter 15 and scale parameter $800 \exp(\gamma_0 x_i + \boldsymbol{\gamma}_1^T \mathbf{b}_i)^{-1/\text{shape}}$, where x_i is a baseline covariate generated from the standard normal distribution. The non-informative censoring times $\mathcal{C}_i$ are generated from a Weibull distribution with $\text{shape} = 5$ and $\text{scale} = 1000$. The observed survival times are defined as $s_i = \min(t_i^*, \mathcal{C}_i)$, with event indicator $\delta_i = I(t_i^* \leq \mathcal{C}_i)$.

The percentage of participants with a failure event during the follow-up period is about 74%, the expected number of observed failure events during the follow-up period is 148, and the duration of the follow-up period is 36 months.

When applying the two-regime method, we consider a logistic model (see model (12) in Section 4) for predicting which censored values are in regime $\mathcal{R}_1$ or $\mathcal{R}_2$. So the membership of censored values is predicted based on some time variables only. Note that in the

simulation results in Section 5.2, we do not evaluate the parameter estimates of model (12) nor assess the goodness of the predictors because the true membership of the censored values is not determined by the model (12) – we generate excessive censoring data (which can be viewed as in $\mathcal{R}_2$) by creating left censoring for measurements in the late period of each vaccination cycle. So the true parameter values of model (12) are unknown although the membership does depend on some time variables.

To compare the *one-regime (OR)* and *two-regime (TR)* methods with existing ones, we also study a *naive joint model (NV)* method, which substitutes the left-censored values by LLOQ, and the *Tobit joint model (TB)*, which assumes that censored values follow the same distribution as the observed values as commonly assumed in the literature. We evaluate the performances with respect to bias and mean square error (MSE). That is, we consider percent relative bias and percent relative MSE as defined as follows (say, for parameter β):

$$\begin{aligned} & - \text{relative bias (\%)} \text{ of } \hat{\beta} = \left| \left(\frac{1}{M} \sum_{m=1}^M \hat{\beta}^{(m)} - \beta \right) / \beta \right| \times 100\%, \\ & - \text{relative MSE (\%)} \text{ of } \hat{\beta} = \frac{1}{M} \sum_{m=1}^M (\hat{\beta}^{(m)} - \beta)^2 / \left| \beta \right| \times 100\%, \end{aligned}$$

where $\hat{\beta}^{(m)}$ is the estimate of β in simulation replication m , $M=200$ is the number of total repetitions in the simulation, and β is the true parameter value.

5.2 Simulation Results

Tables 2–4 summarize the results for data generated under Scenario 1 with 20% and 40% left censored longitudinal data, respectively. We see that, with 20% censored longitudinal data, **the one-regime method and two-regime method perform similarly well and outperform the other methods, since they have smaller values of Bias and MSE and more accurate coverage probabilities. With 40% censored data, the two-regime method performs better than the one-regime method as it gives less biased estimates on average.** In particular, the Tobit joint model performs poorly, suggesting that, if much of the censored data do not follow a parametric distribution, models assuming such a distribution can perform poorly. Therefore the proposed methods may be preferred over the existing naive method and Tobit method.

Tables 5 and 6 summarize the results for data generated under Scenario 2 with 20% and 40% censored longitudinal data, respectively. The purpose of this simulation is to check, if the censored data indeed follow a parametric distribution, how poorly the proposed methods can perform. In this case, as expected, the Tobit joint model, which assumes a parametric distribution for censored values, outperforms the other three methods in general. However, the advantage of the Tobit joint model is minor compared to our proposed methods when the censoring rate is relatively low (see Table 5). In other words, if the censored data indeed follow a parametric distribution, methods without a distributional assumption or only some distributional assumption still perform reasonably well. On the other hand, as discussed in previous sections, parametric distributional assumptions for censored data may be unreasonable in practice and they are not verifiable.

The standard errors of parameter estimates based on the adaptive Gauss-Hermite method are evaluated in Tables 2–6 through coverage probabilities. We see that the coverage probabilities of our methods are generally close to the nominal level.

6 Conclusions and Discussion

In HIV/AIDS studies, censoring in longitudinal data is an important issue since it may reflect the efficacy of a vaccine or viral suppression by an HIV-treatment; moreover variability in the longitudinal data below the LLOQ may be expected to have a different association with the risk of an event than variability above the LLOQ, due either to the variability below the LLOQ being dominated by technical measurement error or from the low value marking a protected individual without additional information in the variation. Similar issues occur in other cohort studies with objective to associate longitudinal biomarkers with event times of interest. Therefore, the choice of appropriate statistical models and methods to address censoring is an important topic. Existing methods in the literature that assume the censored data follow the same distribution as the observed data have disadvantage that this assumption is unverifiable and may be unreasonable due to the nature of the censored data. In this article, we propose one-regime and two-regime methods to describe possibly different distributions of the censored data, with biological justifications. The simulation study shows that the proposed methods work better than existing ones for the scenarios studied, especially when the censoring rate is relatively high.

Another contribution of this article is the computationally efficient h-likelihood method. Although the h-likelihood method is not new, its application in our current setting of joint models for censored longitudinal data and survival data is new. Even for a simple joint model of a LME model and a Cox model, computation based on likelihood methods can be quite challenging if the dimension of random effects is not low (Rizopoulos, 2012). For the current joint models of mixed types of longitudinal data with left censoring and survival data, computation is even more challenging. The proposed h-likelihood appears to be computationally very efficient and performs quite well based on simulations.

The real data analysis shows that the longitudinal characteristics of immune response biomarkers during an HIV vaccine study can be highly correlated with the risk of HIV infection. This may be an important finding for HIV vaccine research since we may use the longitudinal information to predict risk of HIV infection.

Acknowledgements

Research reported in this publication was supported by the National Institute Of Allergy And Infectious Diseases (NIAID) of the National Institutes of Health (NIH) under Award Numbers R37AI054165 and UM1AI068635, and by the Bill and Melinda Gates Foundation Award Number OPP1110049. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or BMGF. The authors thank the participants, investigators, and sponsors of the VAX004, including Global Solutions for Infectious Diseases.

References

- Berk KN (2002). Repeated measures with zeros. *Statistical Methods in Medical Research*, 11(4):303–316. [PubMed: 12197298]

- Centers for Disease Control and Prevention et al. (2012). Principles of vaccination Epidemiol. Prev. Vaccine-Preventable Dis 12th ed. Washington DC: Public Health Foundation, pages 1–9.
- Commenges D, Jolly D, Drylewicz J, Putter H, and Thiébaut R (2011). Inference in hiv dynamics models via hierarchical likelihood. Computational Statistics & Data Analysis, 55(1):446–456.
- Dagne GA and Huang Y (2013). Bayesian semiparametric mixture tobit models with left censoring, skewness, and covariate measurement errors. Statistics in medicine, 32(22):3881–3898. [PubMed: 23553914]
- Gilbert PB, Peterson ML, Follmann D, Hudgens MG, Francis DP, Gurwith M, Heyward WL, Jobes DV, Popovic V, Self SG, et al. (2005). Correlation between immunologic responses to a recombinant glycoprotein 120 vaccine and incidence of hiv-1 infection in a phase 3 hiv-1 preventive vaccine trial. Journal of Infectious Diseases, 191(5):666–677. [PubMed: 15688279]
- Hughes JP (1999). Mixed effects models with censored data with application to hiv rna levels. Biometrics, 55(2):625–629. [PubMed: 11318225]
- Jacqmin-Gadda H, Thiébaut R, Chêne G, and Commenges D (2000). Analysis of left-censored longitudinal data with application to viral load in hiv infection. Biostatistics, 1(4):355–368. [PubMed: 12933561]
- Lachenbruch PA (2002). Analysis of data with excess zeros. Statistical Methods in Medical Research, 11(4):297–302. [PubMed: 12197297]
- Lee Y and Nelder JA (1996). Hierarchical generalized linear models. Journal of the Royal Statistical Society. Series B (Methodological), pages 619–678.
- Lee Y, Nelder JA, and Pawitan Y (2006). Generalized linear models with random effects: unified analysis via H-likelihood. CRC Press.
- Lorimer M and Kiermeier A (2007). Analysing microbiological data: Tobit or not tobit? International journal of food microbiology, 116(3):313–318. [PubMed: 17382420]
- Moulton LH and Halsey NA (1995). A mixture model with detection limits for regression analyses of antibody response to vaccine. Biometrics, pages 1570–1578. [PubMed: 8589241]
- Rizopoulos D (2012). Joint models for longitudinal and time-to-event data: With applications in R. CRC Press.
- Schneider W (2005). Likelihood estimation for censored random vectors. Econometric Reviews, 24(2): 195–217.
- Thiébaut R, Jacqmin-Gadda H, Babiker A, and Commenges D (2005). Joint modelling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of cd4+ cell count and hiv rna viral load in response to treatment of hiv infection. Statistics in medicine, 24(1):65–82. [PubMed: 15523706]
- Wu L (2002). A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to aids studies. Journal of the American Statistical association, 97(460):955–964.
- Yu T, Wu L, and Gilbert PB (2017). A joint model for mixed and truncated longitudinal data and survival data, with application to HIV vaccine studies. Biostatistics (to appear).

Fig. 1. Longitudinal patterns of NAb and MNGNE8 for a few randomly selected vaccine recipients, where the solid lines represent HIV infected participants and the dashed lines represent HIV uninfected participants. The left censored values in NAb are substituted by the LLOQ of 1.48.

Table 1

Estimates of all model parameters.

Model	Par	Naive method			One-regime method			Two-regime method			Tobit method		
		Est	SE	Pvalue	Est	SE	Pvalue	Est	SE	Pvalue	Est	SE	Pvalue
(10)		Estimates of mean parameters											
	β_0	1.57	0.05	0.00	2.32	0.06	0.00	2.12	0.08	0.00	1.18	0.07	0.00
	β_1	1.88	0.06	0.00	0.95	0.07	0.00	1.2	0.09	0.00	2.38	0.09	0.00
	β_2	-0.53	0.02	0.00	-0.31	0.02	0.00	-0.38	0.03	0.00	-0.67	0.03	0.00
	β_3	0.55	0.07	0.00	1.48	0.09	0.00	1.52	0.10	0.00	0.75	0.10	0.00
(11)	η_0				1.98	0.18	0.00						
	η_1				-6.35	0.37	0.00						
	η_2				1.67	0.13	0.00						
	η_3				-0.65	0.28	0.02						
	α_0	-1.73	0.13	0.00	-1.74	0.13	0.00	-1.75	0.13	0.00	-1.74	0.13	0.00
(13)	α_1	0.14	0.01	0.00	0.14	0.01	0.00	0.14	0.01	0.00	0.14	0.01	0.00
	α_2	1.87	0.24	0.00	1.88	0.24	0.00	1.91	0.24	0.00	1.89	0.24	0.00
	α_3	-0.04	0.01	0.00	-0.04	0.01	0.00	-0.05	0.01	0.00	-0.05	0.01	0.00
	γ_0	-0.71	0.32	0.02	-0.65	0.32	0.04	-0.70	0.34	0.04	-0.65	0.30	0.03
	γ_1	-1.03	0.59	0.08	-1.19	0.52	0.02	-1.31	0.61	0.03	-0.80	0.58	0.17
14)	γ_2	1.47	0.47	0.00	1.54	0.43	0.00	1.67	0.46	0.00	1.45	0.45	0.00
Estimates of variance-covariance parameters													
	σ	0.63			0.48			0.51			0.78		
	d_1	0.28			0.25			0.29			0.38		
	d_2	-			-0.78			-			-		
	d_3	0.05			0.06			0.05			0.07		
	d_4	0.09			0.09			0.09			0.09		
	r_{12}	0.58			0.54			0.59			0.56		

Table 2

Simulation results for Scenario 1 with 20% censored longitudinal data (25% from $\mathcal{D}_1$ and 75% from $\mathcal{D}_2$).

Par	True value	Estimate (SSE, SE)			Bias%			MSE%			CP%		
		NV	OR	TR	TB	NV	OR	TR	TB	NV	OR	TR	TB
α_0	-1.65	-1.65 (.08,.09)	-1.69 (.07,.09)	-1.67 (.08,.09)	-1.67 (.07,.09)	0.1	2.2	1.2	1.3	0.4	0.4	0.4	0.3
α_1	0.15	0.15 (.01,.01)	0.15 (.01,.01)	0.15 (.01,.01)	0.15 (.01,.01)	1	0.6	1.2	1.6	0.1	0.1	0.1	0.1
α_2	1.8	1.81 (.08,.1)	1.83 (.09,.1)	1.82 (.08,.1)	1.82 (.08,.1)	0.3	1.7	1	1.1	0.4	0.5	0.4	0.4
α_3	-0.05	-0.05 (.00,.00)	-0.05 (.00,.00)	-0.05 (.00,.00)	-0.05 (.00,.00)	0.8	0.9	1.3	1.2	0	0	0	0
β_0	2	1.91 (.02,.04)	2 (.04,.05)	2 (.04,.05)	1.65 (.04,.05)	4.5	0.2	0.2	17.3	0.4	0.1	0.1	6.1
β_1	1	0.79 (.03,.04)	1 (.03,.04)	1 (.03,.04)	0.9 (.04,.05)	20.6	0.1	0.2	10.1	4.3	0.1	0.1	1.2
β_2	-0.3	-0.26 (.01,.02)	-0.3 (.01,.02)	-0.3 (.01,.02)	-0.29 (.02,.02)	14.7	0.3	0.6	2.4	0.7	0.1	0.1	0.1
β_3	1.5	1.67 (.02,.02)	1.5 (.02,.02)	1.5 (.02,.02)	1.92 (.02,.03)	11.3	0.2	0	28.1	2	0	0	11.8
γ_0	-0.75	-0.71 (.12,.12)	-0.73 (.11,.12)	-0.73 (.12,.13)	-0.73 (.11,.13)	5.5	2.6	3.2	2.3	2	1.7	1.9	1.6
γ_1	-1.5	-1.45 (.22,.16)	-1.44 (.17,.15)	-1.46 (.18,.15)	-1.47 (.17,.15)	3.2	4.2	2.8	1.7	3.3	2.2	2.2	2
γ_2	2	2.01 (.25,.16)	2 (.19,.16)	2.05 (.2,.16)	2.04 (.2,.16)	0.6	0.2	2.3	1.9	3.2	1.8	2.1	2.1

NV: naive method, OR: one-regime method, TR: two-regime method, TB: Tobit joint model, SSE: simulation standard error, SE: mean estimated standard error based on the aGH method, CP: coverage probability.

Table 3

Simulation results for Scenario 1 with 40% censored longitudinal data, with different proportions in $\mathcal{R}_1$ and $\mathcal{R}_2$.

Case A: 13% from $\mathcal{R}_1$ and 87% from $\mathcal{R}_2$																	
Par	True value	Estimate (SSE, SE)				Bias%				MSE%				CP%			
		NV	OR	TR	TB	NV	OR	TR	TB	NV	OR	TR	TB	NV	OR	TR	TB
α_0	-1.65	-1.68 (.08,.09)	-1.7 (.08,.09)	-1.7 (.08,.09)	-1.7 (.08,.09)	1.8	3.1	2.7	3	0.4	0.5	0.5	0.5	0.5	0.5	0.5	0.5
α_1	0.15	0.15 (.01,.01)	0.15 (.01,.01)	0.15 (.01,.01)	0.15 (.01,.01)	2.2	1.5	1.8	2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
α_2	1.8	1.83 (.09,.1)	1.84 (.08,.1)	1.84 (.08,.1)	1.85 (.08,.1)	1.9	2.3	2.1	2.5	0.5	0.4	0.5	0.4	0.5	0.4	0.5	0.4
α_3	-0.05	-0.05 (.00,.00)	-0.05 (.00,.00)	-0.05 (.00,.00)	-0.05 (.00,.00)	1.8	1.8	1.6	2.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
β_0	2	2.17 (.02,.04)	2.01 (.04,.05)	2 (.04,.05)	1.55 (.05,.07)	8.6	0.2	0.2	22.4	1.5	0.1	0.1	0.1	0.1	10.2	0	99
β_1	1	0.3 (.03,.06)	1 (.04,.05)	1 (.03,.04)	0.3 (.06,.1)	69.7	0.1	0.0	70.5	48.7	0.2	0.1	0.1	0.1	50	0	95
β_2	-0.3	-0.1 (.02,.03)	-0.3 (.02,.02)	-0.3 (.02,.02)	-0.1 (.03,.05)	67.5	0.5	0.3	65	13.8	0.1	0.1	0.1	0.1	12.9	0	96
β_3	1.5	0.98 (.01,.03)	1.5 (.02,.02)	1.5 (.02,.02)	1.32 (.02,.05)	34.8	0.3	0.1	12.1	18.1	0.0	0.0	0.0	0.0	2.2	0	99
γ_0	-0.75	-0.76 (.11,.15)	-0.7 (.11,.12)	-0.73 (.12,.12)	-0.81 (.13,.16)	0.8	6.6	2.8	8.2	1.7	1.9	1.8	1.8	2.7	98	94	98
γ_1	-1.5	-1.72 (.19,.18)	-1.26 (.17,.14)	-1.39 (.2,.15)	-1.88 (.28,.18)	14.5	15.9	7.3	25.3	5.5	5.8	3.3	14.9	79	56	78	45
γ_2	2	2.3 (.22,.18)	1.89 (.19,.15)	2.03 (.21,.16)	2.59 (.3,.18)	14.9	5.5	1.7	29.3	6.8	2.3	2.3	21.7	60	79	88	21

NV: naive method, OR: one-regime method, TR: two-regime method, TB: Tobit joint model, SSE: simulation standard error, SE: mean estimated standard error based on the aGH method, CP: coverage probability.

Table 4

Simulation results for Scenario 1 with 40% censored longitudinal data, with different proportions in $\mathcal{R}_1$ and $\mathcal{R}_2$.

Case B: 37% from $\mathcal{R}_1$ and 63% from $\mathcal{R}_2$																	
Par	True value	Estimate (SSE, SE)				Bias%				MSE%				CP%			
		NV	OR	TR	TB	NV	OR	TR	TB	NV	OR	TR	TB	NV	OR	TR	TB
α_0	-1.65	-1.68 (.08,.09)	-1.69 (.08,.09)	-1.69 (.08,.09)	-1.69 (.07,.09)	1.7	2.6	2.3	2.5	0.4	0.5	0.5	0.4	98	97	97	96
α_1	0.15	0.15 (.01,.01)	0.15 (.01,.01)	0.15 (.01,.01)	0.15 (.01,.01)	1.8	1.5	0.8	0.6	0.1	0.1	0.1	0.1	94	93	89	90
α_2	1.8	1.83 (.07,.1)	1.85 (.09,.1)	1.84 (.09,.1)	1.84 (.09,.1)	1.6	2.7	2.2	2.3	0.3	0.5	0.5	0.5	99	96	96	98
α_3	-0.05	-0.05 (.00,.00)	-0.05 (.00,.00)	-0.05 (.00,.00)	-0.05 (.00,.00)	1.6	2.1	1.6	2.1	0.0	0.0	0.0	0.0	98	97	98	98
β_0	2	2.31 (.01,.03)	2.02 (.04,.05)	2.01 (.04,.05)	1.46 (.05,.06)	15.6	0.9	0.4	27.2	4.9	0.1	0.1	14.9	0	98	98	0
β_1	1	0.42 (.02,.04)	1 (.04,.05)	1 (.03,.04)	0.71 (.05,.07)	58	0.5	0.2	29.3	33.7	0.2	0.1	8.8	0	99	99	0
β_2	-0.3	-0.13 (.01,.02)	-0.3 (.02,.02)	-0.3 (.01,.02)	-0.23 (.02,.03)	56.1	0.1	0.2	22.8	9.5	0.1	0.1	1.7	0	97	99	38
β_3	1.5	1.19 (.02,.02)	1.49 (.02,.03)	1.5 (.02,.02)	1.87 (.02,.04)	20.5	0.9	0.1	24.6	6.3	0.0	0.0	9.1	0	98	98	0
γ_0	-0.75	-0.75 (.12,.13)	-0.72 (.11,.12)	-0.74 (.11,.13)	-0.75 (.12,.14)	0.6	3.4	1.8	0.4	2	1.6	1.7	2	97	99	99	97
γ_1	-1.5	-1.58 (.22,.17)	-1.37 (.19,.15)	-1.43 (.22,.15)	-1.56 (.21,.16)	5.4	8.5	4.8	4	3.6	3.4	3.5	3.1	86	77	75	87
γ_2	2	2.1 (.24,.17)	1.97 (.22,.16)	2.05 (.25,.16)	2.21 (.23,.17)	5.2	1.4	2.5	10.4	3.4	2.4	3.2	4.8	80	83	78	73

NV: naive method, OR: one-regime method, TR: two-regime method, TB: Tobit joint model, SSE: simulation standard error, SE: mean estimated standard error based on the aGH method, CP: coverage probability.

Table 5

Simulation result for Scenario 2 with 20% censored longitudinal data.

Par	True value	Estimate (SSE, SE)			Bias%			MSE%			CP%		
		NV	OR	TR	TB	NV	OR	TR	TB	NV	OR	TR	TB
α_0	-1.65	-1.66 (.08,.09)	-1.68 (.08,.09)	-1.67 (.08,.09)	-1.67 (.08,.09)	0.6	1.6	1.4	1.4	0.4	0.4	0.4	0.4
α_1	0.15	0.15 (.01,.01)	0.15 (.01,.01)	0.15 (.01,.01)	0.15 (.01,.01)	0.7	0.8	0.5	1.2	0.1	0.1	0.1	0.1
α_2	1.8	1.81 (.08,.1)	1.83 (.08,.1)	1.82 (.08,.1)	1.82 (.09,.1)	0.7	1.8	0.9	1	0.4	0.4	0.4	0.4
α_3	-0.05	-0.05 (.00,.00)	-0.05 (.00,.00)	-0.05 (.00,.00)	-0.05 (.00,.00)	0.5	1.8	0.9	0.6	0.0	0.0	0.0	0.0
β_0	2	2.49 (.02,.04)	2.02 (.04,.05)	2 (.04,.05)	2.01 (.04,.05)	24.6	1.2	0.2	0.3	12.1	0.1	0.1	0.1
β_1	1	0.67 (.03,.03)	0.99 (.04,.05)	1 (.03,.04)	1 (.03,.04)	33	1.1	0.1	0.1	11	0.1	0.1	0.1
β_2	-0.3	-0.19 (.01,.01)	-0.3 (.01,.02)	-0.3 (.01,.02)	-0.3 (.01,.02)	36.2	1.2	0.3	0.0	4	0.1	0.0	0.0
β_3	1.5	1.17 (.02,.02)	1.48 (.02,.03)	1.5 (.02,.02)	1.5 (.02,.02)	22.3	1.2	0.0	0.0	7.5	0.0	0.0	0.0
γ_0	-0.75	-0.71 (.12,.12)	-0.73 (.11,.12)	-0.74 (.1,.12)	-0.73 (.11,.12)	5	2.1	1.1	2.1	2.1	1.7	1.4	1.5
γ_1	-1.5	-1.4 (.19,.15)	-1.46 (.19,.15)	-1.42 (.17,.15)	-1.45 (.18,.15)	6.6	2.4	5.1	3.5	3.2	2.4	2.2	2.4
γ_2	2	1.98 (.24,.16)	2.04 (.22,.16)	2.02 (.21,.16)	2.03 (.2,.16)	0.8	2	1.2	1.7	2.9	2.5	2.3	2

NV: naive method, OR: one-regime method, TR: two-regime method, TB: Tobit joint model, SSE: simulation standard error, SE: mean estimated standard error based on the aGH method, CP: coverage probability.

Table 6

Simulation results for Scenario 2 with 40% censored longitudinal data.

Par	True value	Estimate (SSE, SE)			Bias%			MSE%			CP%		
		NV	OR	TR	TB	NV	OR	TR	TB	NV	OR	TB	TR
α_0	-1.65	-1.67 (.07,.09)	-1.69 (.07,.09)	-1.68 (.1,.09)	-1.68 (.07,.09)	0.9	2.5	1.9	1.8	0.3	0.4	0.6	0.4
α_1	0.15	0.15 (.01,.01)	0.15 (.01,.01)	0.15 (.01,.01)	0.15 (.01,.01)	1	1.7	0.8	2.2	0.1	0.1	0.1	0.1
α_2	1.8	1.81 (.08,.1)	1.84 (.08,.1)	1.83 (.1,.1)	1.83 (.08,.1)	0.8	2	1.6	1.5	0.4	0.4	0.6	0.4
α_3	-0.05	-0.05 (.00,.00)	-0.05 (.00,.00)	-0.05 (.01,.00)	-0.05 (.00,.00)	0.5	2.5	0.1	0.8	0.0	0.0	0.0	0.0
ρ_0	2	3.03 (.01,.03)	2.11 (.05,.06)	2.02 (.06,.05)	2.01 (.04,.05)	51.7	5.3	1.1	0.7	53.4	0.7	0.2	0.1
ρ_1	1	0.43 (.02,.03)	0.96 (.05,.06)	0.99 (.04,.04)	1 (.04,.04)	57.2	3.9	1.1	0	32.8	0.4	0.2	0.1
ρ_2	-0.3	-0.12 (.01,.01)	-0.29 (.02,.02)	-0.3 (.02,.02)	-0.3 (.02,.02)	60.8	3.6	1.6	0.2	11.1	0.2	0.1	0.1
ρ_3	1.5	0.79 (.03,.01)	1.44 (.03,.04)	1.49 (.03,.02)	1.5 (.02,.02)	47.4	4.1	0.9	0.3	33.7	0.3	0.1	0.0
γ_0	-0.75	-0.71 (.11,.12)	-0.75 (.12,.13)	-0.74 (.11,.12)	-0.75 (.12,.13)	6	0.2	1.6	0.6	1.9	2	1.7	2
γ_1	-1.5	-1.42 (.21,.17)	-1.57 (.22,.16)	-1.41 (.3,.15)	-1.46 (.22,.15)	5.5	4.9	6.2	2.5	3.3	3.6	6.6	3.4
γ_2	2	1.96 (.24,.16)	2.12 (.25,.17)	2.02 (.3,.16)	2.07 (.24,.16)	2	5.9	1	3.6	3	3.8	4.5	3

NV: naive method, OR: one-regime method, TR: two-regime method, TB: Tobit joint model, SSE: simulation standard error, SE: mean estimated standard error based on the aGH method, CP: coverage probability.