In jointly modelling longitudinal and survival data, the longitudinal data may be complex in the sense that they may contain outliers and may be left censored. Motivated from an HIV vaccine study, we propose a robust method for joint models of longitudinal and survival data, where the outliers in longitudinal data are addressed using a multivariate t-distribution for b-outliers and using an M-estimator for e-outliers. We also propose a computationally efficient method for approximate likelihood inference. The proposed method is evaluated by simulation studies. Based on the proposed models and method, we analyze the HIV vaccine data and find a strong association between longitudinal biomarkers and the risk of HIV infection.

1. Introduction. In HIV vaccine studies, joint models of longitudinal and survival data are instrumental, since it is crucial to study the association between longitudinal immune responses biomarkers during repeated immunizations over time and the risk of HIV infection. Joint modelling of longitudinal and survival data has received significant attention in the literature in the past two decades (e.g., Rizopoulos (2012); Elashoff, Li and Li (2015); Wu and Yu (2016)). The following issues complicate joint modelling of HIV vaccine data: (i) some longitudinal biomarker data may be left-censored due to a lower limit of quantification (LLOQ) or below detection limit; (ii) there may be outliers in the longitudinal data; (iii) various immune responses biomarkers may be strongly associated with each other and may be of mixed types such as some being binary and some being continuous; and (iv) standard likelihood estimation of the model parameters can be computationally intensive. There seem to be no existing joint modelling methods which can address all these issues simultaneously.

Our research is motivated by the VAX004 phase III trial. It is a 36-month efficacy study of a candidate vaccine, which contains two recombinant gp120 Envelope proteins – the MN and GNE8 HIV-1 strains, to prevent HIV-1 infection (Flynn et al., 2005). Participants assigned to the vaccine group received the candidate vaccine, with repeated immunizations at months 0, 1, 6, 12, 18, 24, 30 post-enrollment. They were followed until the day of diagnosis of HIV infection or until the end of the study. Meanwhile, their plasma samples were collected at the immunization visits, two weeks after each immunization visit, and the final visit. A number of immune response biomarkers were measured at each visit, including NAb (the titer of neutralizing antibodies to the MN strain of the HIV-1 gp120 Envprotein) and MNGNE8 (the average level of binding antibodies, measured by ELISA, to the MN and GNE8 HIV-1 gp120 Env proteins). Figure 1 shows the trajectories of four immune response biomarkers (Yu, Wu and Gilbert (2018)). Besides the features noted in the previous paragraph, we also see the following features from Figure 1: (a) different biomarkers seem strongly associated with each other since they exhibit similar patterns over time; (b) there may be associations between the

Keywords and phrases: biomarker, outliers, robust joint model, h-likelihood, left-censoring
biomarkers’ longitudinal patterns and the risk of HIV infection; and (c) there seem large variations between individuals. More details are presented in Section 4. Due to the large between-individual variations, mixed effects or random effects models may be desirable for modelling the longitudinal biomarker data.

Fig 1: Longitudinal measurements of four immune response biomarkers. The solid lines show randomly selected subjects with HIV infection by the end of the study, while the dotted lines show randomly selected subjects without HIV infection by the end of the study.

Many authors have reviewed the literature on joint models of longitudinal and survival data (e.g., Rizopoulos (2012); Taylor et al. (2013); Elashoff, Li and Li (2015); Wu and Yu (2016), among others). There has also been active research on longitudinal data with left censoring (e.g., Hughes (1999); Wu (2009); Yu, Wu and Gilbert (2018)), longitudinal data with outliers (e.g., Cantoni and Ronchetti (2001); Sinha (2004); Qin et al. (2016)), and computational tools for joint models (e.g., Rizopoulos (2012); Lee and Nelder (1996); Barrett et al. (2015)). However, to our knowledge, little literature seems available to address all these issues simultaneously.

In this paper, we propose a joint model for longitudinal and survival data, which addresses left censoring and outliers simultaneously. While the methodology for each of these issues is not new, addressing all issues simultaneously in the joint model is not trivial. For (unobserved) left censored longitudinal data, we avoid distributional assumptions by treating them as a point mass, following Yu, Wu and Gilbert (2018). An outlier may be defined as a data point which appears to be inconsistent with the remainder of the data, or a data point which differs significantly from other data. For longitudinal data, the terms e-outliers and b-outliers have appeared in the literature: an e-outlier is an outlying data point among the repeated
measurements within an individual, while a b-outlier is an outlying individual in the sample. In other words, an e-outlier may arise in the model for the within-individual repeated measurements, while a b-outlier may arise in the assumed model/distribution for the random effects. Thus, naturally, we may consider a heavy-tail multivariate \( t \)-distribution for the b-outliers rather than the standard normal distribution assumed for the random effects, and we may use an M-estimator for the e-outliers. An M-estimator downweights potential outlying observations and is widely used in the robust inference literature (e.g., Cantoni and Ronchetti (2001); Sinha (2004); Wu and Qiu (2011)). Note that, if there are too many outliers (say, over 5%), these outliers may be viewed as from a separate population, so we might either study them separately or consider a mixture of two distributions such as a mixture of two normal distributions.

To address the association between different types of longitudinal data, we consider several mixed effects models with correlated or shared random effects. For such a complex joint model, computation for likelihood inference is a major challenge, since the standard EM algorithm may offer very slow or non-convergence. Here we extend the h-likelihood method of Lee and Nelder (1996) for the proposed joint model for approximate likelihood estimation, which is computationally very efficient.

The rest of the paper is organized as follows. In Section 2, we describe the joint model and its likelihood in a general form. Section 3 introduces our proposed models and methods, with asymptotic results, and the proposed approximate estimation method. In Section 4, we analyze the HIV vaccine data in details. Section 5 shows simulation results evaluating the proposed method. We conclude the article with some discussion in Section 6.

2. The Joint Model and Its Likelihood.

2.1. The models. In this section, we present the models in general forms so that they can be applied in other applications. In the motivating HIV vaccine study described in the previous section, the longitudinal biomarkers are highly correlated, as can be seen from the similar patterns in Figure 1. In practice, some continuous longitudinal biomarker data are sometimes dichotomized for biological reasons. Thus, to simplify the presentation, here we consider one continuous longitudinal biomarker \( Y^* \) and one binary longitudinal biomarker \( Z \) (taking values 0 or 1). These two different types of biomarkers may be associated with each other, as noted earlier. In addition, the biomarker \( Y^* \) value may be left censored due to an assay’s lower limit of quantification (LLOQ) \( d \) (or below detection limit \( d \)). We denote the observed (un-censored) \( Y^* \) value by \( Y \). Another complication in the data is that some observed \( Y \) data may contain outliers. Let \( C \) be the left censoring indicator of \( Y^* \) such that \( C = 1 \) if \( Y^* \) value is below LLOQ (i.e., \( Y^* < d \)) and \( C = 0 \) otherwise. Occasionally, to simplify presentation, we may not strictly distinguish \( Y \) and \( Y^* \) if the context is clear. We also use bold-face \( y \) to denote a column vector, \( y^T \) denote its transpose, and \( y \) denote a scalar. We use similar notation for other variables.

Note that the value of the left-censoring indicator variable \( C \) at time \( t \) is determined by the longitudinal biomarker \( Y^* \) value if \( Y^* \) is measured at time \( t \). However, for the motivating HIV study, the biomarker \( Y^* \) value is measured infrequently over time, so the values of the censoring status \( C \) at many time points are not available when \( Y^* \) is not measured, especially at time points close to HIV infection. On the other hand, the longitudinal patterns of the censoring status \( C \), especially at time points close the HIV infection, may be associated with the risk of HIV infection, as shown in the data analysis and other studies. Thus, by modelling the censoring indicator \( C \) based on observed data, we may be able to “predict” the left censoring status \( C \) at time \( t \) based on its observed patterns and other available information, even when \( Y^* \) is not measured at time \( t \). As noted earlier, some longitudinal biomarkers in the study are
highly correlated. So we may use this information to help predicting the left-censoring status \(C\) at time \(t\) when \(Y^*\) is not measured. We may incorporate this information by allowing the random effects in the \(C\) model to be correlated with the random effects in the models for other biomarkers (e.g., \(Z\)) or directly use other biomarker as predictors in the model for \(C\). Here we consider correlated or shared random effects.

Suppose that there are \(n\) individuals in the study, with \(n_i\) repeated measurements on the longitudinal biomarkers for individual \(i,\ i = 1, 2, \ldots, n\). Some biomarker values may be missing at scheduled times, and the missing data are assumed to be missing at random. Let \(Y_{ij}\) be the biomarker \(Y\) for individual \(i\) at time \(t_{ij}\) and \(y_{ij}\) be its observed value, \(i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n_i\). We define \(C_{ij}, Z_{ij}\) and \(c_{ij}, z_{ij}\) similarly. The observed data are \(\{(y_{ij}, c_{ij}, z_{ij}),\ i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n_i\}\).

We consider a linear mixed effects (LME) model for the continuous \(Y\)-data, and generalized linear mixed effects (GLME) models for the binary \(Z\)-data and \(C\)-data. Specifically, we consider the following GLME models for the binary variable \(Z_{ij}\)

\[
(1) \quad \text{logit}(P(Z_{ij} = 1)) = w_{ij}^T \alpha + u_{ij}^T b_{ii} \equiv \mu_{Z_{ij}}, \quad i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n_i,
\]

where \(w_{ij}\) and \(u_{ij}\) are covariate vectors containing time, \(\alpha\) is a vector of fixed effect parameters, \(b_{ii}\) is a vector of random effects. For simplicity, we suppress the dimensions of these vectors, with the understanding that the dimensions match in the model. We use similar notation for other models.

For the observed (un-censored) data \(Y\), i.e., given \(C_{ij} = 0\), we consider the following LME model

\[
(2) \quad Y_{ij} = x_{ij}^T \beta + v_{ij}^T b_{2i} + \epsilon_{ij} \equiv \mu_{Y_{ij}} + \epsilon_{ij}, \quad i = 1, 2, \ldots, n,\ j = 1, \ldots, n_i,
\]

where \(x_{ij}\) and \(v_{ij}\) contain covariates including time, \(\beta\) contains fixed effect parameters, \(b_{2i}\) contains random effects, and \(\epsilon_{ij}\) is an error term with \(\epsilon_{ij} \sim N(0, \sigma^2)\). We assume that the observed data \(Y\) follow a truncated normal distribution with density function \(f(y_{ij}|c_{ij} = 0, \beta, b_{2i}, \sigma) = \frac{1}{\sigma} \phi \left( \frac{y_{ij} - \mu_{Y_{ij}}}{\sigma} \right)\). Note that, here we only model the observed (uncensored) \(Y\)-data, not the unobserved left-censored \(Y^*\)-data which are below detection limit \(d\). Thus, we avoid making un-verifiable model or distributional assumptions for the unobserved left-censored data. The unobserved left-censored data are treated as point mass with \(f(y^*_{ij}|c_{ij} = 1, \beta, b_{2i}, \sigma) = 1\) as shown in the joint likelihood (5) in Section 2.2. This may lead to some loss of information, compared with a model assuming a (unverifiable) distribution for the censored values. However, we also avoid assuming a model for the unobserved left-censored \(Y^*\)-data which cannot be verified and is often unreasonable since the unobserved below-detection \(Y^*\)-values may behave very differently than the observed above-detection \(Y\)-values.

In model (2), we actually model \(Y|C = 0\), so we also need to model \(C\), similar to the idea of mixture models. As noted earlier, for the left-censoring indicator \(C\), while its value is determined by the value of \(Y^*\) if \(Y^*\) is measured, the values of \(Y^*\) are often un-available at times of interest due to its infrequent measurement schedules. Since the values of \(Y\) are highly correlated with the values of other biomarkers and exhibit similar periodic patterns (see Figure 1), we may use information from other observed biomarkers and the observed longitudinal pattern of left-censoring statuses to possibly “predict” the censoring status of \(Y^*\) when the values of \(Y^*\) are unavailable. The information from other observed biomarkers may be incorporated by allowing the random effects from these biomarker models to be correlated with the random effects for the \(C\)-model, since these random effects summarize the individual-specific characteristics of the corresponding longitudinal trajectories. Thus, we consider the following GLME model for the binary censoring indicator \(C_{ij}\)

\[
(3) \quad \text{logit}(P(C_{ij} = 1)) = \phi_{ij}^T \eta + \psi_{ij}^T b_{3i} \equiv \mu_{C_{ij}}, \quad i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n_i,
\]
where $\phi_{ij}$ and $\psi_{ij}$ contain covariates including time, $\eta$ contains fixed effect parameters, and $b_{3i}$ contains random effects. Note that, as shown below, the random effects $b_{3i}$ in model (3), which reflect the individual characteristics of the censoring process, are assumed to be correlated with the random effects $b_{1i}$ and $b_{2i}$ in models (1) and (2) respectively, allowing information from other biomarker data to help predicting the censoring status when it is not known. Such censoring information may be predictive for the risk of HIV infection.

The values of $Y$, $Z$, and $C$ are correlated. Since $Y_{ij}$ is a continuous random variables while $Z_{ij}$ and $C_{ij}$ are binary random variables, the usual multivariate mixed effects models are not applicable. Here, we incorporate the association by assuming that the random effects in the three models are correlated or shared. This is a reasonable assumption since the three variables exhibit similar individual-specific longitudinal patterns. Specifically, we assume that

$$b_i = (b_{1i}, b_{2i}, b_{3i})^T \sim f(b_i | \Sigma), \quad i = 1, 2, \ldots, n,$$

where $\Sigma$ is an arbitrary covariance matrix incorporating the correlations among the random effects, and $f(b_i | \Sigma)$ is typically a multivariate normal distribution with zero mean and covariance matrix $\Sigma$ but it can also be a multivariate $t$-distribution for robust inference. Note that $b_i$ only contains the distinct elements in $(b_{1i}, b_{2i}, b_{3i})$.

In the HIV vaccine study, our main interest is in the risk of event “HIV infection” and if this risk is associated with the longitudinal biomarker data. At the end of the study, some subjects may not experience HIV infection, so their event times are right censored. We assume that the right censoring is non-informative. Let $\{(s_i, \delta_i) | i = 1, \ldots, n\}$ denote the observed survival (or event time) data, where $s_i$ is either the event time or the censoring time, and $\delta_i$ is the event indicator such that $\delta_i = 1$ if subject $i$ is infected and $\delta_i = 0$ otherwise. For the survival data, we consider the following Cox proportional hazards model

$$h_i(t) = h_0(t) \exp(x_{0i}^T \lambda_0 + b_i^T \lambda_1) \equiv h_0(t) \exp(\mu_{si}), \quad t \geq 0, \quad i = 1, 2, \ldots, n,$$

where $h_i(t)$ is the hazard/risk of HIV infection for subject $i$ at time $t$, $h_0(t)$ is an unspecified baseline hazard, $x_{0i}$ contains baseline covariates, and $\lambda_0$ and $\lambda_1$ are vectors of fixed parameters. Thus, the risk of HIV infection $h_i(t)$ is assumed to be associated with the random effects $b_i$ in the longitudinal biomarker models. This assumption may be reasonable since the random effects in the longitudinal models reflect the individual-specific characteristics of the corresponding longitudinal processes. Such a joint model is called shared-parameter model in the joint model literature, which is perhaps the most commonly assumed joint model for longitudinal and survival data.

In the above models, the three longitudinal models (1), (2), (3) are linked through correlated or shared random effects, and these random effects are used as “predictors” in the survival model (4). The specifications of these models are motivated by the HIV vaccine data described in Section 1. Since the four models are associated, inference of all the model parameters can be based on the joint likelihood of all the observed data, as described in the next section.

2.2. The likelihood. Let $\lambda = (\lambda_0^T, \lambda_1^T)^T$ be a collection of parameters in the survival model (4), let $\theta = (\alpha^T, \beta^T, \eta^T, \lambda^T)^T$ be a collection of all the mean parameters in all the longitudinal and survival models, let $\text{vec}(\Sigma)$ be a vector of unconstrained parameters that determines $\Sigma$ based on spherical parameterization, and let $\xi = (\sigma, \text{vec}(\Sigma))$ be a collection of all the distinct variance-covariance parameters. Let $f(\cdot)$ denote a generic density function. We assume that, conditioning on the random effects, the longitudinal variables and the survival
variable are independent of each other. The (joint) likelihood of all the observed longitudinal data and survival data based on the assumed models can be written as

\[
L(\theta, \xi | y, c, z, s, \delta) = \prod_{i=1}^{n} \int f(z_i | \alpha, b_1) f(y_i | c_i, \beta, b_2, \sigma) f(c_i | \eta, b_3) \times f(s_i, \delta_i | \lambda, b) f(b_i | \Sigma) dB_i,
\]

where \( f(z_i | \alpha, b_1) = \prod_{j=1}^{n_i} \left( \exp(\mu_{zi}) \right)^{z_{ij}} \left[ 1 + \exp(\mu_{zi}) \right]^{-1}, \) \( f(y_i | c_i, \beta, b_2, \sigma) = \prod_{j=1}^{n_i} \left[ \phi\left( \frac{y_{ij} - \mu_{zi}}{\sigma} \right) \right]^{c_{ij}}, \) \( f(c_i | \eta, b_3) = \prod_{j=1}^{n_i} \left[ \exp(\mu_{ci}) \right]^{c_{ij}} \left[ 1 + \exp(\mu_{ci}) \right]^{-1}, \) \( f(s_i, \delta_i | \lambda, b) = \left[ h_i(s_i) \right]^{k_i} \exp(-\int_0^{s_i} h_i(u) du), \) and \( \phi() \) is the probability density function (pdf) of a standard normal distribution.

The MLEs of all the parameters can be obtained by maximizing the above likelihood, which leads to solving the following estimating equations

\[
S(\alpha) = E\left\{ \frac{\partial \log f(z | \alpha, b_1)}{\partial \alpha} \right\} | y, c, z, s, \delta = 0,
\]

\[
S(\beta) = E\left\{ \frac{\partial \log f(y | c, \beta, b_2, \sigma)}{\partial \beta} \right\} | y, c, z, s, \delta = 0,
\]

\[
S(\eta) = E\left\{ \frac{\partial \log f(c | \eta, b_3)}{\partial \eta} \right\} | y, c, z, s, \delta = 0,
\]

\[
S(\lambda) = E\left\{ \frac{\partial \log f(s, \delta | \lambda, b)}{\partial \lambda} \right\} | y, c, z, s, \delta = 0,
\]

\[
S(\sigma) = E\left\{ \frac{\partial \log f(y | c, \beta, b_2, \sigma)}{\partial \sigma} \right\} | y, c, z, s, \delta = 0,
\]

\[
S(vec(\Sigma)) = E\left\{ \frac{\partial \log f(b | \Sigma)}{\partial vec(\Sigma)} \right\} | y, c, z, s, \delta = 0.
\]

Before discussing computational issues of solving the above estimating equations, we note that the MLEs of the parameters in the mixed effects models are very sensitive to outliers in the longitudinal data (Sinha (2004), Wu (2009)). In other words, ignoring outliers in the data may lead to misleading parameter estimates. In the next section, we discuss robust estimates of the model parameters.

### 3. Robust Estimation.

#### 3.1. A robust method. For longitudinal data, there are two types of outliers: e-outliers (outlying observations within an individual) and b-outliers (outlying individuals). Since it can often be challenging to detect outliers in longitudinal data, robust methods which automatically incorporate outliers can be very useful. In the following, we focus on outliers in the continuous data of \( Y \). The method can be conceptually extended to outliers in the binary data of \( Z \), as shown in Sinha (2004). For potential outliers in the \( Y \)-data, we propose to handle e-outliers based on the M-estimators and handle b-outliers based on heavy-tail distributions such as the multivariate t-distribution. Since the longitudinal data of \( Y \) and \( Z \) exhibit similar between-individual variations, b-outliers in the \( Y \)-data may also imply b-outliers in the \( Z \)-data.
An M-estimator for e-outliers. For e-outliers in the $Y$-data, we consider an M-estimator method to downweigh the influence of outliers in the estimating equation (7) for estimating $\beta$, as in Sinha (2004). Note that the estimating equation (7) of $\beta$ can be re-written as

$$S(\beta) = E\left(\sum_{i=1}^{n} \sum_{j=1}^{n_i} \left\{ (1 - c_{ij}) \frac{y_{ij} - \mu_{ij}}{\sigma^2} x_{ij} \right\} \mid y, c, z, s, \delta \right) = 0.$$  

We apply a bounded function $\psi()$ (e.g. the Huber’s function $\psi_c(x) = \max\{-c, \min(x, c)\}$, $c > 0$ with turning point $c$) on the term $\frac{y_{ij} - \mu_{ij}}{\sigma}$ and obtain the following robust estimating equation for $\beta$

$$\tilde{S}(\beta) = E\left(\sum_{i=1}^{n} \sum_{j=1}^{n_i} \left\{ (1 - c_{ij}) \frac{1}{\sigma} \psi\left(\frac{y_{ij} - \mu_{ij}}{\sigma}\right) x_{ij} \right\} - Q(\beta, b_2) \mid y, c, z, s, \delta \right) = 0,$$

where $Q(\beta, b_2) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} E\left\{ (1 - c_{ij}) \frac{1}{\sigma} \psi\left(\frac{y_{ij} - \mu_{ij}}{\sigma}\right) x_{ij} \mid b_{2i} \right\}$ is a correction term to ensure Fisher’s consistency of an estimator of $\beta$. When $\psi(x) = \psi_c(x)$ (the Huber’s function), we may choose the value of $c$ to reflect different degrees of robustness. When $c = \infty$, the robust M-estimator reduces to the usual MLE. Common choices of $c$ are $c = 1.5$ or $c = 2$.

Similarly, for the variance parameter $\sigma^2$ in the model for the $Y$-data, we may consider the following robust estimating equation

$$\tilde{S}(\sigma) = E\left(\sum_{i=1}^{n} \sum_{j=1}^{n_i} (1 - c_{ij}) \psi^2(r_{ij}) - \sum_{i=1}^{n} \sum_{j=1}^{n_i} E\left\{ (1 - c_{ij}) \psi^2(r_{ij}) \mid b_{2i} \right\} \mid y, c, z, s, \delta \right) = 0,$$

where $r_{ij} = (y_{ij} - \mu_{ij})/\sigma$.

A multivariate t-distribution for b-outliers. The standard distributional assumption for the random effects is the multivariate normal distribution. When the variations between individuals suggest potential b-outliers, a multivariate t-distribution may be more desirable since it has heavier tails than the multivariate normal distribution. Thus, for robust inference, we assume that the random effects follow a multivariate t-distribution with degrees of freedom $\nu$ and covariance matrix $\Sigma$, i.e.,

$$b_i \sim t_{\nu} \left(0, \frac{\nu - 2}{\nu} \Sigma\right), \quad \nu > 2, \quad i = 1, 2, \cdots, n.$$  

We may then choose the value of the degrees of freedom $\nu$ to reflect different degrees of robustness: the smaller the value of $\nu$, the heavier the tails of the t-distribution. When $\nu = \infty$, the distribution $t_{\nu} \left(0, \frac{\nu - 2}{\nu} \Sigma\right)$ reduces to the multivariate normal distribution $N(0, \Sigma)$. With sufficient data, we may estimate the degrees of freedom $\nu$ from the data using the maximum likelihood method. For small samples, we may set it a priori at some sensible values (often between 3 and 9) and perform sensitivity analysis (Lange, Little and Taylor, 1989). Based on Lucas (1997), the protection against outliers is preserved only if the degree of freedom parameter is fixed. In the data analysis, we assume $\nu$ to be fixed and plot the profile likelihood of $\nu$ to choose the most appropriate value of $\nu$ and we also perform sensitivity analysis.

3.2. Asymptotic results. We show some asymptotic results of the robust estimate of $\beta$, following similar proofs in the literature (e.g. Sinha (2004)). These asymptotic results can be used to obtain standard errors of the robust estimates in data analysis, when the sample size is reasonably large. When the sample size is small, bootstrap method may be used to obtain the standard errors of robust parameter estimates. For other parameter estimates, standard asymptotic results for MLEs may be used.
Let \( D_i = (y_i^T, c_i^T, z_i^T, s_i, \delta_i)^T \) be a collection of all the longitudinal and survival response data of individual \( i, i = 1, 2, \cdots, n \). Let

\[
g'(\beta, D_i) = E\left\{ \sum_{j=1}^{n_i} (1 - c_{ij}) \frac{1}{\sigma} \psi\left( \frac{y_{ij} - \mu y_{ij}}{\sigma} \right) x_{ij} - \sum_{j=1}^{n_i} E((1 - c_{ij}) \frac{1}{\sigma} \psi\left( \frac{y_{ij} - \mu y_{ij}}{\sigma} \right) x_{ij} | b_{2i}) \bigg| D_i \right\},
\]

and \( \tilde{S}(\beta) = \sum_{i=1}^{n} g'(\beta, D_i) \). Let \( g(\beta, D_i) \) be a function whose derivative function with respect to \( \beta \) is \( g'(\beta, D_i) \) and \( G_n(\beta, D) = \frac{1}{n} \sum_{i=1}^{n} g(\beta, D_i) \). Then we have \( G_n'(\beta, D) = \frac{1}{n} \sum_{i=1}^{n} g'(\beta, D_i) = \frac{1}{n} \tilde{S}(\beta) \). Let \( G_n(\beta) = E(G_n(\beta, D)) = \frac{1}{n} \sum_{i=1}^{n} E(g(\beta, D_i)) = E(g(\beta, D)) \).

**Theorem.** Let \( \hat{\beta}_n \) be the robust estimator obtained by solving the estimating equation (12), and let \( \beta^* \) be the true value of \( \beta \). Under the regularity assumptions R1–R7 given in the web-based Supplemental Materials, we have the following results:

(a) Consistency: \( \hat{\beta}_n \xrightarrow{p} \beta^* \) a.s. as \( n \to \infty \),

(b) Asymptotic normality:

\[
\sqrt{n}(\hat{\beta}_n - \beta^*) \xrightarrow{d} N(0, \Lambda^{-1} \Omega \Lambda^{-1}), \quad \text{as } n \to \infty,
\]

where

\[
\Lambda = \lim_{n \to \infty} \Lambda_n = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} E\left( \left. \frac{\partial g'(\beta, D_i)}{\partial \beta^T} \right|_{\beta=\beta^*} \right),
\]

and

\[
\Omega = \lim_{n \to \infty} \Omega_n = \lim_{n \to \infty} \frac{1}{n} E \left[ \left( \sum_{i=1}^{n} g'(\beta^*, D_i) \right) \left( \sum_{i=1}^{n} g'(\beta^*, D_i) \right)^T \right].
\]

The proofs are given in the web-based Supplemental Materials.

The above asymptotic normality result (b) can be used for inference of parameter \( \beta \), such as confidence interval and hypothesis testing. Let \( \hat{\alpha}_n, \hat{\eta}_n, \hat{\lambda}_n \) be the parameter estimates obtained by solving likelihood score equations (6), (8), and (9), respectively. Then these estimates have the usual properties of consistency and asymptotic normality of MLEs.

3.3. A computationally efficient approximate method. The estimating equations in the previous sections involve intractable integrations with respect to the (unobserved) random effects \( b_i = (b_{1i}, b_{2i}, b_{3i}) \). The dimension of the random effects \( b_i \) is usually high (at least 3), so it can be computationally challenging to solve these estimating equations since the commonly used numerical integration methods or Monte Carlo methods can be computationally intensive. In the following, we propose an approximate method based on h-likelihood method, which is computationally efficient (Lee and Nelder (1996)). As an attractive computational tool, the h-likelihood method has becoming increasingly popular for likelihood estimation of mixed effects models – see Lee, Nelder and Pawitan (2018) for a recent review. Essentially, it uses Laplace approximations to intractable integrals in the likelihoods for mixed effects models, with adjustments to reduce bias, which produce approximate maximum likelihood estimates (MLEs) for the mean parameters and approximate restricted MLEs (REMLs) for the variance-covariance (dispersion) parameters. It is particularly useful for likelihood estimation involving high-dimensional integrations with respect to the random effects, since it is computationally much more efficient than the corresponding Monte Carlo or numerical methods and it performs well in various settings. In this paper, we propose an h-likelihood method for robust inference in our joint model setting and evaluate its performance via simulations.
The log h-likelihood function based on the assumed models can be written as
\[
\ell_h(\theta, \xi|y, c, z, s, \delta) = \log f(z|\alpha, b_1) + \log f(y|c, \beta, b_2, \sigma) + \log f(c|\eta, b_3)
\]
\[
+ \log f(s, \delta|\lambda, b) + \log f(b|\Sigma),
\]
where \( z = (z_1^T, z_2^T, \cdots, z_n^T)^T \) and similar notation for other variables. Let \( \theta_{-}\beta \) be the collection of all the mean parameters except for \( \beta \). Beginning with some starting values \((\theta^{(0)}, \xi^{(0)})\), we propose the following iterative procedure: at iteration \( k (k = 1, 2, \cdots) \), we do the following steps

- **Step 1:** Given \( \theta^{(k)} \) and \( \xi^{(k)} \), obtain updated estimates of the random effects \( b^{(k+1)} \) by maximizing \( \ell_h \) with respect to \( b \);
- **Step 2:** Given \( \xi^{(k)} \), \( b^{(k+1)} \), and \( \hat{\theta}_{-}\beta^{(k)} \), obtain updated robust estimate of \( \beta^{(k+1)} \) by solving estimating equation (12), where the the expectation is approximated using samples drawn based on the Metropolis-Hastings algorithm.
- **Step 3:** Given \( \hat{\xi}^{(k)} \), \( \hat{b}^{(k+1)} \), and \( \hat{\beta}_{-}\beta^{(k+1)} \), obtain updated estimates of \( \theta_{-}\beta \) by maximizing the following adjusted profile h-likelihood with respect to \( \theta_{-}\beta \):
  \[
p_{\theta_{-}\beta} = \left( \ell_h - 0.5 \log \frac{H(\ell_h, b)}{2\pi} \right)_{\xi^{(k)}, b=\hat{b}^{(k+1)}, \beta=\hat{\beta}^{(k+1)}},
\]
where \( H(\ell_h, b) = -\frac{\partial^2 \ell_h}{\partial b^T \partial b} \).
- **Step 4:** Given \( \hat{b}^{(k+1)} \) and \( \hat{\beta}_{-}\beta^{(k+1)} \), obtain updated estimates of \( \sigma^{(k+1)} \) by solving the estimating equation (13), where the expectation is approximated using samples drawn based on the Metropolis-Hastings algorithm. Then, we obtain the estimates of the rest of the parameters in \( \xi \) (denoted as \( \xi_{-}\sigma \)) by maximizing the following adjusted profile h-likelihood with respect to \( \xi_{-}\sigma \),
  \[
p_{\xi_{-}\sigma} = \left( \ell_h - 0.5 \log \frac{H[\ell_h(\theta, b)]}{2\pi} \right)_{\theta=\hat{\theta}^{(k+1)}, b=\hat{b}^{(k+1)}, \sigma=\hat{\sigma}^{(k+1)}},
\]
where \( H[\ell_h(\theta, b)] = -\left( \frac{\partial^2 \ell_h}{\partial \theta^T \partial \theta} + \frac{\partial^2 \ell_h}{\partial \beta^T \partial \beta} \right) \).
- **Step 5:** Given \( \hat{b}^{(k+1)}, \hat{\theta}_{-}\beta^{(k+1)}, \hat{\xi}^{(k+1)} \), we obtain an updated nonparametric estimate of the baseline hazard as follows
  \[
h_0^{(k+1)}(t) = \sum_{i=1}^n \frac{\sum_{i=1}^n \exp \left( x_{i0}^T \lambda_0^{(k+1)} + (b_i^{(k+1)})^T \lambda_i^{(k+1)} \right) I(s_i \geq t)}{\sum_{i=1}^n \exp \left( x_{i0}^T \lambda_0^{(k+1)} + (b_i^{(k+1)})^T \lambda_i^{(k+1)} \right)},
\]
where \( I(\cdot) \) is an indicator function. If the survival data are modelled by a parametric model such as a Weibull model, the parameter estimates can be obtained in a similar way as in the previous steps.

Iterating the above Steps 1–5 until convergence, we obtain the approximate estimates of the parameters and random effects. After convergence, we may estimate the covariance matrices of the parameter estimates by calculating the derivatives of the estimating equations (6), (12), (8), and (9),
\[
\tilde{\text{cov}}(\alpha) = [-\partial S(\alpha)/\partial \alpha^T]^{-1}|_{\alpha=\cdot}, \quad \tilde{\text{cov}}(\hat{\beta}) = [-\partial \hat{S}(\beta)/\partial \beta^T]^{-1}|_{\beta=\hat{\beta}},
\]
\[
\tilde{\text{cov}}(\eta) = [-\partial S(\eta)/\partial \eta^T]^{-1}|_{\eta=\cdot}, \quad \tilde{\text{cov}}(\lambda) = [-\partial S(\lambda)/\partial \lambda^T]^{-1}|_{\lambda=\cdot}.
\]
The estimated standard errors of the parameter estimates are the square root of the diagonal elements of the covariance matrices.

4.1. The Data. In this section, we return to the VAX004 phase III trial dataset briefly described in Section 1. The dataset contains 194 participants in total. By the end of the study, 21 (10.8%) subjects acquired HIV infection during the follow-up, and the times to infection diagnosis range from 43 to 954 days. A number of longitudinal variables are collected throughout the study, but most of them are highly correlated with correlations over 0.9. Thus, we choose two representative variables of primary interest: NAb and MNGNE8. The variable MNGNE8 is dichotomized based on its median values, as done in Yu, Wu and Gilbert (2018).

From Figure 1, we see that data on the longitudinal variables exhibit similar periodic patterns, due to repeated administrations of the vaccine. Moreover, variable NAb is left-censored with a lower limit of quantification (LLOQ) of 1.477 and censoring rate of 27%. The observed NAb data range from 1.79 to 4.82 with a mean of 3.13 and a standard deviation of 0.59. Figure 2 shows the longitudinal trajectories of NAb of some selected participants, with left-censoring and possible outliers. The two points indicated in blue boxes in Figure 2 are examples of possible outliers, because they do not appear to follow the general patterns of sawtooth waves. In HIV vaccine trials, the values of an immune response typically increase sharply right after each vaccination, and then start to decrease about two weeks after the vaccination. So the point outlier 1, measured about 23 weeks after vaccination, is supposed to be below the previous measurement point, and the point outlier 2, measured about 2 weeks after vaccination, is supposed to be above the previous measurement point. Thus, these two points are possible e-outliers in the data. As discussed in Section 1, outliers are often model-dependent and may not always be easily detected. Thus, robust methods such as the M-estimators and t-distributions are useful since they automatically incorporate potential outliers. For longitudinal data, e-outliers refer to outlying observations among within-individual repeated measurements, while b-outliers refer to outlying individuals in the sample. As exploratory analysis, some outliers may be detected by plotting standardized residuals and estimated random effects (Zheng, Fung and Zhu, 2013; Koller, 2016), or by inspecting the distributions of the Mahalanobis distances based on the estimates (Gill, 2000; Waternaux, Laird and Ware, 1989; Copt and Victoria-Feser, 2006).

A main objective of the study is to investigate possible association between the risk of HIV infection and the immune response longitudinal measurements. From Figure 1, it appears that such association may exist, but it needs to be confirmed by modelling the data. Thus, a joint model for the longitudinal data and the survival (time-to-infection) data is desirable. A main challenge in the data analysis is to handle outliers and left censoring in the longitudinal data simultaneously. In the following, we analyze the dataset based on the proposed joint model and the estimation method.

4.2. The models. 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 its sample median 0.57 and \( Z_{ij} = 0 \) otherwise (Yu, Wu and Gilbert (2018)). Let \( Y_{ij} \) be the observed (uncensored) NAb value (i.e. \( Y_{ij} > \text{LLOQ} \)), and let \( C_{ij} \) be the left-censoring indicator of NAb such that \( C_{ij} = 1 \) if NAb is left-censored and \( C_{ij} = 0 \) otherwise.

As we can see from Figure 1, the longitudinal trajectories of \( Y \) and \( Z \) are closely associated with time variables, which may be defined in different ways, such as scheduled measurement times, vaccination times, times since most recent vaccinations, and time intervals between two consecutive vaccinations. Thus, in modelling these longitudinal data, we may treat these time variables as covariates. More specifically, we define the following time variables (in months): (i) \( \text{sampledays} \), denoted by \( t_{ij} \), which is the measurement time of individual \( i \) at \( j \)-th measurement; (ii) the vaccination time of individual \( i \) at \( k \)-th vaccination, denoted by \( t_{ik}^{V} \); (iii)
**dosedays**, denoted by $t_{d_{ij}}$, which is the time from the most recent vaccination to the current measurement time $t_{ij}$, i.e., $t_{d_{ij}} = t_{ij} - t^V_{ik}$ where $t^V_{ik} \leq t_{ij} \leq t^V_{ik+1}$; and (iv) the time between two consecutive vaccinations within which $t_{ij}$ lies, denoted by $\Delta_{ij}$, i.e., $\Delta_{ij} = t^V_{i,k+1} - t^V_{ik}$ with $t^V_{ik} \leq t_{ij} \leq t^V_{i,k+1}$. For measurement time $t_{ij}$ after the final vaccination, we define $\Delta_{ij}$ as the time between the final vaccination and the final measurement time. In data analysis, to avoid too small or too large parameter estimates, we re-scale some time variables to different time units (year/week): $t^*_{ij} = t_{ij}/12$ (in years) and $t^*_{d_{ij}} = t_{d_{ij}} \times 30/7$ (in weeks).

Next, we consider several mixed effects models for modelling the longitudinal biomarker data, due to the large between-individual variations. Since the longitudinal measurement schedules are infrequent in this study, nonparametric or semiparametric mixed effects models may not be suitable as they require more frequent measurements. As shown in Figure 1, the longitudinal biomarker data exhibit clear periodic patterns due to repeated vaccinations. Thus, we may use simple periodic functions, such as the $\sin()$ function, to partially capture these periodic patterns. For the remaining variations in the data, we consider linear and quadratic functions of time and choose the models using standard model selection criteria such as AIC or BIC values and likelihood ratio test. For models which are similar based on these criteria, we choose the ones which are simpler and make more scientific sense. As noted in the previous paragraph, to simplify the presentations of the models, we use different time variables $t_{ij}, t^*_{ij}, t_{d_{ij}}, t^*_{d_{ij}},$ and $\Delta_{ij}$ – they are illustrated in Figure 1 in the web-based Supplemental Materials. The details of the model selections and the goodness-of-fit of the models are also given in the web-based Supplemental Materials. As a result, we obtain the following final models for the longitudinal biomarker data:

\begin{align}
\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_1 b_{1i} + d_2 b_{2i} t_{ij}, \\
Y_{ij} &= \beta_0 + \beta_1 t^*_{ij} + \beta_2 t^2_{ij} + \beta_3 \sin\left(\frac{\pi}{\Delta_{ij}} \times t_{d_{ij}}\right) + d_3 b_{1i} + \epsilon_{ij}, \quad \text{for } Y_{ij} \geq \text{LLOQ}, \\
\text{logit}(P(C_{ij} = 1)) &= \eta_0 + \eta_1 t^*_{ij} + \eta_2 t^2_{ij} + \eta_3 \sin\left(\frac{\pi}{\Delta_{ij}} \times t_{d_{ij}}\right) + d_4 b_{1i},
\end{align}
where the $d_k$’s are dispersion/variance parameters, with identification conditions $d_1 > 0, d_2 > 0$, and $\epsilon_{ij}$ is an error term with $\epsilon_{ij} \sim N(0, \sigma^2)$. In the above models, the random effect $b_{1i}$ in the $Z$-model also appears in the models for $Y$ and $C$ since the three longitudinal trajectories exhibit very similar between-individual variations and thus have closely related random effects. We assume that the random effects

$$b_i = (b_{1i}, b_{2i})^T \sim t_\nu \left( 0, \frac{\nu - 2}{\nu} \Sigma \right)$$

with degree of freedom $\nu$ and covariance matrix $\Sigma = \left( \begin{array}{cc} 1 & r_{12} \\ r_{12} & 1 \end{array} \right)$, $|r_{12}| < 1$.

The random effects $b_i$ in the above longitudinal models represent individual-specific characteristics of the longitudinal trajectories. We are interested in investigating if such individual-specific characteristics may be associated with the risk of HIV infection. Therefore, for the survival data, we consider the following Cox model

$$h_i(t|x_i, b_i) = h_0(t) \exp \{ \lambda_0 x_i + \lambda_1 b_{1i} + \lambda_2 b_{2i} \},$$

where $x_i$ is the standardized value of the immune response GNE8_CD4 of individual $i$ at the initial visit, and the random effects $b_i$ are used as predictors in the survival model. The significances of the estimates of $\lambda_1$ and $\lambda_2$ indicate possible association between the longitudinal processes and risks of HIV infections.

4.3. Data analysis results. The longitudinal and survival models (15)–(18) are linked through shared random effects. We use the proposed joint model and the robust estimation method for estimating the model parameters simultaneously, incorporating outliers and left-censored values. We choose the values of the turning point $c$ in the M-estimator and the degree of freedom parameter $\nu$ in the $t$-distribution based on the plot of the profile likelihood. The profile log-likelihood plot of $\nu$ given different values of $c$ is shown in the web-based Supplemental Materials. The turning points $c = 1.5$ and $c = 2$ return very similar approximate log-likelihood values, with the latter slightly better than the former. Given a fixed $c$, the approximate log-likelihood generally has an increasing pattern over the degree of freedom parameter $\nu$. This implies that there may be few or no $b$-outliers or the impact of $b$-outliers is minor. The combination of $c = 2$ and $\nu = \infty$ (i.e. $b_i \sim N(0, \Sigma)$) gives the largest approximate log-likelihood among all.

Table 1 shows the parameter estimation results based on the non-robust (NR) method and the robust method with $c = 2$ and $\nu = \infty$ (R1), $c = \infty$ and $\nu = 3$ (R2), and $c = 2$ and $\nu = 3$ (R3), respectively. One of the main objectives of this analysis is to check if the parameter estimates and their standard errors are affected by the potential outliers in the longitudinal data. We see that the two methods produce somewhat similar estimates and standard errors, with the differences between them relatively minor. This suggests that the impact of potential outliers on the parameter estimates may be small in this setting. Thus, we can be reasonably confident about these results and the implied conclusions. Note that the key parameters of interest here are the $\lambda_k$’s since they link the risk of HIV infection to the random effects from the longitudinal models which reflect the individual-specific characteristics of the longitudinal processes. The negative estimates of $\lambda_1$ suggest that the larger the values of NAb ($Y$) and/or the larger the probability of MNGNE8 ($Z$) above its median, the lower the risk of HIV infection. The positive estimates of $\lambda_2$ suggest that a larger increasing trend in MNGNE8 over time is associated with a higher risk of HIV infection.

We also conducted model diagnostics and present the results in the web-based Supplemental Materials. For the linear mixed effect model of NAb, we check its goodness-of-fit by the fitted-versus-observed value plots. For the logistic mixed effect models of $C$ and MNGNE8ind, we plotted their ROC curves. For the survival model, we plotted the Kaplan-Meier estimates versus fitted survival curves and also the Schoenfeld residual plot. All these
plots can be found in the web-based Supplemental Materials. Overall, the proposed models fit the data reasonably well without major concerns.

### Table 1

**Parameter estimates based on robust and non-robust methods.**

<table>
<thead>
<tr>
<th>Model</th>
<th>Par.</th>
<th>Estimate</th>
<th>Std.Error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NR  R1   R2  R3</td>
<td>NR  R1   R2  R3</td>
<td>NR  R1   R2  R3</td>
</tr>
<tr>
<td><strong>Estimates of mean parameters</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model (15)</td>
<td>(\alpha_0)</td>
<td>-1.55</td>
<td>-1.62</td>
<td>-1.59</td>
</tr>
<tr>
<td></td>
<td>(\alpha_1)</td>
<td>0.11</td>
<td>0.12</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>(\alpha_2)</td>
<td>1.73</td>
<td>1.81</td>
<td>1.79</td>
</tr>
<tr>
<td></td>
<td>(\alpha_3)</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>(\beta_0)</td>
<td>2.42</td>
<td>2.38</td>
<td>2.42</td>
</tr>
<tr>
<td>Model (16)</td>
<td>(\beta_1)</td>
<td>0.85</td>
<td>0.89</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>(\beta_2)</td>
<td>-0.28</td>
<td>-0.29</td>
<td>-0.28</td>
</tr>
<tr>
<td></td>
<td>(\beta_3)</td>
<td>1.34</td>
<td>1.47</td>
<td>1.34</td>
</tr>
<tr>
<td></td>
<td>(\beta_4)</td>
<td>0.20</td>
<td>0.21</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>(\gamma_0)</td>
<td>2.03</td>
<td>1.94</td>
<td>1.95</td>
</tr>
<tr>
<td>Model (17)</td>
<td>(\eta_1)</td>
<td>-6.56</td>
<td>-6.23</td>
<td>-6.25</td>
</tr>
<tr>
<td></td>
<td>(\eta_2)</td>
<td>1.73</td>
<td>1.64</td>
<td>1.65</td>
</tr>
<tr>
<td></td>
<td>(\eta_3)</td>
<td>-0.67</td>
<td>-0.64</td>
<td>-0.66</td>
</tr>
<tr>
<td></td>
<td>(\lambda_0)</td>
<td>-0.71</td>
<td>-0.70</td>
<td>-0.74</td>
</tr>
<tr>
<td>Model (18)</td>
<td>(\lambda_1)</td>
<td>-0.91</td>
<td>-1.07</td>
<td>-1.41</td>
</tr>
<tr>
<td></td>
<td>(\lambda_2)</td>
<td>1.64</td>
<td>1.60</td>
<td>1.92</td>
</tr>
</tbody>
</table>

**Estimates of dispersion parameters**

|         | \(\sigma\) | 0.50  | 0.49  | 0.50  | 0.49  |
|         | \(d_1\) | 0.35  | 0.29  | 0.23  | 0.23  |
|         | \(d_2\) | 0.05  | 0.06  | 0.07  | 0.07  |
|         | \(d_3\) | -0.88 | -0.79 | -0.75 | -0.75 |
|         | \(r_{12}\) | 0.52  | 0.50  | 0.61  | 0.61  |

NR: non-robust method with \(c = \infty\) and \((b_{11}, b_{22})^T \sim N(0, \Sigma)\).
R1: robust method with \(c = 2\) and \((b_{11}, b_{22})^T \sim N(0, \Sigma)\).
R2: robust method with \(c = \infty\) and \((b_{11}, b_{22})^T \sim t_3(0, \frac{1}{3} \Sigma)\).
R3: robust method with \(c = 2\) and \((b_{11}, b_{22})^T \sim t_3(0, \frac{1}{3} \Sigma)\).

### 5. Simulation Studies

In this section, we evaluate the proposed robust method and h-likelihood method via comprehensive simulation studies. The models and the true parameter values are chosen to be similar to those in the real data analysis (see Section 4.2). However, different percentages of two types of outliers are randomly created, and the proposed robust methods with various degrees of robustness are used to fit the data to evaluate and compare their performances. The details are given below.

#### 5.1. Simulation design

We simulate the data as follows. Given random effects \((b_{11}, b_{22})^T \sim N(0, \Sigma)\), we generate binary longitudinal data \(Z_{ij}\) and \(C_{ij}\) from Bernoulli distributions with probabilities

\[
\begin{align*}
    p_{Z_{ij}} &= \exp(\mu_{Z_{ij}})(1 + \exp(\mu_{Z_{ij}}))^{-1}, \\
    p_{C_{ij}} &= \exp(\mu_{C_{ij}})(1 + \exp(\mu_{C_{ij}}))^{-1},
\end{align*}
\]

respectively, where

\[
\begin{align*}
    \mu_{Z_{ij}} &= \alpha_0 + \alpha_1 t_{ij} + \alpha_2 \sin(\frac{\pi}{\Delta_{ij}}) t_{d_{ij}} + \alpha_3 t_{d_{ij}}^2 + d_1 b_{1i} + d_2 b_{2i} t_{ij}, \\
    \mu_{C_{ij}} &= \eta_0 + \eta_1 t_{ij} + \eta_2 t_{ij}^2 + \eta_3 \sin(\frac{\pi}{\Delta_{ij}}) t_{d_{ij}} + d_3 b_{1i}.
\end{align*}
\]
When \( C_{ij} = 0 \) (i.e., when \( Y_{ij} \) is not left censored or \( Y_{ij} \geq d \)), we generate \( Y_{ij} \) from a truncated normal distribution, i.e., \( Y_{ij} \sim N(\mu_{Y_{ij}}, \sigma^2) \) with a lower bound of \( d \), where

\[
\mu_{Y_{ij}} = \beta_0 + \beta_1 t_{ij}^* + \beta_2 t_{ij}^{*2} + \beta_3 \sin\left(\frac{\pi}{\Delta_{ij}} t_{ij}\right) + d_4 b_{1i}.
\]

We generate the survival times \( T_i \) from a Weibull distribution with shape parameter 15 and scale parameter 800 \( \exp(\lambda_0 x_i + \lambda_1 b_{1i} + \lambda_2 b_{2i})^{-1/15} \), where \( x_i \) is a baseline covariate generated from \( N(0,1) \). Moreover, we generate the non-informative right censored survival times \( C_i \) from a Weibull distribution with shape parameter 5 and scale parameter 1000, in order to achieve similar right censoring rates of the survival times as in the real data. The observed survival data \((s_i, \delta_i)\) are \( s_i = \min(T_i, C_i) \), with \( \delta_i = I(T_i \leq C_i) \) being the event indicator. The true parameter values in all the models are set as \( \alpha = (-1.65, 0.15, 1.8, -0.05)^T \), \( \beta = (2, 1, -0.3, 1.5)^T \), \( \eta = (1, -3, 0.9, -3.5)^T \), \( \lambda = (-0.75, -1.5, 2)^T \), \((d_1, d_2, d_3, d_4) = (0.4, 0.15, -1.5, 0.5)\), \( d = 1.5 \), \( \sigma = 0.5 \), and \( \Sigma = (0.5, 0.5) \).

To create outliers in \( Y \), we consider the following three scenarios: (i) \textit{Scenario 1} (e-outliers): We randomly select a subset (say 5%) of the simulated data \( Y \) and add a large number (say \( \pm \sigma \)) to the original values; (ii) \textit{Scenario 2} (b-outliers): We randomly select a subset (say 5%) of individuals and add a large number (say \( \pm 10 \)) to the generated random effects \( b_{1i} \); (iii) \textit{Scenario 3} (e-outliers + b-outliers): We add large numbers to a set (say 2.5%) of the observations of \( Y \) and to a set (say 2.5%) of the random effects \( b_{1i} \).

We evaluate the parameter estimates by relative bias (rBias) and relative root mean square error (rRMSE):

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

where \( \hat{\theta}^{(m)} \) is the estimate of \( \theta \) in simulation replication \( m \), \( M = 100 \) is the number of replications in all the simulations, and \( \theta \) is the true parameter value. To check if the standard error formulas based on the h-likelihood method given at the end of Section 3.3 are reliable or not, we compute two types of standard errors for each parameter estimate: average of the standard errors based on the h-likelihood method from simulation repetitions (denoted by SE), and the sample standard error of the estimates from simulation repetitions without using the h-likelihood formula (denoted by SSE). If the SE’s from the h-likelihood method are reliable, they should be close to the corresponding SSE’s. Finally, we report the estimated coverage probabilities of the 95% confidence intervals for the corresponding parameters using the h-likelihood-based SE’s.

5.2. Simulation results. Tables 2 – 4 provide the main simulation results: Table 2 summarizes the simulation results when the random effects \( b_{1i} \)’s contain 5% outliers by adding \( \pm 10 \); Table 3 summarizes the simulation results when there are 5% e-outliers in the \( Y \)-data by adding \( \pm 5\sigma \); Table 4 summarizes the simulation results when there are 2.5% outliers in the random effects \( b_{1i} \)’s by adding \( \pm 5 \) and 2.5% e-outliers in the measurements of \( Y \) by adding \( \pm 5\sigma \). The overall censoring rates for \( Y \) are about 13.5%, 15%, and 14.5% for Tables 2 – 4 respectively. More simulation results with different percentages of outliers can be found in the web-based Supplemental Materials. As noted earlier, in the proposed joint model, the key parameters of interest are \( \lambda_1 \) and \( \lambda_2 \) since they indicate if the individual longitudinal trajectories are associated with risk of HIV infection, which is our main objective. Thus, in the following we focus on these two parameters in the discussion of simulation results.
Table 2 shows that the robust method (R1 and R2) for b-outliers generally returns smaller relative biases (rBias) and relative rooted-mean-square-errors (rRMSE) in parameter estimates than the non-robust (NR) method, where the robust method is based on a multivariate $t_\nu$-distribution with degrees of freedom $\nu$. To check model mis-specification and perform sensitivity analysis, two cases are considered for the robust method: R1 ($\nu = 5$) and R2 ($\nu = 3$). We see that R2 is preferred for $\hat{\lambda}_1$ and $\hat{\lambda}_2$. This is because $t_3$ in R2 has heavier tails than $t_5$ in R1, so $t_3$ can accommodate more outliers. Note that the standard errors (SE) of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ in the survival model based on the proposed method are under-estimated, since the corresponding SE’s are smaller than the corresponding SSE’s (the sample SE from simulations). Thus, the corresponding coverage probabilities of the 95% confidence intervals are below the nominal level. This problem has been recognized in the joint-model literature when the baseline hazard in the survival model is completely unspecified (Hsieh, Tseng and Wang (2006); Rizopoulos (2012); Yu, Wu and Gilbert (2018)). In practice, we may consider bootstrap methods to obtain standard errors or consider a parametric survival model such as a Weibull survival model. Table 2 shows the standard errors and coverage probabilities based on the nonparametric bootstrap method for method R2, denoted by SE* and CP* respectively, and Table 7 in the web-based Supplemental Materials provides simulation results when a Weibull survival model is used for the case of $c = \infty$ and $b_i \sim t_3(0, \frac{1}{3}\Sigma)$. We see that the coverage probabilities of the key parameters $\lambda_1$ and $\lambda_2$ based on bootstrap are now close to the nominal level 0.95.

Table 3 shows that, in the presence of e-outliers only, the robust methods substantially outperform the non-robust method, with much smaller rBiases and rRMSEs for the key parameter estimates $\lambda_1$ and $\lambda_2$. To check model mis-specification and perform sensitivity analysis, we again consider two cases for the robust method with different turning point $c$ for the M-estimator: R1: $c = 2$ and R2: $c = 1.5$. We see that R1 leads to smaller rBiases and rRMSEs, which is consistent with the common choice of $c = 2$ for M-estimators in the literature of robust methods. Again, the coverage probabilities are smaller than the nominal level 95% due to under-estimated standard errors, as mentioned earlier. Table 8 in the web-based Supplemental Materials provides the simulation results when a Weibull survival model is used with $c = 2$ and $b_i \sim N(0, \Sigma)$. The robust method leads to reasonable coverage probabilities for $\lambda_1$ and $\lambda_2$ when a parametric survival model is considered.

Table 4 shows the simulation results with the presence of both b-outliers and e-outliers. The results are consistent with those in Tables 2 and 3 in terms of biases and coverage probabilities. The rRMSEs values are comparable across different methods.

From these simulation results, we can conclude that the proposed robust method performs well and is preferred to the non-robust method in the presence of outliers. The advantages of the robust method are more substantial for e-outliers than for b-outliers. This is not surprising since joint models are often less sensitive to the mis-specification of the random effect distribution (Rizopoulos, 2012). As noted above, one issue of the proposed robust method is the under-estimated standard errors for the parameters in the Cox PH model, which may lead to lower coverage rates and is a common issue in this joint model setting. On the other hand, the non-robust method, which ignores outliers, may inflate the standard errors and thus produce higher coverage rates, but it produces larger biases and MSEs in general. The issue of under-estimated standard errors may be addressed by bootstrap methods or considering a parametric survival model, such as a Weibull survival model, instead of Cox PH model.

6. Discussion. We have proposed a robust method for joint models of mixed types of longitudinal data and survival data, where the longitudinal data may contain outliers and left censoring. We address the e-outliers by an M-estimator and the b-outliers by a multivariate $t$-distribution. Moreover, we have proposed a computationally efficient approximate method
Table 2
Simulation results with 5% b-outliers.

<table>
<thead>
<tr>
<th>Par</th>
<th>True value</th>
<th>Estimate</th>
<th>SE</th>
<th>SSE</th>
<th>rBias%</th>
<th>rRMSE%</th>
<th>CP%</th>
</tr>
</thead>
<tbody>
<tr>
<td>α₀</td>
<td>-1.65</td>
<td>-1.70</td>
<td>-1.70</td>
<td>-1.69</td>
<td>0.09</td>
<td>0.09</td>
<td>0.10</td>
</tr>
<tr>
<td>α₁</td>
<td>0.15</td>
<td>0.16</td>
<td>0.16</td>
<td>0.15</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>α₂</td>
<td>1.80</td>
<td>1.84</td>
<td>1.84</td>
<td>1.11</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>α₃</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.05</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>β₀</td>
<td>2.00</td>
<td>2.01</td>
<td>2.00</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>β₁</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>β₂</td>
<td>-0.30</td>
<td>-0.30</td>
<td>-0.30</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>β₃</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>η₀</td>
<td>1.00</td>
<td>0.97</td>
<td>0.98</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>η₁</td>
<td>-3.00</td>
<td>-2.99</td>
<td>-2.97</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>η₂</td>
<td>0.90</td>
<td>0.89</td>
<td>0.89</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>η₃</td>
<td>-3.50</td>
<td>-3.52</td>
<td>-3.53</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>λ₀</td>
<td>-0.75</td>
<td>-0.77</td>
<td>-0.75</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>λ₁</td>
<td>-1.50</td>
<td>-1.59</td>
<td>-1.54</td>
<td>-1.45</td>
<td>0.22</td>
<td>0.20</td>
<td>0.19</td>
</tr>
<tr>
<td>λ₂</td>
<td>2.00</td>
<td>2.14</td>
<td>2.09</td>
<td>0.19</td>
<td>0.18</td>
<td>0.17</td>
<td>0.17</td>
</tr>
</tbody>
</table>

SE: average standard error. SSE: sample standard error; CP: coverage probability.

NR: non-robust method with \( b_i \sim N(0, \Sigma) \) and \( c = \infty \).
R1: robust method with \( b_i \sim t(\nu, \frac{1}{2}, \Sigma) \) and \( c = \infty \).
R2: robust method with \( b_i \sim t(\nu, \frac{1}{2}, \Sigma) \) and \( c = 5 \).

Table 3
Simulation results with 5% e-outliers.

<table>
<thead>
<tr>
<th>Par</th>
<th>True value</th>
<th>Estimate</th>
<th>SE</th>
<th>SSE</th>
<th>rBias%</th>
<th>rRMSE%</th>
<th>CP%</th>
</tr>
</thead>
<tbody>
<tr>
<td>α₀</td>
<td>-1.65</td>
<td>-1.69</td>
<td>-1.69</td>
<td>-1.68</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>α₁</td>
<td>0.15</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>α₂</td>
<td>1.80</td>
<td>1.85</td>
<td>1.85</td>
<td>1.11</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>α₃</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.05</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>β₀</td>
<td>2.00</td>
<td>1.94</td>
<td>1.98</td>
<td>0.04</td>
<td>0.03</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>β₁</td>
<td>1.00</td>
<td>1.04</td>
<td>1.01</td>
<td>0.07</td>
<td>0.05</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td>β₂</td>
<td>-0.30</td>
<td>-0.32</td>
<td>-0.31</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>β₃</td>
<td>1.50</td>
<td>1.55</td>
<td>1.51</td>
<td>0.04</td>
<td>0.03</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>η₀</td>
<td>1.00</td>
<td>1.01</td>
<td>0.99</td>
<td>0.13</td>
<td>0.12</td>
<td>0.12</td>
<td>0.11</td>
</tr>
<tr>
<td>η₁</td>
<td>-3.00</td>
<td>-3.08</td>
<td>-3.04</td>
<td>-3.04</td>
<td>0.25</td>
<td>0.25</td>
<td>0.24</td>
</tr>
<tr>
<td>η₂</td>
<td>0.90</td>
<td>0.94</td>
<td>0.92</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
<td>0.11</td>
</tr>
<tr>
<td>η₃</td>
<td>-3.50</td>
<td>-3.53</td>
<td>-3.49</td>
<td>-3.50</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>λ₀</td>
<td>-0.75</td>
<td>-0.80</td>
<td>-0.78</td>
<td>-0.77</td>
<td>0.16</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>λ₁</td>
<td>-1.50</td>
<td>-1.71</td>
<td>-1.57</td>
<td>-1.61</td>
<td>0.23</td>
<td>0.19</td>
<td>0.20</td>
</tr>
<tr>
<td>λ₂</td>
<td>2.00</td>
<td>2.28</td>
<td>2.11</td>
<td>0.34</td>
<td>0.28</td>
<td>0.27</td>
<td>0.33</td>
</tr>
</tbody>
</table>

SE: average standard error. SSE: sample standard error; CP: coverage probability.

NR: non-robust method with \( b_i \sim N(0, \Sigma) \) and \( c = \infty \).
R1: robust method with \( b_i \sim N(0, \Sigma) \) and \( c = 2 \).
R2: robust method with \( b_i \sim N(0, \Sigma) \) and \( c = 1.5 \).
### Table 4: Simulation results with 2.5% b-outliers and 2.5% e-outliers.

<table>
<thead>
<tr>
<th>Par</th>
<th>True Value</th>
<th>Estimate</th>
<th>SE</th>
<th>SSE</th>
<th>Bias%</th>
<th>rRMSe%</th>
<th>CP%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NR</td>
<td>R1</td>
<td>R2</td>
<td>R3</td>
<td>NR</td>
<td>R1</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>-1.65</td>
<td>-1.68</td>
<td>-1.69</td>
<td>-1.69</td>
<td>-0.09</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>-0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>1.80</td>
<td>1.84</td>
<td>1.83</td>
<td>1.84</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>-0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>2.00</td>
<td>1.99</td>
<td>1.98</td>
<td>1.97</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>1.00</td>
<td>1.01</td>
<td>1.02</td>
<td>1.01</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.30</td>
<td>-0.30</td>
<td>-0.31</td>
<td>-0.30</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>1.50</td>
<td>1.51</td>
<td>1.52</td>
<td>1.51</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>1.00</td>
<td>0.99</td>
<td>1.01</td>
<td>1.01</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>-3.00</td>
<td>-3.04</td>
<td>-3.02</td>
<td>-3.02</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>0.90</td>
<td>0.92</td>
<td>0.90</td>
<td>0.90</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>$\gamma_3$</td>
<td>-3.50</td>
<td>-3.50</td>
<td>-3.53</td>
<td>-3.54</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>$\lambda_0$</td>
<td>-0.75</td>
<td>-0.77</td>
<td>-0.77</td>
<td>-0.78</td>
<td>0.15</td>
<td>0.15</td>
<td>0.14</td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>-1.50</td>
<td>-1.61</td>
<td>-1.55</td>
<td>-1.56</td>
<td>0.20</td>
<td>0.17</td>
<td>0.18</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>2.00</td>
<td>2.15</td>
<td>2.11</td>
<td>2.13</td>
<td>0.27</td>
<td>0.19</td>
<td>0.19</td>
</tr>
</tbody>
</table>

NR: non-robust method with $b_i \sim N(0, \Sigma)$ and $c = \infty$.
R1: robust method with $b_i \sim t_{(c, 0.5)}(\Sigma)$ and $c = 2$.
R2: robust method with $b_i \sim \alpha_t(0, \Sigma)$ and $c = 1.5$.
R3: robust method with $b_i \sim t_{(c, 0.5)}(\Sigma)$ and $c = 2$.

SE: average standard error. SSE: sample standard error; CP: coverage probability.

Based on the h-likelihood for parameter estimation. We find that parameter estimates in joint models of longitudinal and survival data are very sensitive to e-outliers but are less sensitive to b-outliers in the longitudinal data. Thus, in data analysis using joint models, we should focus more on the outlying observations among the repeated measurements within an individual than outlying individuals.

Based on the proposed method, we analyze the motivating HIV vaccine data and find a strong association between the individual-specific longitudinal immune response trajectories and the risk of HIV infection. This finding is important since we may then possibly predict the risk of HIV infection based on early longitudinal trajectories of potential biomarkers, avoiding the more expensive approach of collecting all the longitudinal data later in the study. An advantage of using random effects from the longitudinal models as predictors in the survival model is that such an approach avoids the complications of the longitudinal values being censored or missing or mis-measured at event times. Moreover, the random effects may be viewed as summaries of the longitudinal trajectories. For this specific HIV vaccine data, the analysis results based on the robust method and the non-robust method seem close. However, the proposed robust method allows us to be more confident about the analysis results and conclusions since left censoring and potential outliers are addressed. Moreover, simulation results show that the robust method may perform much better than the non-robust method in some cases. Thus, the proposed method may be useful for other datasets. Finally, the proposed method may be used as a tool for sensitivity analysis to see if analysis results are sensitive to left censoring and potential outliers.

In this study, since the biomarkers are highly correlated, it may be difficult to recognize which of the biomarkers leads to the event or perhaps the biomarkers jointly lead to the event. As we have done here, a simple approach would be to choose one or two representative biomarkers in the modelling, since a multivariate longitudinal model for all the biomarkers may involve too many parameters, making the joint model over complicated. In future research, we might consider a single underlying latent process which governs all the longitudi-
dinal biomarkers. We might also consider a principal component analysis for the longitudinal biomarkers and then choose one or two longitudinal principal components in the joint model.

We have assumed a point mass for the censored values in order to avoid a unverifiable parametric distributional assumption for the censored values. This may lead to some loss of information, compared to a method assuming a unverifiable parametric distribution for the censored values (e.g., Hughes (1999)). However, since no information regarding the distribution of the censored values is available, there is actually little loss of information, so avoiding a unverifiable distributional assumption may be a reasonable approach.

Kong and Nan (2016) considered a left-censored covariate in regression models for cross-sectional data. They used survival analysis methods for the left-censored covariate after converting the left-censoring to right censoring, similar to the methods used in survival analysis for right-censored data. They assumed an accelerated failure time (AFT) model for the transformed covariate, and included other observed predictors (covariates) in the AFT model, without distributional assumption for the error term of the AFT model. Then, they used the nonparametric Kaplan-Meier estimator to estimate the residual distribution. While they did not explicitly assume a point mass for the censored values, the Kaplan-Meier estimator of the error distribution is based on the observed data of the covariate with censoring, and they used other observed data as predictors in the AFT model to improve the estimation. They claim that the efficiency gain of their method depends on the strength of the association between the covariate with censoring and other predictors. Essentially, this is similar to our idea of using other observed predictors to possibly predict the censoring status of $Y^*$ when $Y^*$ is not measured, based on a logistic regression model. Neither methods assume a distribution for the censored values. A comparison of the two methods will be reported separately.

Although there has been extensive research in joint models in the last two decades, computation remains a main challenge for statistical inference of joint models. The approximate method based on the h-likelihood appears to work reasonably well and is computationally efficient. The h-likelihood is essentially a modification of the Laplace approximation. Such an approximation often performs well for continuous data but may perform less satisfactory for discrete data such as binary data. Its performance may also depend on the frequencies of the repeated measurements. We plan to investigate the approximate method further via more extensive simulation studies.

The models in the paper are motivated from a real dataset from an HIV vaccine study, so they may also be applicable in other similar studies. While the models appear to be complicated, the main longitudinal models are in fact a linear mixed effects (LME) model for the continuous response and a logistic mixed effects model for the binary response. Both types of models are commonly used in practice. The LME model in the data analysis appears to be complicated since it incorporates the periodic patterns of the longitudinal data due to repeated vaccinations over time, but it is still an LME model. Left censoring and outliers are common problems in many longitudinal studies. Therefore, the proposed models are more generally applicable, and the LME model may be simpler in other applications.

7. Software. R code and a sample data set are available at https://github.com/oliviayu/robust-HHJMs or upon request to the first author.
SUPPLEMENTARY MATERIAL

Supplementary material for “Robust joint modelling of left-censored longitudinal data and survival data, with application to HIV vaccine studies”

The supplementary material includes four sections: A: Proof of asymptotic properties; B: Model selection procedure; C: Model Diagnostics of the Real Data Analysis; and D: Additional simulation results.

REFERENCES


