A joint model for mixed and truncated longitudinal data and survival data, with application to HIV vaccine studies

TINGTING YU*, LANG WU

Department of Statistics, University of British Columbia, 3182 Earth Sciences Building, 2207 Main Mall, Vancouver, BC V6T 1Z4, Canada
tingting.yu@stat.ubc.ca

PETER B. GILBERT

Department of Biostatistics, University of Washington, F-600, Health Sciences Building, Box 357232, Seattle, WA 98195-7232, USA and Vaccine and Infectious Disease and Public Health Sciences Divisions, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M2-C200, Seattle, WA 98109, USA

SUMMARY

In HIV vaccine studies, a major research objective is to identify immune response biomarkers measured longitudinally that may be associated with risk of HIV infection. This objective can be assessed via joint modeling of longitudinal and survival data. Joint models for HIV vaccine data are complicated by the following issues: (i) left truncations of some longitudinal data due to lower limits of quantification; (ii) mixed types of longitudinal variables; (iii) measurement errors and missing values in longitudinal measurements; (iv) computational challenges associated with likelihood inference. In this article, we propose a joint model of complex longitudinal and survival data and a computationally efficient method for approximate likelihood inference to address the foregoing issues simultaneously. In particular, our model does not make unverifiable distributional assumptions for truncated values, which is different from methods commonly used in the literature. The parameters are estimated based on the h-likelihood method, which is computationally efficient and offers approximate likelihood inference. Moreover, we propose a new approach to estimate the standard errors of the h-likelihood based parameter estimates by using an adaptive Gauss–Hermite method. Simulation studies show that our methods perform well and are computationally efficient. A comprehensive data analysis is also presented.

Keywords: Adaptive Gauss–Hermite; Cox model; h-Likelihood; Lower limit of quantification; Mixed-effect model; Shared-parameter model.

1. INTRODUCTION

In preventive HIV vaccine efficacy trials, participants are randomized to receive a series of vaccinations or placebos and are followed until the day of being diagnosed with HIV infection or until the end

*To whom correspondence should be addressed.

of study follow-up. We are often interested in the times to HIV infection. Meanwhile, blood samples are repeatedly collected over time for each participant in order to measure immune responses induced by the vaccine, such as CD4 T cell responses. A major research interest in HIV vaccine studies is to identify potential immune response biomarkers for HIV infection. For example, for many infectious diseases antibodies induced by a vaccine can recognize and kill a pathogen before it establishes infection; therefore high antibody levels are often associated with a lower risk of pathogen infection. Since longitudinal trajectories of some immune responses are often associated with the risk of HIV infection, in statistical analysis it is useful to jointly model the longitudinal and survival data. Moreover, such joint models can be used to address measurement errors and non-ignorable missing data in the longitudinal data.

There has been active research on joint models of longitudinal and survival data in recent years. Lawrence Gould *and others* (2015) have given a comprehensive review in this field. Rizopoulos *and others* (2009) proposed a computational approach based on the Laplace approximation for joint models of continuous longitudinal response and time-to-event outcome. Bernhardt *and others* (2014) discussed a multiple imputation method for handling left-truncated longitudinal variables used as covariates in AFT survival models. Król *and others* (2016) considered joint models of a left-truncated longitudinal variable, recurrent events, and a terminal event. The truncated values of the longitudinal variable were assumed to follow the same normal distributions as the untruncated values. Other recent work includes Fu and Gilbert (2017), Barrett *and others* (2015), Elashoff *and others* (2015), Chen *and others* (2014), Taylor *and others* (2013), Rizopoulos (2012b), and Zhu *and others* (2012). Analysis of HIV vaccine trial data offers the following new challenges: (i) some longitudinal data may be left truncated by a lower limit of quantification (LLOQ) of the biomarker assay, and the common approach of assuming that truncated values follow parametric distributions is unverifiable and may be unreasonable for vaccine trial data; (ii) the longitudinal multivariate biomarker response data are intercorrelated and may be of mixed types such as binary and continuous; (iii) the longitudinal data may exhibit *periodic patterns* over time, due to repeated administrations of the HIV vaccine; (iv) some longitudinal biomarkers may have measurement errors and missing data; and (v) the computation associated with likelihood inference can be very intensive and challenging. A comprehensive statistical analysis of HIV vaccine trial data requires us to address all the foregoing issues *simultaneously*. Therefore, despite extensive literature on joint models, new statistical models and methods are in demand.

In this article, we propose innovative models and methods to address the above issues. The contributions of the paper are: (i) when longitudinal data are left truncated, we propose a new method that does not assume any parametric distributions for the truncated values, which is different from existing approaches in the literature that unrealistically assume truncated values to follow the same distributions as those for the observed values; (ii) for longitudinal data with left truncation, the observed values are assumed to follow a *truncated* normal distribution; (iii) we incorporate the associations among several longitudinal responses of mixed types by linking the longitudinal models with shared and correlated random effects (Rizopoulos, 2012b); and (iv) we address the computational challenges of likelihood inference by proposing a computationally very efficient approximate method based on the h-likelihood method (Lee *and others*, 2006). It is known that, when the baseline hazard in the Cox survival model is completely unspecified, the standard errors of the parameter estimates in the joint models may be underestimated (Rizopoulos, 2012a; Hsieh *and others*, 2006). To address this issue, we also propose a new approach to estimate the standard errors of parameter estimates.

The article is organized as follows. In Section 2, we introduce the HIV vaccine trial data that motivates the research. In Section 3, we describe the proposed models and methods, which address all of the issues discussed above simultaneously. Section 4 presents analysis of the HIV vaccine trial data. Section 5 shows simulation studies to evaluate the proposed models and methods. We conclude the article with some discussion in Section 6.

2. THE HIV VACCINE DATA

Our research is motivated by the VAX004 trial, which is a 36-month efficacy study of a candidate vaccine to prevent HIV-1 infection, which contained two recombinant gp120 Envelope proteins: the MN and GNE8 HIV-1 strains (Flynn and others, 2005). One of the main objectives in the trial was to assess immune response biomarkers measured in vaccine recipients for their association with the incidence of HIV infection. It was addressed using plasma samples collected at the immunization visits, 2 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) to measure several immune response variables in vaccine recipients. Eight immune response variables were measured in total, most of which are highly correlated with each other and some have up to 16% missing data. We focus on a subset of these variables that may be representative and have low rates of missing data. In particular, we use the *NAb* and the *MNGNE8* variables, where *NAb* is the titer of neutralizing antibodies to the MN strain of the HIV-1 gp120 Env protein and *MNGNE8* is the average level of binding antibodies (measured by ELISA) to the MN and GNE8 HIV-1 gp120 Env proteins (Gilbert and others, 2005). Due to space limitation, the details of the clinical research questions, participant selection procedure, and other immune response variables are described in Section 1 in the [Supplementary material](#) available at *Biostatistics* online.

The data set we consider has 194 participants in total, among whom 21 participants acquired HIV infection during the trial with time of infection diagnosis ranging from day 43 to day 954 and event rate of 10.8%. The average number of repeated measurements over time is 12.6 per participant. Moreover, *NAb* has a LLOQ of 1.477, and about 27% of *NAb* measurements are below the LLOQ (left-truncated). To minimize the potential for bias due to missing data on the immune response biomarkers, the models adjust for the dominant baseline prognostic factor for HIV-1 infection—baseline behavioral risk score, which is grouped into three categories: 0, 1, or 2 for risk score 0, 1–3, 4–7, respectively.

Figure 1 shows the longitudinal trajectories of the immune responses of a few randomly selected participants, where the left truncated values in *NAb* are substituted by the LLOQ. The value of an immune response typically increases sharply right after each vaccination, and then starts to decrease about 2 weeks after the vaccination. Such patterns are shown as the reverse sawtooth waves in Figure 1. We see that participants HIV-infected later on seem to have lower values of *MNGNE8* and *NAb* than those uninfected by the end of the study. In particular, for *MNGNE8*, there seem to be clear differences between HIV-infected

Fig. 1. Longitudinal trajectories of two immune response variables for a few randomly selected VAX004 vaccine recipients, where the solid lines represent pre-infection trajectories of participants who acquired HIV infected and the dashed lines represent trajectories for participants who never acquired HIV infection. The left truncated values in *NAb* are substituted by the LLOQ of 1.477.

and uninfected participants, separated by the median value. The figures show that the longitudinal patterns of some immune responses seem to be associated with HIV infection, motivating inference via joint models of the longitudinal and survival data. In addition, some immune responses are highly correlated over time and are of mixed types, so we should also incorporate the associations among different types of longitudinal variables. Moreover, due to substantial variations across subjects, mixed effects models may be useful. The random effects in mixed effects models can serve several purposes: (i) they represent individual variations or individual-specific characteristics of the participants; (ii) they incorporate the correlation among longitudinal measurements for each participant; and (iii) they may be viewed as summaries of the individual profiles. Therefore, mixed effects models seem to be a reasonable choice for modeling the HIV vaccine trial data.

More results of the exploratory data analysis are given in Sections 2 and 3 in the [Supplementary material](#) available at *Biostatistics* online, including the summary statistics of the immune responses, Kaplan–Meier plot of the time to HIV infection, and longitudinal trajectories of NAb and MNGNE8 with the time variable shifted and aligned at the event times.

3. JOINT MODELS AND INFERENCE

3.1. The longitudinal, truncation, and survival models

3.1.1. Models for longitudinal data of mixed types. In the following, we denote by Y a random variable, y its observed value, $f(y)$ a generic density function, with similar notation for other variables. For simplicity of presentation, we consider two correlated longitudinal variables, Y and Z , where Y is continuous and subject to left truncation due to LLOQ, and Z is binary or count (e.g. dichotomized variable or number of CD4 T cells). The models can be easily extended to more than two longitudinal processes. Let d be the LLOQ of Y and C be the truncation indicator of Y such that $C = 1$ if $y \leq d$ and $C = 0$ otherwise.

For the continuous longitudinal variable Y , after possibly some transformations such as a log-transformation, we may assume that the *untruncated* data of Y follow a *truncated normal distribution*. We consider a linear or nonlinear mixed effects (LME or NLME) model for the observed values of y_{ij} given that $y_{ij} \geq d$, that is,

$$Y_{ij}|Y_{ij} \geq d = \Psi_{ij}^T \boldsymbol{\beta} + \Phi_{ij}^T \mathbf{b}_{1i} + \epsilon_{ij}, \quad \text{or} \quad Y_{ij}|Y_{ij} \geq d = g(\Psi_{ij}, \Phi_{ij}, \boldsymbol{\beta}, \mathbf{b}_{1i}) + \epsilon_{ij},$$

$$\mathbf{b}_{1i} \stackrel{iid}{\sim} N(\mathbf{0}, \mathbf{D}_1), \quad \epsilon_i \stackrel{iid}{\sim} N(\mathbf{0}, R_i), \quad i = 1, \dots, n, \quad j = 1, \dots, n_i, \quad (3.1)$$

where Y_{ij} is the longitudinal variable of participant i at time t_{ij} , Ψ_{ij} and Φ_{ij} are vectors of covariates, $\boldsymbol{\beta}$ contains fixed parameters, $\mathbf{b}_{1i}$ contains random effects, $g(\cdot)$ is a known nonlinear function, $\mathbf{D}_1$ and R_i are covariance matrices, and $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{in_i})^T$ are random errors independent of $\mathbf{b}_{1i}$. A LME model is usually an empirical model while an NLME model is a mechanistic model widely used in HIV viral dynamics (Wu, 2009). We assume that $R_i = \sigma^2 I_i$, i.e. the within-individual repeated measurements are independent conditional on the random effects.

In model (3.1), the observed Y_{ij} 's given the random effects and the condition " $Y_{ij} \geq d$ " (or $C_{ij} = 0$) are assumed to be normally distributed, so it is reasonable to assume the Y_{ij} follow a *truncated normal distribution* (Mehrotra and others, 2000). For the truncated Y_{ij} values (i.e. $Y_{ij} \leq d$), any parametric distributional assumptions are unverifiable, although most existing literature makes such assumptions for convenience of likelihood inference. Moreover, the truncated values are unlikely to follow normal distributions in most cases, since the Y_{ij} values at least must be positive while a normal random variable can take any real values. Thus, it is more reasonable to assume the truncated normal distribution for the observed Y_{ij} values and leave the distribution of the truncated Y_{ij} values completely unspecified. The

density function of “ $Y_{ij}|\mathbf{b}_{1i}, c_{ij} = 0$ ” is given as [Mehrotra and others \(2000\)](#)

$$f(y_{ij}|c_{ij} = 0, \mathbf{b}_{1i}, \boldsymbol{\beta}, \sigma) = \frac{1}{\sigma} \phi\left(\frac{y_{ij} - \mu_{ij}}{\sigma}\right) \left[1 - \Phi\left(\frac{d - \mu_{ij}}{\sigma}\right)\right]^{-1}, \quad (3.2)$$

where $\mu_{ij} = E(y_{ij}|\mathbf{b}_{1i})$, $\phi(\cdot)$ is the probability density function of the standard normal distribution $N(0, 1)$ and $\Phi(\cdot)$ is the corresponding cumulative distribution function.

For the discrete longitudinal variable Z , we consider the following generalized linear mixed effects model (GLMM)

$$q(E(Z_{ik})) = \mathbf{x}_{ik}^T \boldsymbol{\alpha} + \mathbf{u}_{ik}^T \mathbf{b}_{2i}, \quad i = 1, \dots, n_i, k = 1, \dots, m_i, \quad (3.3)$$

where Z_{ik} is the longitudinal variable of participant i at time t_{ik} , $q(\cdot)$ is a known link function, $\mathbf{x}_{ik}$ and $\mathbf{u}_{ik}$ are vectors of covariates, $\boldsymbol{\alpha}$ are fixed parameters, $\mathbf{b}_{2i}$ is a vector of random effects with $\mathbf{b}_{2i} \sim N(\mathbf{0}, \mathbf{D}_2)$, and $Z_{ik}|\mathbf{b}_{2i}$ is assumed to follow a distribution in the exponential family.

The longitudinal data may contain intermittent missing data and dropouts. We assume that the intermittent missing data and dropouts are missing at random. The fact that the missing data are biomarkers measuring immune responses to the vaccine (and not variables such as toxicity that could obviously be related to missed visits or dropout), and that the vaccine has a large safety data base showing it is not toxic, makes this assumption plausible.

3.1.2. A new approach for truncated longitudinal data. When the Y values are truncated, a common approach in the literature is to assume that the truncated values continue to follow the normal distribution assumed for the observed values ([Hughes, 1999](#); [Wu, 2002](#)). However, such an assumption is unverifiable and may be unreasonable in some cases, as noted earlier. In particular, when the truncation rate is high, the normality assumption is even less reasonable as the truncation rate can be much larger than the left-tail probability of the normal distribution for the observed data. For example, Figure 2 displays histograms of NAb for two participants, where the left truncated data are substituted by the LLOQ of 1.477. The truncation rates, 27% for participant 1 and 33% for participant 2, seem much larger than the left-tail probabilities of the assumed distributions for the observed data.

Fig. 2. Histograms of NAb of two VAX004 vaccine recipients, where the left-truncated data are substituted by the LLOQ of 1.477.

Here we propose a different approach: we do not assume any parametric distributions for the truncated values, but instead we conceptually view the truncated values as a point mass or cluster of unobserved values below the LLOQ without any distributional assumption. Note that, although the truncation status C_{ij} can be determined by the Y_{ij} values, in HIV vaccine studies, many biomarkers are measured infrequently over time, due to both budget and practical considerations, while some other variables can be measured more frequently. For this reason, when the Y values are not measured, we can roughly predict the truncation status of the Y value based on other measured variables that are associated with Y , including time. It is important to predict the truncation status of Y values, since left-truncated Y values have important implications (e.g. a positive immune response may be needed for protection by vaccination). A model for the truncation indicator C_{ij} can help make reasonable predictions of Y_{ij} when such predictions are needed. Therefore, we assume the following model for the truncation indicator C_{ij} :

$$\text{logit}(P(C_{ij} = 1)) = \mathbf{w}_{ij}^T \boldsymbol{\eta} + \mathbf{v}_{ij}^T \mathbf{b}_{3i}, \quad \mathbf{b}_{3i} \sim N(\mathbf{0}, \mathbf{D}_3), \quad (3.4)$$

where $\mathbf{w}_{ij}^T$ and $\mathbf{v}_{ij}^T$ contain covariates, $\boldsymbol{\eta}$ contains fixed parameters, and $\mathbf{b}_{3i}$ contains random effects. The contribution of the longitudinal data of Y for individual i to the likelihood given the random effects is $f(y_{ij}|c_{ij} = 0, \mathbf{b}_{1i}, \boldsymbol{\beta}, \sigma) \times f(c_{ij}|\mathbf{b}_{3i}, \boldsymbol{\eta})$, where $f(y_{ij}|c_{ij} = 0, \mathbf{b}_{1i}, \boldsymbol{\beta}, \sigma)$ is given by (3.2) and $P(c_{ij} = 0|\mathbf{b}_{3i}, \boldsymbol{\eta}) = (1 + \exp(\mathbf{w}_{ij}^T \boldsymbol{\eta} + \mathbf{v}_{ij}^T \mathbf{b}_{3i}))^{-1}$.

Another use of model (3.4) is modeling non-ignorable or informative missing data in the longitudinal Y data. When longitudinal data have both left-truncated data and non-ignorable missing data, we should consider two separate models similar to (3.4). Here we do not consider the issue of non-ignorable missing data, but the models and methods can be easily extended to handle missing data. In fact, left truncated data may be viewed as non-ignorable missing data.

3.1.3. Association between mixed types of longitudinal variables. Different immune response variables are typically highly correlated and may be of different types, such as one being continuous and another one being binary. The exact structures of the associations among different longitudinal variables may be complicated. However, we can reasonably assume that the variables are associated through shared or correlated random effects from different models. This is a reasonable assumption, since the random effects represent individual deviations from population averages and can be interpreted as *unobserved or latent individual characteristics*, such as individual genetic information or health status, which govern different longitudinal processes. This can be seen from Figure 1 where different immune response variables within the same individual exhibit similar patterns over time, including the truncation process. Therefore, we assume that $\mathbf{b}_i = (\mathbf{b}_{1i}^T, \mathbf{b}_{2i}^T, \mathbf{b}_{3i}^T)^T \sim N(\mathbf{0}, \Sigma)$, where Σ is an arbitrary covariance matrix. Note that we allow the random effects in the longitudinal models to be different since the longitudinal trajectories of different variables may exhibit different between-individual variations (as measured by random effects), especially for different types of longitudinal variables such as binary and continuous variables.

3.1.4. A Cox model for time-to-event data. The times to HIV infection may be related to the longitudinal patterns of the immune responses and left-truncated statuses. The specific nature of this dependence may be complicated. There are several possibilities: (i) the infection time may depend on the current immune response values at infection times; (ii) the infection time may depend on past immune response values; and (iii) the infection time may depend on summaries or key characteristics of the longitudinal or truncation trajectories. Here we consider case (iii) for the following reasons: (a) the random effects may be viewed as summaries of individual-specific longitudinal trajectories; (b) the immune response values may be truncated due to lower detection limits; and (c) this approach is also widely used in the joint model literature. Since the random effects in the longitudinal models may be interpreted as “summaries” or

individual-specific characteristics of the longitudinal processes, we may use random effects from the longitudinal models as “covariates” in the survival model. Such an approach is commonly used in the literature and is often called “shared parameter models” (Wulfsohn and Tsiatis, 1997; Rizopoulos, 2012b).

Let T_i^* be the time to HIV infection, C_i be the right-censoring time, $S_i = \min\{T_i^*, C_i\}$ be the observed time, and $\delta_i = I(T_i^* \leq C_i)$ be the event indicator. We assume the censoring is non-informative and consider a Cox model for the observed survival data $\{(s_i, \delta_i), i = 1, \dots, n\}$,

$$h_i(t) = h_0(t) \exp \left(\mathbf{x}_{si}^T \boldsymbol{\gamma}_0 + \mathbf{b}_i^T \boldsymbol{\gamma}_1 \right), \quad (3.5)$$

where $h_0(t)$ is an unspecified baseline hazard function, $\mathbf{x}_{si}$ contains baseline covariates of individual i , and $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$ are vectors of fixed parameters. In model (3.5), the parameters $\boldsymbol{\gamma}_1$ link the risk of HIV infection at time t to the random effects in the longitudinal or truncation models, which allow us to check if individual-specific characteristics of the longitudinal immune responses are associated with the risk of HIV infection. We assume that the survival data and the longitudinal data are conditionally independent given the random effects.

3.2. An approximate method for likelihood inference

We consider the likelihood method for parameter estimation and inference for the above models. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \boldsymbol{\eta}^T, \boldsymbol{\gamma}_0^T, \boldsymbol{\gamma}_1^T)^T$ be the collection of all mean parameters and $\boldsymbol{\xi} = (\sigma, \text{vec}(\Sigma))$ be the collection of variance–covariance (dispersion) parameters. The (joint) likelihood for all the observed longitudinal data and time-to-infection data is given by

$$L(\boldsymbol{\theta}, \boldsymbol{\xi}) = \prod_{i=1}^n \int \left\{ f(\mathbf{y}_i | \mathbf{c}_i = 0, \mathbf{b}_i, \boldsymbol{\beta}, \sigma) f(\mathbf{z}_i | \mathbf{b}_{2i}, \boldsymbol{\alpha}) f(\mathbf{c}_i | \mathbf{b}_{3i}, \boldsymbol{\eta}) f(s_i, \delta_i | h_0, \mathbf{b}_i, \boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1) f(\mathbf{b}_i | \Sigma) \right\} d \mathbf{b}_i.$$

Since the dimension of the random effects $\mathbf{b}_i$ is often high and some density functions can be highly complicated, evaluation of the above integral can be a major challenge. The common approach based on the Monte Carlo EM algorithm can offer potential difficulties such as very slow or even non-convergence (Hughes, 1999). Numerical integration methods such as the Gaussian Hermite (GH) quadrature method can also be very tedious. Therefore, in the following we consider an approximate method based on the h-likelihood, which can be computationally much more efficient while maintaining reasonable accuracy (Lee and others, 2006; Ha and others, 2003; Molas and others, 2013). Its performance in the current context will be evaluated by simulations later.

Essentially, the h-likelihood method uses Laplace approximations to the intractable integral in the likelihood. A first-order Laplace approximation can be viewed as the GH quadrature method with one node. So a Laplace approximation can be less accurate than the GH quadrature method with more than one node. However, when the dimension of the integral is high, a Laplace approximation can be computationally much less intensive than the GH quadrature method whose computational intensity grows exponentially with the dimension of the integral. Moreover, it produces approximate MLEs for the mean parameters and approximate restricted maximum likelihood estimates (REMLs) for the variance–covariance (dispersion) parameters. For the models (3.1), (3.3)–(3.5) in the previous section, the log h-likelihood function is given by

$$\begin{aligned} \ell_h = \sum_{i=1}^n \ell_{hi} = \sum_{i=1}^n \left\{ \log f(\mathbf{y}_i | \mathbf{c}_i = 0, \mathbf{b}_i, \boldsymbol{\beta}, \sigma) + \log f(\mathbf{z}_i | \mathbf{b}_{2i}, \boldsymbol{\alpha}) + \log f(\mathbf{c}_i | \mathbf{b}_{3i}, \boldsymbol{\eta}) \right. \\ \left. + \log f(s_i, \delta_i | h_0, \mathbf{b}_i, \boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1) + \log f(\mathbf{b}_i | \Sigma) \right\}. \end{aligned} \quad (3.6)$$

Based on [Ha and others \(2003\)](#) and [Molas and others \(2013\)](#), we propose the following estimation procedure via the h-likelihood. Beginning with some starting values $(\hat{\xi}^{(0)}, \hat{\theta}^{(0)}, h_0^{(0)})$, we iterate the steps below:

- 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 (3.6) with respect to $\mathbf{b}$;
- Step 2: Given $(\hat{\xi}^{(k)}, \hat{h}_0^{(k)}, \hat{\mathbf{b}}^{(k+1)})$, obtain updated estimates of the mean parameters $\boldsymbol{\theta}^{(k+1)}$ by maximizing the following adjusted profile h-likelihood as in [Lee and Nelder \(1996\)](#) with respect to $\boldsymbol{\theta}$:

$$p_{\theta} = \left(\ell_h - 0.5 \log \left| \frac{H(\ell_h, \mathbf{b})}{2\pi} \right| \right) \Bigg|_{\xi=\hat{\xi}^{(k)}, h_0=\hat{h}_0^{(k)}, \mathbf{b}=\hat{\mathbf{b}}^{(k+1)}}, \quad \text{where } H(\ell_h, \mathbf{b}) = -\partial^2 \ell_h / \partial \mathbf{b}^T \partial \mathbf{b}.$$

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

$$p_{\xi} = \left(\ell_h - 0.5 \log \left| \frac{H[\ell_h, (\boldsymbol{\theta}, \mathbf{b})]}{2\pi} \right| \right) \Bigg|_{h_0=\hat{h}_0^{(k)}, \boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(k+1)}, \mathbf{b}=\hat{\mathbf{b}}^{(k+1)}},$$

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

- 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.

By iterating the above four steps until convergence, we can obtain approximate MLEs for the mean parameters, approximate REMLs for the variance-covariance parameters, empirical Bayes estimates of the random effects, and a nonparametric estimate of the baseline hazard function. To set starting values, we may first fit the models separately and then choose the resulting parameter estimates as the starting values for $\xi^{(0)}, \theta^{(0)}, h_0^{(0)}$. More details are described in Section 4.3. The standard errors of the parameter estimates can be obtained based on

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\theta}}, \hat{\mathbf{b}}) = H^{-1}[\ell_h, (\hat{\boldsymbol{\theta}}, \hat{\mathbf{b}})] = - \begin{pmatrix} \frac{\partial^2 \ell_h}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} & \frac{\partial^2 \ell_h}{\partial \boldsymbol{\theta} \partial \mathbf{b}^T} \\ \frac{\partial^2 \ell_h}{\partial \mathbf{b} \partial \boldsymbol{\theta}^T} & \frac{\partial^2 \ell_h}{\partial \mathbf{b} \partial \mathbf{b}^T} \end{pmatrix}^{-1} \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}, \mathbf{b}=\hat{\mathbf{b}}}.$$

That is, the estimated variances of $\hat{\boldsymbol{\theta}}$ can be chosen to be the diagonal elements of the top left corner of the matrix $\widehat{\text{Cov}}(\hat{\boldsymbol{\theta}}, \hat{\mathbf{b}})$ ([Lee and Nelder, 1996](#); [Ha and others, 2003](#)).

As mentioned in Section 1, the standard errors of parameter estimates may be under-estimated when the baseline hazard $h_0(t)$ is unspecified (Rizopoulos, 2012a; Hsieh and others, 2006). A bootstrap method for obtaining standard errors is a good choice, but it is computationally intensive. Thus, here we propose a new approach to estimate the standard errors of parameter estimates based on the adaptive Gauss–Hermite (aGH) method (Rizopoulos, 2012a; Hartzel and others, 2001; Pinheiro and Bates, 1995). The basic idea is as follows. After convergence of the above steps, we can approximate the score function of θ for subject i by the following:

$$\begin{aligned} S_i(\theta) &= \frac{\partial}{\partial \theta^T} \log \int \left\{ f(\mathbf{y}_i | \mathbf{c}_i = 0, \mathbf{b}_{1i}, \boldsymbol{\beta}, \sigma) f(\mathbf{z}_i | \mathbf{b}_{2i}, \boldsymbol{\alpha}) f(\mathbf{c}_i | \mathbf{b}_{3i}, \boldsymbol{\eta}) f(\mathbf{s}_i, \boldsymbol{\delta}_i | h_0, \mathbf{b}_i, \boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1) f(\mathbf{b}_i | \Sigma) \right\} d\mathbf{b}_i \\ &= \int f(\mathbf{b}_i | \mathbf{y}_i, \mathbf{c}_i, \mathbf{z}_i, \mathbf{s}_i, \boldsymbol{\delta}_i, \boldsymbol{\theta}, \sigma, \Sigma, h_0) \times A(\boldsymbol{\theta}, \mathbf{b}_i) d\mathbf{b}_i \\ &\propto \sum_{k=1}^{q^K} \left\{ \pi_k \times f(\mathbf{b}_i^{(k)} | \mathbf{y}_i, \mathbf{c}_i, \mathbf{z}_i, \mathbf{s}_i, \boldsymbol{\delta}_i, \boldsymbol{\theta}, \hat{\sigma}, \hat{\Sigma}_i, \hat{h}_0) \times e^{\mathbf{z}_i^{(k)T} \mathbf{z}_i^{(k)}} \times A(\boldsymbol{\theta}, \mathbf{b}_i^{(k)}) \right\}, \end{aligned}$$

where q is the dimension of the random effects, K is the number of quadrature points for each random effect, π_k are weights for the original GH nodes $\mathbf{z}_i^{(k)}$, $\mathbf{b}_i^{(k)} = \hat{\mathbf{b}}_i + \sqrt{2\hat{\Sigma}_i} \mathbf{z}_i^{(k)}$ with $\hat{\Sigma}_i$ being the upper triangular factor of the Cholesky decomposition of $\hat{\Sigma}_i = (-\partial^2 \ell_{hi} / \partial \mathbf{b}_i \partial \mathbf{b}_i^T)^{-1} |_{\mathbf{b}_i = \hat{\mathbf{b}}_i}$, and $A(\boldsymbol{\theta}, \mathbf{b}_i^{(k)}) = \partial \ell_{hi} / \partial \boldsymbol{\theta}^T |_{\sigma = \hat{\sigma}, h_0 = \hat{h}_0, \Sigma = \hat{\Sigma}_i, \mathbf{b}_i = \mathbf{b}_i^{(k)}}$. Then, the standard errors of the parameter estimates can be estimated based on $\widehat{\text{Var}}(\hat{\boldsymbol{\theta}}) = \left\{ -\sum_{i=1}^n \frac{\partial S_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\}^{-1} \Big|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}}$. In practice, we can calculate $\partial S_i(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ numerically using the central difference approximation (Rizopoulos, 2012a).

4. DATA ANALYSIS

4.1. HIV vaccine data and new time variables

In this section, we analyze the VAX004 data set described in Section 2, based on the proposed models and methods. Our objective is to check if individual-specific longitudinal characteristics of immune responses are associated with the risk of HIV infection. A comprehensive analysis may be infeasible due to space limitation, but we will focus on the essential features of the data. Since the immune response variables are mostly highly correlated, we choose two variables, “MNGNE8” and “NAb”, which may represent the key features of the longitudinal immune response data. Note that some variables are often conveniently converted to binary data for simpler clinical interpretations. Here, we let Z_{ij} be the dichotomized MNGNE8 data such that $Z_{ij} = 1$ if the MNGNE8 value of individual i at time t_{ij} is larger than the sample median 0.57 and $Z_{ij} = 0$ otherwise. Let Y_{ij} be the original NAb value of individual i at time t_{ij} . Recall that 27% of the original NAb values are below this variable’s LLOQ (i.e. left truncated).

A unique feature of vaccine trial data is that the longitudinal immune response data typically exhibit *periodic patterns*, due to repeated administration of the vaccine. This can be clearly seen in Figure 1. Statistical modeling must incorporate these features. Here we use a simple periodic function $\sin(\cdot)$ to empirically capture the periodic patterns and further define the following time variables (in months): (i) the time from the beginning of the study to the current scheduled measurement time, denoted by t_{ij} ; (ii) the time from the most recent immunization to the current scheduled measurement time, denoted by t_{dij} (so $t_{dij} \leq t_{ij}$); and (iii) the time between two consecutive vaccine administrations, denoted by Δ_{ij} , so there will be at least one t_{dij} and one t_{ij} between Δ_{ij} and $\Delta_{i,j+1}$. For measurement time t_{ij} scheduled after the final vaccination, we define Δ_{ij} as the time between the final vaccination and the final measurement time. These different time variables are needed in modeling the longitudinal trajectories. Figure 3 gives an example

Fig. 3. Illustration of three time variables in VAX004. The cross symbols indicate the measurement times of subject i . The dashed vertical lines show the scheduled times of vaccinations and the end of study (i.e. month 0, 1, 6, 12, 18, 24, 30, 36), where the black dashed lines represent the times when subject i received vaccines and the gray dashed lines represent the times when subject i missed the scheduled vaccinations. The arrow lines represent the time periods t_{ij} , t_{dij} , Δ_{ij} of the sixth and eighth measurements with $j = 6$ and $j = 8$, respectively.

of how different time variables are defined for a randomly chosen participant i . Recall that vaccinations are scheduled at months 0, 1, 6, 12, 18, 24, 30, and the study ends at month 36. For this participant i , s/he receives the first four vaccinations, but then drops out from the study before receiving the fifth vaccination. There are eight measurements over time in total, denoted by the cross symbols, where the measurement times may be different from the vaccination times. Suppose that the sixth measurement is taken at month 9, i.e. $t_{i6} = 9$, then we have $t_{di6} = t_{i6} - 6 = 3$, the difference between the sixth measurement time and the latest vaccination time (Vac 3 at month 6) for this participant, and $\Delta_{i6} = 12 - 6 = 6$, since the sixth measurement happens between the third vaccination (Vac 3 at month 6) and the fourth vaccination (Vac 4 at month 12). To avoid very large or small parameter estimates, we also re-scale the times as follows: $t_{dij}^* = t_{dij} * 30/7$ (in weeks) and $t_{ij}^* = t_{ij}/12$ (in years).

4.2. Models

Based on rationales discussed in Sections 3 and 4.1, we consider empirical models for the continuous and binary longitudinal data and survival model. The longitudinal models are selected based on AIC values (see details in Section 3 in the [Supplementary material](#) available at *Biostatistics* online). For the NAb data with 27% truncation, we model the untruncated data by using the LME model

$$Y_{ij} | C_{ij} = 0 = \beta_0 + \beta_1 t_{ij}^* + \beta_2 t_{ij}^{*2} + \beta_3 \sin\left(\frac{\pi}{\Delta_{ij}} t_{dij}\right) + \beta_4 risk1_i + \beta_5 risk2_i + d_1 b_{1i} + \epsilon_{ij}, \quad (4.1)$$

where $risk1_i$ and $risk2_i$ are categories 1 and 2 of baseline behavioral risk score, the random effects $b_{1i} \sim N(0, 1)$, d_1 is the variance parameter, ϵ_{ij} follows a truncated normal distribution with mean 0 and variance σ^2 , and $Y_{ij}|C_{ij} = 0$ follows a *truncated normal distribution*. To ensure identifiability of the models, we assume that $d_1 > 0$. We only consider a random intercept in the model because adding more random effects does not substantially reduce AIC values while making the models more complicated.

We also model the truncation indicator, C_{ij} , of NAb to find possible associations of truncation with the time variables and other covariates and to predict the truncation status of NAb at times when NAb values

are unavailable. The selected model is given as below,

$$\text{logit}(P(C_{ij} = 1)) = \eta_0 + \eta_1 t_{ij}^* + \eta_2 t_{ij}^{*2} + \eta_3 \sin\left(\frac{\pi}{\Delta_{ij}} t_{dij}\right) + \eta_4 \text{risk}1_i + \eta_5 \text{risk}2_i + \eta_6 b_{1i}, \quad (4.2)$$

which shares the same random effect as the NAb model (4.1), since these two processes seem to be highly correlated with each other. In many studies, the Y_{ij} and C_{ij} values are measured sparsely and we can use model (4.2) to predict the truncation status of Y_{ij} at times when Y-measurements are unavailable.

For the binary MNGNE8 data, variable selections by AIC values lead to the model

$$\text{logit}(P(Z_{ij} = 1)) = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 \sin\left(\frac{\pi}{\Delta_{ij}} \times t_{dij}\right) + \alpha_3 t_{dij}^* + \alpha_4 b_{1i} + d_2 b_{2i} t_{ij}, \quad (4.3)$$

where d_2 is the variance parameter with $d_2 > 0$ and the individual characteristics are incorporated via random slope $b_{2i} \sim N(0, 1)$ and random intercept b_{1i} shared by models (4.1)–(4.2).

The association among the longitudinal models is incorporated through shared and correlated random effects from different models: $\mathbf{b}_i = (b_{1i}, b_{2i})^T \sim N(0, \Sigma)$, with $\Sigma = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}$ and $-1 \leq r_{12} \leq 1$. Note that the random effect b_{1i} is shared by all the longitudinal models, since all the immune response longitudinal data exhibit similar individual-specific patterns and the random effect b_{1i} for the continuous NAb data best summarizes these patterns. For example, when a participant has a high baseline measurement of NAb, s/he likely also has a high baseline value of MNGNE8 and a low baseline probability that NAb is left truncated.

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

$$h_i(t|x_i, \mathbf{b}_i) = h_0(t) \exp\left\{\gamma_0 x_i + \gamma_1 \text{risk}1_i + \gamma_2 \text{risk}2_i + \gamma_3 b_{1i} + \gamma_4 b_{2i}\right\}, \quad (4.4)$$

where x_i is the measurement of GNE8_CD4 (i.e. blocking of the binding of the GNE8 HIV-1 gp120 Env protein to soluble CD4) for individual i on the first day of the study after the first immunization, rescaled to have a mean of 0 and a standard deviation of 1. We call x_i the standardized baseline GNE8_CD4. Since the analysis in this section is exploratory in nature, for simplicity we ignore other covariates.

4.3. Parameter estimates, model diagnostics, and new findings

We estimate model parameters using the proposed h-likelihood method. As a comparison, we also use the two-step method, which fits each longitudinal model separately and obtains random effect estimates in the first step and then in the second step the random effects in the Cox model are simply substituted by their estimates from the first step. The results of the two-step method are obtained using the R packages *lme4* and *survival*. The drawbacks of the two-step method are: (i) it may under-estimate the standard errors of the parameter estimates in the survival model, since it fails to incorporate the estimation uncertainty in the first step; (ii) it fails to incorporate the associations among the longitudinal variables; and (iii) it may lead to biased estimates of longitudinal model parameters when longitudinal data are terminated by event times and/or truncated longitudinal data are simply replaced by the LLOQ or half this limit (Wu, 2009). Table 1 summarizes estimation results based on the above two methods. Algorithms based on the h-likelihood method were terminated when the relative change became less than 10^{-2} in the estimates or 10^{-3} in the approximated log-likelihood. Since our main objective is to investigate if individual-specific characteristics of the longitudinal immune responses are associated with the risk of HIV infection, we mainly focus on $\hat{\gamma}_3$ and $\hat{\gamma}_4$ in the survival model (4.4) as these parameters link the random effects to the hazard of HIV infection.

Table 1. Estimates of all model parameters in VAX004

Model	Par	Two-step method			H-likelihood method		
		Est	SE	p-value	Est	SE ^a _(SE^b)	p-value ^a _(p-value^b)
<i>Estimates of mean parameters</i>							
LME model (4.1) for NAb	β_0	1.57	0.05	< 0.01	2.35	0.06 _(0.07)	< 0.01 _(<0.01)
	β_1	1.89	0.05	< 0.01	0.95	0.06 _(0.08)	< 0.01 _(<0.01)
	β_2	−0.54	0.02	< 0.01	−0.31	0.02 _(0.03)	< 0.01 _(<0.01)
	β_3	0.55	0.05	< 0.01	1.46	0.07 _(0.10)	< 0.01 _(<0.01)
	β_4	0.004	0.04	0.93	−0.04	0.02 _(0.05)	0.15 _(0.49)
	β_5	−0.27	0.12	0.03	−0.10	0.08 _(0.18)	0.18 _(0.56)
Truncation model (4.2)	η_0	2.09	0.17	< 0.01	1.94	0.20 _(0.24)	< 0.01 _(<0.01)
	η_1	−6.52	0.31	< 0.01	−6.27	0.29 _(0.39)	< 0.01 _(<0.01)
	η_2	1.71	0.11	< 0.01	1.65	0.10 _(0.14)	< 0.01 _(<0.01)
	η_3	−0.64	0.22	< 0.01	−0.61	0.22 _(0.30)	< 0.01 _(0.04)
	η_4	−0.09	0.14	0.50	−0.03	0.14 _(0.22)	0.84 _(0.90)
	η_5	1.15	0.38	< 0.01	0.96	0.37 _(0.64)	0.01 _(0.13)
GLMM (4.3) for MNGNE8	α_0	−1.60	0.10	< 0.01	−1.68	0.10 _(0.14)	< 0.01 _(<0.01)
	α_1	0.11	0.01	< 0.01	0.14	0.01 _(0.01)	< 0.01 _(<0.01)
	α_2	1.76	0.19	< 0.01	1.82	0.18 _(0.26)	< 0.01 _(<0.01)
	α_3	−0.05	0.01	< 0.01	−0.05	0.01 _(0.02)	< 0.01 _(<0.01)
Survival model (4.4)	γ_0	−0.72	0.23	< 0.01	−0.79	0.24 _(0.37)	< 0.01 _(0.03)
	γ_1	0.70	0.56	0.21	−0.28	0.39 _(0.69)	0.48 _(0.69)
	γ_2	1.46	1.14	0.20	1.86	1.07 _(1.33)	0.08 _(0.16)
	γ_3	−0.01	0.24	0.96	−1.69	0.32 _(0.92)	< 0.01 _(0.07)
	γ_4	0.13	0.23	0.58	2.37	0.31 _(0.80)	< 0.01 _(<0.01)
<i>Estimates of variance-covariance parameters</i>							
	σ	0.66			0.48		
	d_1	0.22			0.54		
	η_6	−0.89			−1.60		
	α_4	0.43			0.004		
	d_2	0.05			0.19		
	r_{12}	0.37			0.76		

SE^a and p-value^a: Standard error and p-value based on the h-likelihood method.

SE^b and p-value^b: Standard error and p-value based on the newly proposed method with 4 quadrature points.

From Table 1, we see that the two methods lead to quite different results, especially the estimates of γ_3 and γ_4 in the survival model that are our main focus in this analysis. For the two-step method, the estimates of γ_3 and γ_4 are near zero with confidence intervals including zero, not supporting that individual-specific immune response longitudinal trajectories are associated with the risk of HIV infection. However, these parameter estimates based on the proposed joint model with the h-likelihood method lead to different conclusions. Both $\hat{\gamma}_3$ and $\hat{\gamma}_4$ are highly significant based on the standard errors estimated by the joint model with the h-likelihood method (denoted as SE^a), suggesting that individual-specific immune response longitudinal trajectories are highly associated with the risk of HIV infection. Since the standard errors based on the h-likelihood method may be under-estimated (Hsieh and others, 2006; Rizopoulos, 2012b), as discussed earlier, we also calculate the standard errors using the proposed method based on

the aGH method, and the results with four quadrature points are given as SE^b in the table. We see that, based on the new standard errors, the p-value for testing $H_0 : \gamma_3 = 0$ is slightly larger than 0.05 while that for testing $H_0 : \gamma_4 = 0$ is still highly significant. Therefore, we may conclude that individual-specific immune response longitudinal trajectories are associated with the risk of HIV infection. This conclusion is unavailable based on the two-step method.

The negative estimate of γ_3 suggests that higher NAb values are associated with a lower risk of HIV infection, and the positive estimate of γ_4 suggests that large increases in MNGNE8 over time are associated with a higher risk of HIV infection. Specifically, there is an estimated 81.6% decrease (i.e. $\exp(\hat{\gamma}_3) - 1$) in the hazard/risk with a one unit increase in the individual effect b_{1i} and an estimated 10.6 times increase (i.e. $\exp(\hat{\gamma}_4)$) in the hazard/risk with a one unit increase in the individual-specific slope b_{2i} , holding other covariates constant. These findings are original, since they are unavailable based on the two-step method, and show the important contribution of the proposed joint model and the h-likelihood method.

The joint model method and the two-step method have consistent significances of the parameters in the longitudinal models (4.1)–(4.3), except for β_5 and η_5 . By the two-step method, the tests for $H_0 : \beta_5 = 0$ and for $H_0 : \eta_5$ yield significant p-values, suggesting that participants with baseline behavioral risk score in category 2 (i.e. $\text{risk2} = 1$) have significantly lower NAb values than other participants. By the joint model method, on the other hand, such a negative association is not statistically significant. For model (4.1), the mean square error (MSE) based on the joint model is 0.296, while the MSE based on the two-step method is 0.403.

The model diagnostics are conducted to check the assumptions and goodness-of-fit of the models. The results are listed in Section 4 in the [Supplementary material](#) available at *Biostatistics* online. Overall, the assumptions hold and the models fit the data well. The data used in this example may be requested through a concept proposal to the owner of the data—Global Solutions in Infectious Diseases.

5. SIMULATION STUDIES

In this section, we conduct three simulation studies to evaluate the proposed joint model with the h-likelihood method. The models and their true parameter values in the simulation studies are chosen to be similar to the estimated values in the models for real data in the previous section.

5.1. Simulation study 1

Conditional on the random effects, the binary data Z_{ij} are generated from a Bernoulli distribution with probabilities $p_{ij} = \exp(\xi_{ij}) / (1 + \exp(\xi_{ij}))^{-1}$, where $\xi_{ij} = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 \sin(\frac{\pi}{\Delta_{ij}} t_{dij}) + \alpha_3 t_{dij}^* + \alpha_4 b_{1i} + d_2 b_{2i} t_{ij}$. For the continuous data Y_{ij} , we first randomly generate y_{ij} from a normal distribution $Y_{ij} | \mathbf{b}_{1i} \sim N(\mu_{ij}, \sigma^2)$ with $\mu_{ij} = \beta_0 + \beta_1 t_{ij}^* + \beta_2 t_{ij}^{*2} + \beta_3 \sin(\frac{\pi}{\Delta_{ij}} t_{dij}) + d_1 b_{1i}$. Then we create truncations so that y_{ij} is observed if $y_{ij} > LLOQ$ and truncated otherwise and we choose $LLOQ = 2$. The random effects $\mathbf{b}_i = (b_{1i}, b_{2i})$ are generated from a multivariate normal distribution $N(0, \Sigma)$. The true values of the parameters are set to be: $\boldsymbol{\alpha}^T = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4) = (-1.65, 0.15, 1.8, -0.05, 0.4)$, $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2, \beta_3) = (2, 1, -0.3, 1.5)$, $\sigma = 0.5$, $d_1 = 0.5$, $d_2 = 0.15$, $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$. The survival times T_i are generated from a Weibull distribution with shape parameter of 15 and scale parameter of $800 \exp(\gamma_0 X_i + \gamma_1 b_{1i} + \gamma_2 b_{2i})^{-1/15}$, where X_i is a baseline covariate generated from the standard normal distribution. The non-informative censoring times C_i are generated from a Weibull distribution with shape parameter of 5 and scale parameter of 1000. The true parameter values are given as $\boldsymbol{\gamma}^T = (\gamma_0, \gamma_1, \gamma_2) = (-0.75, -1.5, 2)$.

Table 2. Simulation results with bi-weekly longitudinal measurements based on the two-step (TS) method and the h-likelihood (HL) method

Model	Par	True value	Estimate		SSE		rBias (%)		rMSE (%)		Coverage probability (%)		
			TS	HL	TS	HL	TS	HL	TS	HL	TS	HL ^a	HL ^b
(4.1)	β_0	2.00	2.14	2.01	0.04	0.06	7.05	0.55	1.07	0.16	15	94	98
	β_1	1.00	0.88	1.00	0.04	0.04	11.55	0.20	1.47	0.20	13	94	99
	β_2	-0.30	-0.26	-0.30	0.02	0.02	13.07	0.00	0.60	0.12	32	94	99
	β_3	1.50	1.41	1.49	0.02	0.02	5.70	0.37	0.52	0.04	1	94	98
(4.3)	α_0	-1.65	-1.64	-1.64	0.10	0.10	0.57	0.59	0.62	0.63	94	95	99
	α_1	0.15	0.15	0.15	0.02	0.02	2.02	2.45	0.18	0.17	94	89	96
	α_2	1.80	1.80	1.78	0.11	0.10	0.26	0.84	0.61	0.61	96	95	99
	α_3	-0.05	-0.05	-0.05	0.01	0.01	1.80	2.68	0.06	0.06	93	93	97
(4.4)	γ_0	-0.75	-0.68	-0.74	0.16	0.17	9.44	0.74	4.00	3.76	88	85	97
	γ_1	-1.50	-1.10	-1.44	0.22	0.29	26.91	3.99	13.96	5.92	38	65	84
	γ_2	2.00	1.53	2.02	0.23	0.31	23.40	1.10	13.70	4.88	34	69	87

HL^a: Coverage probability based on the h-likelihood method.HL^b: Coverage probability based on the newly proposed method for standard errors with 4 quadrature points.

5.2. Simulation studies 2 and 3

To better evaluate the performance of the h-likelihood method, we conducted two additional simulation studies: (i) a joint model with higher dimensions of random effects (Study 2, four random effects); and (ii) a joint model with a parametric survival model (Study 3, a Weibull survival model). For the parametric joint model in Study 3, we also estimate the model parameters using the aGH method as comparison. Due to space limitation, we put the details of these two simulation studies in Sections 5 and 6 in the [Supplementary material](#) available at *Biostatistics* online.

5.3. Simulation results and discussions

We compare the performance of the methods based on the relative bias and MSE of the parameter estimates, which are defined as follows (say, for parameter β):

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

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

Table 2 summarizes the results of Simulation Study 1 ($M = 500$) when the longitudinal measurements are collected bi-weekly. The proposed h-likelihood method outperforms the two-step method as it returns much less biased estimates for most of the parameters. As for the bias in $\hat{\alpha}$, it is known that the h-likelihood method may perform less satisfactorily for logistic mixed effects models (Kuk and Cheng, 1999; Waddington and Thompson, 2004). The standard errors of $\hat{\gamma}_j$'s seem to be underestimated by the h-likelihood method. This problem has been reported elsewhere (Hsieh and others, 2006; Rizopoulos, 2012b). However, our newly proposed method based on the aGH method with 4 quadrature points returns

coverage probabilities much closer to the nominal coverage probabilities. The results of Simulation Studies 2 and 3 are given in Tables S5 and S6 in the [Supplementary material](#) available at *Biostatistics* online. The main conclusions are consistent with those from Table 2. In Simulation Study 3, the $\hat{\gamma}_j$'s based on the h-likelihood method are much less biased, though with slightly larger MSEs, than those based on the aGH method, while for β_3 , α_0 , and α_2 , the aGH method has less biased estimates than the h-likelihood method (see Table S6 in the [Supplementary material](#) available at *Biostatistics* online). Synthesizing the simulation results, we conclude that the proposed h-likelihood method, with the new approach of estimating the standard errors, performs reasonably well. Its performance remains consistent with higher dimensions of random effects and parametric survival models. Although it is sometimes less accurate than the aGH method, it is computationally much more efficient.

6. DISCUSSION

In this article, we have considered a joint model for mixed types of longitudinal data with left truncation and a survival model and proposed a new method to handle the left-truncation in longitudinal data. A main advantage of this method, compared with existing methods in the literature (e.g. [Hughes, 1999](#)), is that it does not make any untestable distributional assumption for the truncated data that are below a measurement instruments LLOQ. Different types of longitudinal data are assumed to be associated via shared and correlated random effects. We have also proposed an h-likelihood method for approximate joint likelihood inference, which is computationally much more efficient than the aGH method. Moreover, we have proposed a new method to better estimate the standard errors of parameter estimates from the h-likelihood method. Based on a MacBook Pro Version 10.11.4, the average computing times of the h-likelihood method were 2.7 min for the semiparametric joint model with 2 random effects and 21.9 min for the semiparametric joint model with 4 random effects, respectively. For the parametric joint model with 2 random effects, the average running time of the h-likelihood method was 9.1 min, much faster than the aGH method that takes 28.4 min.

Analysis of the real HIV vaccine data based on the proposed method shows that the individual-specific characteristics of longitudinal immune response, summarized by random effects in the models, are highly associated with the risk of HIV infection. This finding is quite interesting and helpful to designing future HIV vaccine studies. We have also proposed a model for the left-truncation indicator of the longitudinal immune response data and showed that the left-truncation status follows certain patterns as functions of time. Such a model can be used to predict the left-truncation status (below LLOQ status) of some longitudinal immune response values when measurement schedules are infrequent or sparse.

The joint model in this article may be extended in several directions. For example, the Cox model may be replaced by an accelerated failure time model or survival model for interval censored data or competing risks data. The association among different types of longitudinal processes may also be modeled in other ways such as shared latent processes. In addition, the dropouts in the real data may be associated with longitudinal patterns, so we may consider incorporating missing data mechanisms into the joint models in future research. Research for these extensions will be reported separately.

7. SOFTWARE

Software in the form of R code and a sample input data set are available at <https://github.com/oliviayu/HHJMs>.

SUPPLEMENTARY MATERIAL

[Supplementary material](#) is available at <http://biostatistics.oxfordjournals.org>.

ACKNOWLEDGMENTS

The authors thank the reviewers for the thoughtful comments to help improve the article greatly. 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 trial, including Global Solutions for Infectious Diseases. *Conflict of Interest*: None declared.

FUNDING

National Institute Of Allergy And Infectious Diseases of the National Institutes of Health (NIH) (Award Numbers R37AI054165 and UM1AI068635); and Bill and Melinda Gates Foundation (BMGF) (Award Number OPP1110049).

REFERENCES

- BARRETT, J., DIGGLE, P., HENDERSON, R. AND TAYLOR-ROBINSON, D. (2015). Joint modelling of repeated measurements and time-to-event outcomes: flexible model specification and exact likelihood inference. *Journal of the Royal Statistical Society: Series B* **77**, 131–148.
- BERNHARDT, P. W., WANG, H. J., AND ZHANG, D. (2014). Flexible modeling of survival data with covariates subject to detection limits via multiple imputation. *Computational Statistics and Data Analysis*, **69**, 81–91.
- CHEN, Q., MAY, R. C., IBRAHIM, J. G., CHU, H., AND COLE, S. R. (2014). Joint modeling of longitudinal and survival data with missing and left-censored time-varying covariates. *Statistics in Medicine*, **33**, 4560–4576.
- ELASHOFF, R. M., LI, G., AND LI, N. (2015). *Joint Modeling of Longitudinal and Time-to-Event Data*. Boca Raton, FL: Chapman & Hall/CRC.
- FLYNN, N., FORTHALL, D., HARRO, C., JUDSON, F., MAYER, K., PARA, M., AND GILBERT, P.; The rgp120 HIV Vaccine Study Group (2005). Placebo-controlled phase 3 trial of recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *Journal of Infectious Diseases*, **191**, 654–65.
- FU, R. AND GILBERT, P. B. (2017). Joint modeling of longitudinal and survival data with the cox model and two-phase sampling. *Lifetime Data Analysis*, **23**, 136–159.
- GILBERT, P. B., PETERSON, M. L., FOLLMANN, D., HUDGENS, M. G., FRANCIS, D. P., GURWITH, M., HEYWARD, W. L., JOBES, D. V., POPOVIC, V., SELF, S. G., 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**, 666–677.
- HA, I. D., PARK, T., AND LEE, Y. (2003). Joint modelling of repeated measures and survival time data. *Biometrical Journal*, **45**, 647–658.
- HARTZEL, J., AGRESTI, A., AND CAFFO, B. (2001). Multinomial logit random effects models. *Statistical Modelling*, **1**, 81–102.
- HSIEH, F., TSENG, Y.-K., AND WANG, J.-L. (2006). Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics*, **62**, 1037–1043.
- HUGHES, J. P. (1999). Mixed effects models with censored data with application to hiv rna levels. *Biometrics*, **55**, 625–629.
- KRÓL, A., FERRER, L., PIGNON, J.-P., PROUST-LIMA, C., DUCREUX, M., BOUCHÉ, O., MICHIELS, S., AND RONDEAU, V. (2016). Joint model for left-censored longitudinal data, recurrent events and terminal event: predictive abilities of tumor burden for cancer evolution with application to the ffcd 2000–05 trial. *Biometrics*, **72**, 907–916.
- KUK, A. Y. AND CHENG, Y. W. (1999). Pointwise and functional approximations in monte carlo maximum likelihood estimation. *Statistics and Computing*, **9**, 91–99.

- LAWRENCE GOULD, A., BOYE, M. E., CROWTHER, M. J., IBRAHIM, J. G., QUARTEY, G., MICALLEF, S., AND BOIS, F. Y. (2015). Joint modeling of survival and longitudinal non-survival data: current methods and issues. report of the dia bayesian joint modeling working group. *Statistics in Medicine*, **34**, 2181–2195.
- LEE, Y. AND NELDER, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series B*, **58**, 619–678.
- LEE, Y., NELDER, J. A., AND PAWITAN, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Boca Raton, FL: Chapman & Hall/CRC.
- MEHROTRA, K. G., KULKARNI, P. M., TRIPATHI, R. C., AND MICHALEK, J. E. (2000). Maximum likelihood estimation for longitudinal data with truncated observations. *Statistics in Medicine*, **19**, 2975–2988.
- MOLAS, M., NOH, M., LEE, Y., AND LESAFFRE, E. (2013). Joint hierarchical generalized linear models with multivariate gaussian random effects. *Computational Statistics and Data Analysis*, **68**, 239–250.
- PINHEIRO, J. C. AND BATES, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, **4**, 12–35.
- RIZOPOULOS, D. (2012a). Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive gaussian quadrature rule. *Computational Statistics and Data Analysis*, **56**, 491–501.
- RIZOPOULOS, D. (2012b). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Boca Raton, FL: Chapman & Hall/CRC.
- RIZOPOULOS, D., VERBEKE, G., AND LESAFFRE, E. (2009). Fully exponential laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society: Series B*, **71**, 637–654.
- TAYLOR, J. M., PARK, Y., ANKERST, D. P., PROUST-LIMA, C., WILLIAMS, S., KESTIN, L., BAE, K., PICKLES, T., AND SANDLER, H. (2013). Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics*, **69**, 206–213.
- WADDINGTON, D. AND THOMPSON, R. (2004). Using a correlated probit model approximation to estimate the variance for binary matched pairs. *Statistics and Computing*, **14**, 83–90.
- 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**, 955–964.
- WU, L. (2009). *Mixed Effects Models for Complex Data*. Boca Raton, FL: Chapman & Hall/CRC.
- WULFSOHN, M. S. AND TSIATIS, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330–339.
- ZHU, H., IBRAHIM, J. G., CHI, Y.-Y., AND TANG, N. (2012). Bayesian influence measures for joint models for longitudinal and survival data. *Biometrics*, **68**, 954–964.

[Received November 3, 2016; revised August 9, 2017; accepted for publication August 23, 2017]