



Analysis of Censored Longitudinal Data with Skewness and a Terminal Event

Xiao Su & Sheng Luo

To cite this article: Xiao Su & Sheng Luo (2016): Analysis of Censored Longitudinal Data with Skewness and a Terminal Event, Communications in Statistics - Simulation and Computation, DOI: [10.1080/03610918.2016.1157181](https://doi.org/10.1080/03610918.2016.1157181)

To link to this article: <http://dx.doi.org/10.1080/03610918.2016.1157181>

 View supplementary material 

 Accepted author version posted online: 21 Mar 2016.

 Submit your article to this journal 

 View related articles 

 View Crossmark data 

Analysis of Censored Longitudinal Data with Skewness and a Terminal Event

Xiao Su,¹ Sheng Luo^{1,*}

¹Department of Biostatistics, School of Public Health, The University of Texas Health Science Center at Houston, 1200 Pressler St., Houston, Texas 77030, USA

**email*: sheng.t.luo@uth.tmc.edu

In HIV/AIDS study, the measurements viral load are often highly skewed and left-censored because of a lower detection limit. Furthermore, a terminal event (e.g., death) stops the follow-up process. The time to terminal event may be dependent on the viral load measurements. In this article, we present a joint analysis framework to model the censored longitudinal data with skewness and a terminal event process. The estimation is carried out by adaptive Gaussian quadrature techniques in SAS procedure NLMIXED. The proposed model is evaluated by a simulation study and is applied to the motivating Multicenter AIDS Cohort Study (MACS).

Key Words: Detection limit; Informative censoring; Joint model; Skew distributions; Tobit model.

1 Introduction

In many AIDS studies, the infection and progression of human immunodeficiency virus type 1 (HIV-1) are usually measured by viral load (plasma HIV-1 RNA copies) and CD4 cell count (the number of CD4+ T lymphocytes per volume of blood). Viral load is a measurement of the severity of a viral infection, and can be calculated by estimating the amount of virus in plasma. To model the viral load trajectory as a function of CD4 count and other risk factors, linear and non-linear mixed-effects models have been widely used (e.g., Davidian, 1995; Huang and Dagne, 2012). Despite the improvement of measurement technology, viral load measurements are still subject to censoring due to limits of detection (LOD), e.g., left censoring due to a lower LOD at 50 copies/ml in ultra sensitive assay (Schockmel et al., 1997). To address this issue of censored longitudinal response variables, a common practice is to impute the censored values by the LOD or some values such as half of LOD. These ad hoc imputation methods may produce biased estimates, standard errors, and prediction (e.g., Hughes, 1999; Thiébaud et al., 2006). Alternatively, Tobit models (e.g., Tobin, 1958; Lynn, 2001; Sattar et al., 2011) that treat all censored measurements as missing values of a latent variable are often used. The Tobit models explicitly incorporate into the likelihood function both the probability that an observation is below LOD and the probability distribution of an observation given that it is above the LOD. The Tobit models, which assume normal distributions for random errors, usually provide consistent parameter estimates when the normality assumption is satisfied.

However, the viral load measurements are often heavily skewed, even after some transformation. The Tobit models lack robustness against departure from the normality and outliers (e.g., Sahu et al., 2003; Bandyopadhyay et al., 2012) and give inconsistent results when the normality assumption for the random errors is violated (Dagne and Huang, 2012). Thus, it is essential to replace the normal distributions with some more flexible skewed distributions. Azzalini (1985) has considered a skew-normal (SN) distribution and studied its properties. Multivariate SN distributions have been studied in the literature (Sahu et al., 2003) and applied to several kinds of

models (e.g., Ghosh et al., 2007; Huang and Dagne, 2011; Bandyopadhyay et al., 2012). However, there is no literature discussing the SN distribution in the Tobit models. Furthermore, during the course of AIDS studies, the follow-up of some individuals is stopped by terminal events such as death, dropout due to adverse event (AE) or severe adverse event (SAE), or some other events. Because the terminal events may be related to the individuals' viral load measurement, the terminal mechanism is outcome-dependent. The dependent terminal event time is often termed "dependent censoring" or "informative censoring". Ignoring the dependent censoring leads to biased estimates (Henderson et al., 2000). To address this issue, joint analysis of survival with repeated measures has been increasingly common (e.g., Wulfsohn and Tsiatis, 1997; Henderson et al., 2000). Tsiatis and Davidian (2004) and Yu et al. (2004) give excellent reviews of joint modeling research. To the best of our knowledge, there is no work done to simultaneously account for skewness and dependent terminal events in the framework of Tobit models. It is not clear how these two features may interact and simultaneously influence the inferential procedures. The goal of this article is to investigate the covariate effects in a Tobit model when these two features exist in a longitudinal prospective study. Specifically, we replace the normality assumption for random errors by multivariate skew-normal (SN) distributions and propose a joint model framework to account for the dependent terminal events.

The remainder of the article is organized as follows. In Section 2, we describe the dataset that motivates this research. Section 3 introduces the skew-normal distributions and proposes a joint model for the skewed viral load response variable subject to dependent terminal events, in addition to the estimation procedure. Section 4 provides an extensive simulation study to assess the performance of the proposed joint model. Section 5 applies the proposed model to the motivating dataset. Section 6 gives some concluding remarks and discussions.

2 Motivating Data

This article is motivated by the Multicenter AIDS Cohort study (MACS), an ongoing prospective study of the natural and treated histories of HIV-1 infection in homosexual and bisexual men. The study participants had baseline and semiannual follow-up visits. The collected variables include patient characteristics (demographic data, medical history), physical examination, AIDS-related conditions (e.g., blood count, viral load, CD4 count), and repository storage (e.g., serum, plasma, urine, semen, etc.). Our data analysis is based on 1,339 participants with complete data in viral load, CD4 count, ethnicity, date of birth, time to death/censoring.

To study HIV disease progression, we are interested in the relationship between viral load measurements and CD4 cell count over time while adjusting for other variables (e.g., ethnicity, age, time). These virologic and immunologic markers have been shown to predict progression to AIDS (Mellors et al., 1997). As pointed out by Huang and Dagne (2011), CD4 cell count, viral load, or both may be treated as responses in AIDS studies. In this article, viral load is used as a primary endpoint and CD4 cell count is viewed as a covariate to help predict virologic responses. One caveat here is that the viral load measurements are subject to left censoring due to a lower limit of detection (LOD) at 50 copies/ml. The viral load measurements below this LOD are not accurate. Therefore, we censor these measurements (3,035 out of 14,445 measurements, or 21.0%) at the LOD of 50. Moreover, the viral load measurements are skewed, even after log transformation (mean, 8.509; median, 9.379). To visualize this, Figure 1 (panels a and b) display the density histogram and associated Q-Q plots for the repeated and uncensored viral load measurements (in natural log scale), which reveals some degree of left skewness (estimated skewness is -0.562). Panels c and d present the histogram and Q-Q plots of the estimates of residuals, obtained after fitting a Tobit model (9) to the MACS dataset. These two plots reveal some left-skewness of the residuals. To this end, some more flexible skewed distributions are to be used for the random errors.

During the course of the MACS study, there are 973 (69.5%) deaths that occurred. It is observed

that the viral load measurements are associated with the survival time. The left panel of Figure 2 shows plot of mean log viral load values over time for MACS participants with follow-up time less than 16 years (1,030 participants, dotted line) and more than 16 years (309 participants, solid line). Participants with shorter follow-up time tend to have higher log viral load values, indicating that participants with more severe viral infection are more likely to experience death. The right panel of Figure 2 displays Kaplan-Meier curves showing the difference in time to death for participants with high (median log viral load > 9.38 , dashed line) and low (median log viral load ≤ 9.38 , solid line) viral load measurements. Participants with higher viral load values have much shorter time to death than the ones with lower viral load values (log-rank test $p < 0.001$). These two plots manifest the strong correlation between the viral load values and the time to death.

3 Statistical models and likelihood inference

3.1 The multivariate skew-normal distribution

We begin with an introduction of multivariate skew-normal (SN) distributions. Following the notation of the SN family in Azzalini (2005) and Arellano-Valle et al. (2006), a p -dimensional random vector $\mathbf{Y} = (y_1, \dots, y_p)'$ has a p -dimensional SN distribution with a $p \times 1$ location vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)'$, a $p \times p$ positive definite dispersion matrix $\boldsymbol{\Sigma}$, and a $p \times 1$ skewness parameter vector $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_p)'$, if its density is given by

$$SN(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = 2\phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi_p(\boldsymbol{\lambda}'\boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})), \quad (1)$$

where $\phi_p(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\Phi_p(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ are the probability distribution function (PDF) and the cumulative distribution function (CDF), respectively, of the p -variate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1}$. We denote $\mathbf{Y} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ for a random vector with the above density function. The mean and covariance of \mathbf{Y} are $E(\mathbf{Y}) = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\boldsymbol{\Delta}$ and $Var(\mathbf{Y}) = \boldsymbol{\Sigma} - \frac{2}{\pi}\boldsymbol{\Delta}\boldsymbol{\Delta}'$, respectively, with $\boldsymbol{\Delta} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}$ and $\boldsymbol{\delta} = \boldsymbol{\lambda}/\sqrt{1 + \boldsymbol{\lambda}'\boldsymbol{\lambda}}$. Note that if the

skewness vector $\lambda = \mathbf{0}$, then the density of \mathbf{Y} reduces to $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Moreover, if $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)'$, $\lambda_1 = \dots = \lambda_p = \lambda$, and $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_p$ with \mathbf{I}_p being a $p \times p$ identity matrix, then $\mathbf{Y} \sim SN_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p, \lambda \mathbf{1}_p)$ where $\mathbf{1}_p$ is a $p \times 1$ vector of 1. Equivalently, the univariate formulation of (1) is $y_k \sim SN(\mu_k, \sigma^2, \lambda)$ for $k = 1, \dots, p$, which is the case to be used in this article. This univariate SN distribution has the PDF

$$f(y_k | \mu_k, \sigma^2, \lambda) = 2\phi\left(\frac{y_k - \mu_k}{\sigma}\right)\Phi\left(\lambda \frac{y_k - \mu_k}{\sigma}\right), \quad (2)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the PDF and CDF, respectively, of the standard normal distribution. The univariate SN distribution has the CDF

$$F(y_k | \mu_k, \sigma^2, \lambda) = \Phi\left(\frac{y_k - \mu_k}{\sigma}\right) - 2T\left(\frac{y_k - \mu_k}{\sigma}, \lambda\right), \quad (3)$$

where $T(h, a) = \frac{1}{2\pi} \int_0^a \frac{e^{-\frac{1}{2}h^2(1+x^2)}}{1+x^2} dx$ with values between 0 and 1 is Owen's T function (Owen, 1956). The mean and variance of y_k are $\mu_k + \sigma\delta \sqrt{\frac{2}{\pi}}$ and $\sigma^2(1 - \frac{2\delta^2}{\pi})$, respectively, with $\delta = \frac{\lambda}{1+\lambda^2}$. As a special case, the density of y_k reduces to $N(\mu_k, \sigma^2)$, when the skew parameter $\lambda = 0$.

3.2 Model and notation

Let y_{ij}^* be the response value (e.g., viral load) for individual i ($i = 1, \dots, N$) at time t_{ij} ($j = 1, \dots, n_i$), subject to censoring due to upper and lower limits of detection (LOD), where n_i is the number of repeated measurements for individual i . For simplicity, we only consider left censoring with a lower limit d , but our method can be easily extended to right censoring, double censoring, and even time-varying LOD. The observed value of y_{ij}^* is y_{ij} , where $y_{ij} \geq d$. Let $\mathbf{y}_i^* = (y_{i1}^*, \dots, y_{in_i}^*)'$ and $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})'$. To model the covariate effects, we consider a linear mixed effects (LME) model with a SN distribution.

$$\mathbf{y}_i^* = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}_i + \boldsymbol{\epsilon}_i, \quad (4)$$

with the assumption that

$$\mathbf{u}_i \sim^{\text{iid}} N_q(0, \boldsymbol{\Sigma}_u) \text{ and } \boldsymbol{\epsilon}_i \sim^{\text{iid}} SN_{n_i}(0, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\lambda}), \quad (5)$$

where \mathbf{X}_i is the $n_i \times p$ design matrix corresponding to the fixed effects (e.g., CD4 count, age, and time, etc), $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, \mathbf{Z}_i is the $n_i \times q$ design matrix corresponding to the $q \times 1$ vector of random effects \mathbf{u}_i , and $\boldsymbol{\epsilon}_i$ is the $n_i \times 1$ vector of random errors following the SN distribution with $n_i \times n_i$ dispersion matrix $\boldsymbol{\Sigma}_\epsilon$ and $n_i \times 1$ skewness parameter vector $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_{n_i})'$. We refer to models (4) and (5) as the multivariate skew-normal Tobit model.

Next, we discuss the model formulation under the univariate SN distribution. If $\lambda_1 = \dots = \lambda_{n_i} = \lambda$, and $\boldsymbol{\Sigma}_\epsilon = \sigma_\epsilon^2 \mathbf{I}_{n_i}$ with \mathbf{I}_{n_i} being a $n_i \times n_i$ identity matrix, then $\boldsymbol{\epsilon}_i \sim SN_{n_i}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_{n_i}, \lambda \mathbf{1}_{n_i})$ where $\mathbf{1}_{n_i}$ is a $n_i \times 1$ vector of 1. Equivalently, ϵ_{ij} follows univariate SN distribution $SN(0, \sigma_\epsilon^2, \lambda)$ for $j = 1, \dots, n_i$, which is the case used in this article. The skewness parameter λ indicates the degree of skewness (skewness increases as λ increases in absolute value), with positive λ being right-skewed and negative λ being left-skewed. If $\lambda = 0$, then $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. Thus, when $\epsilon_{ij} \sim SN(0, \sigma_\epsilon^2, \lambda)$,

$$y_{ij}^* \sim SN(\mu_{ij}, \sigma_\epsilon^2, \lambda) \text{ with } \mu_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{u}_i, \quad (6)$$

where \mathbf{X}_{ij} and \mathbf{Z}_{ij} are the j th row of matrices \mathbf{X}_i and \mathbf{Z}_i , μ_{ij} is the conditional mean given the random effects \mathbf{u}_i . The response value y_{ij}^* has PDF $f(y_{ij}^*|\mu_{ij}, \sigma_\epsilon^2, \lambda)$ as formulated in (2) and CDF $F(y_{ij}^*|\mu_{ij}, \sigma_\epsilon^2, \lambda)$ as formulated in (3). We refer to model (6) as the univariate skew-normal Tobit model.

Let t_i denote the time to a terminal event for individual i , δ_i (1 if the terminal event is observed and 0 otherwise) denote the censoring indicator for t_i , and \mathbf{W}_i denotes vector of possible risk factors. Under the Cox proportional hazard model, the hazard of a terminal event is

$$h(t_i) = h_0(t_i) \exp(\mathbf{W}_i\boldsymbol{\gamma} + \boldsymbol{\nu}\mathbf{u}_i), \quad (7)$$

where $h_0(\cdot)$ is the baseline hazard function, \mathbf{u}_i is the shared random effects accounting for the correlation between the longitudinal and survival processes and $\boldsymbol{\nu}$ measures their association. We

consider two types of baseline hazard function: Weibull distribution (e.g., $h_0(t_i) = \alpha \lambda_D t_i^{\alpha-1}$) and piecewise constant function. Lawless and Zhan (1998) and Feng et al. (2005) illustrated that models using a piecewise constant baseline hazard yield good estimators for both fixed effects and frailty. Piecewise constant baseline hazard function has been widely used in the literature (Liu and Huang, 2009). Given a set of fixed time points $0 = \tau_0 < \tau_1 < \dots < \tau_m$, and the baseline hazard vector $\mathbf{g} = (g_0, g_1, \dots, g_{m-1})$, we define the piecewise constant hazard function as $h_0(t_i) = \sum_{l=0}^{m-1} g_l I_l(t_i)$, with $I_l(t_i) = 1$ if $\tau_l \leq t_i < \tau_{l+1}$ and 0 otherwise. The survival function $S(t_i) = \exp[-\int_0^{t_i} h(s) ds]$. The unknown parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Sigma}_u, \sigma_\epsilon, \lambda, \boldsymbol{\gamma}, \boldsymbol{\nu}, \alpha, \lambda_D)$ if $h_0(t_i)$ follows a Weibull distribution or $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Sigma}_u, \sigma_\epsilon, \lambda, \boldsymbol{\gamma}, \boldsymbol{\nu}, \mathbf{g})$ if $h_0(t_i)$ follows a piecewise constant function. Under conditional independence assumption (given random effects \mathbf{u}_i , the longitudinal and survival processes are independent), the conditional likelihood of observing data $\mathbf{D}_i = (\mathbf{y}_i, t_i, \delta_i)$ is

$$L_i(\boldsymbol{\theta}|\mathbf{u}_i, \mathbf{D}_i) = \left[\prod_{j=1}^{n_i} f(y_{ij}|\mu_{ij}, \sigma_\epsilon^2, \lambda)^{1-I(y_{ij}=d)} F(d|\mu_{ij}, \sigma_\epsilon^2, \lambda)^{I(y_{ij}=d)} \right] \cdot h(t_i)^{\delta_i} S(t_i). \quad (8)$$

We refer to the proposed joint models $L_i(\boldsymbol{\theta}|\mathbf{u}_i, \mathbf{D}_i)$ assuming the skew-normal and normal distributions for ϵ_{ij} as models JM_{SN} and JM_{N} , respectively. Moreover, we consider the reduced models assuming independence between the longitudinal outcomes and survival time (i.e., $\boldsymbol{\nu} = 0$). We refer to the reduced models assuming the skew-normal and normal distributions as models RM_{SN} and RM_{N} , respectively.

3.3 Maximum likelihood estimation

The marginal likelihood for one individual is $L_i(\boldsymbol{\theta}|\mathbf{D}_i) = \int L_i(\boldsymbol{\theta}|\mathbf{u}_i, \mathbf{D}_i) f(\mathbf{u}) d\mathbf{u}$, where $f(\mathbf{u})$ is $N_{n_i}(0, \boldsymbol{\Sigma}_u)$. This likelihood function involves an integral with respect to random effects and the integral cannot be evaluated analytically. Numerical integration such as Laplace approximation (Liu et al., 2008) or Gaussian quadrature (Liu and Huang, 2009) can be used for estimation. Liu et al. (2010) pointed out that both methods generally perform well, with the Laplace approximation being much faster in high dimensional random effects settings. However, Laplace approximation is

not yet available in commercial software packages for fitting nonlinear mixed effects models, while Gaussian quadrature can be conveniently implemented in SAS procedure NLMIXED. Thus, adaptive Gaussian quadrature is adopted to approximate the integral. In numerical analysis, a quadrature rule is an approximation of the definite integral by a weighted sum of function values at specified points within the domain of integration. Moreover, the adaptive Gaussian quadrature method accounts for the shape of the likelihood when placing quadrature points, which is more efficient and provides a better approximation than the non-adaptive Gaussian quadrature with equally spaced points (Lesaffre and Spiessens, 2001). In addition to accurate parameter estimates and available standard error estimates, the adaptive Gaussian quadrature method possesses the advantage of easy implementation because SAS procedure NLMIXED only requires inputting the likelihood (conditional on random effects) explicitly and the approximation of the marginal likelihood can be directly maximized. To facilitate easy reading and implementation of the proposed models, a sample SAS code for fitting the proposed model with a piecewise constant baseline hazard function has been presented in the Web Supplement.

One caveat here is that the likelihood in model (8) involves Owen's T function as defined in model (3), which is unavailable in SAS. To solve this problem, we use a C program to compute Owen's T function and call it from SAS. This C program (`asa076.c`) can be downloaded from Dr. John Burkardt's website at http://people.sc.fsu.edu/~jburkardt/c_src/asa076/asa076.html. We compile this C program and a header file (`asa076.h`) into a shared library in the format of `dll` (Windows-based dynamic-link library). Then we use SAS procedure `PROTO` to link SAS to the shared library and define a SAS function (in the name of `OwenT`) using the `FCMP` procedure. Then this user-defined SAS function can be called whenever the Owen's T function needs to be computed. Please refer to the Web Supplement for the details of coding and refer to the SAS help file at <http://support.sas.com/kb/40/562.html> for details of calling C programs from SAS.

We next illustrate how to give appropriate initial values for the parameter vector θ . For the regression parameters β and random effects variance Σ_u in Eq (6), we fit a linear mixed model (using `proc mixed` or `proc glimmix` in SAS) and use the estimates as the initial values. For the

regression parameters γ in Eq (7), we fit a parametric Cox model (using `proc phreg` in SAS) and use the estimates as the initial values.

4 Simulation Study

In this section we report results from an extensive simulation study of four settings to compare the performance of the proposed joint models and reduced models. In each setting, we generate 3,000 datasets with sample size $N = 500$. The simulated data structure is similar to but simpler than the MACS study. The joint model that we adopt is

$$y_{ij}^* = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 t_{ij} + u_i + \epsilon_{ij},$$

$$h_i(t_i) = h_0(t_i) \exp(\gamma_1 x_{i1} + \gamma_2 x_{i2} + \nu u_i),$$

where y_{ij} is subject to left censoring at $d = 4.5$.

We assume the repeated measurements y_i^* are observed at five time points (t_{ij} with $j = 1, \dots, n_i = 5$). Covariates for both submodels include a continuous variable x_{i1} sampled from standard normal distribution, a binary variable x_{i2} which takes value 0 or 1 each with probability 0.5, and time variable t_{ij} to denote the linear time trend. The random errors ϵ_{ij} are simulated from $N(0, \sigma_\epsilon^2)$ or $\epsilon_{ij} \sim SN(0, \sigma_\epsilon^2, \lambda)$ with various σ_ϵ and λ . The random effects u_i are simulated from $N(0, \sigma_u^2)$ with various σ_u^2 .

The baseline hazard functions for terminal events are $h_0(t_i) = 0.0157t_i^{-0.57}$ (Settings I, II, and III) and $h_0(t_i) = 2t_i$ (Setting IV), respectively. The independent censoring time censors about 30% of the total individuals. We apply the estimation framework in Section 3.3 to obtain inference. The simulation results presented in Tables 1 to 4 report the bias, the standard error (SE) of the parameter estimates, the mean of the standard error estimates (SEM), and coverage probability (CP) of the 95% confidence interval for each parameter of interest under the joint and reduced models defined in Section 3.

In Setting I, there is no correlation between the survival time and the longitudinal outcomes

(i.e., $\nu = 0$) and the longitudinal data have either normal (upper table) or skew-normal (lower table) distributions. Table 1 suggest that when the longitudinal data have either normal or skew-normal distributions, both the reduced and joint models generate comparable results, i.e., the bias is negligible, SE is generally close to SEM, and the confidence interval coverage probabilities are reasonably close to 95%. Under model overparameterization, the estimates of ν from both joint models JM_N and JM_{SN} are correctly close to zero.

In Setting II, the death hazard shares random effects u_i with the longitudinal model with $\nu = 0.3$, indicating that higher repeated measures (e.g., viral load) are associated with a higher death rate. The longitudinal data have either normal (upper table) or skew-normal (lower table) distributions. Table 2 suggests that in Setting II with some correlation (i.e., $\nu = 0.3$), both joint models provide estimates of all parameters with negligible bias, SE being generally close to SEM, CP being reasonably around the nominal value. In contrast, the reduced models give biased estimates (bias is generally one order of magnitude larger than the joint model counterparts) and low coverage probabilities, especially for the Tobit model regression parameters β , σ_u^2 , and the survival regression parameter γ .

In Setting III, the terminal event is dependent on the skew-normally distributed longitudinal outcome with skewness $\lambda = -1$ and $\sigma_\epsilon = 1$. The results in Table 3 suggest that in the proposed joint model framework, the correct model JM_{SN} provides reasonable estimates, while model JM_N also provides reasonable estimates for all parameters except the Tobit model intercept β_0 (because of the intercept shift from the SN distribution, i.e., $E(\epsilon) = \sigma_\epsilon \delta \sqrt{2/\pi} = -0.399$). Moreover, because the skewness in longitudinal measurements is modeled by the parametric skew-normal distribution, it is essential to assess the robustness of the proposed joint models under model misspecification. Table 4 displays the results of fitting the joint models JM_{SN} and JM_N in Setting IV when the random errors are simulated from a skew-t distribution (degree of freedom $df=5$ and skewness $\lambda = 1.5$). The results in Table 4 suggest that model JM_{SN} , by accounting for the skewness, provides reasonable estimates for all parameters (except β_0) with relatively small bias and SE being generally close to SEM. In comparison, model JM_N gives severely biased estimates for the Tobit model regression

parameter β and the random effects variance. The poor estimate of β_0 from model JM_{SN} should not be surprising because of the incorrect assumption of the random errors distribution.

From the simulation study, we conclude that in the presence of independent terminal events, the joint models provide results comparable to the reduced models and the estimate of the parameter ν is correctly close to zero, when the longitudinal data have either normal or skew-normal distributions. Under the dependent terminal mechanism, the joint models provide more accurate estimates for all parameters than their reduced model counterparts, when the longitudinal data have either normal or skew-normal distributions. Furthermore, when the random errors distribution is skewed but its distribution is misspecified, the proposed joint model JM_{SN} , by accounting for the skewness, still provides reasonably accurate estimates for all parameters except the longitudinal regression intercept.

5 Real Data Analysis

We apply our model to the MACS dataset. Among the 1,339 participants included in this analysis, the number of visits varies (range, 2 to 46; mean, 10.8; median, 7.0). There are 973 (69.5%) deaths in the dataset. Our analysis includes the following covariates: longitudinal CD4 (range, 0 – 790; mean 232.3), BCD4 (baseline CD4 count; range, 6 – 790; mean, 340.1), ethnicity (1, white; 0, otherwise; 90.0% of white), age (baseline age; range, 18 – 66; mean, 35.4), time (in years; range, 1 – 23; mean, 8.2, median, 7). The covariate age is centered at the mean and divided by 10. The covariates CD4 and BCD4 are transformed by square root. The response variable is viral load (in nature log scale) and the survival event is time to death in years. The model for the log-transformed viral load is

$$y_{ij}^* = \beta_0 + \beta_1 \sqrt{CD4_{ij}} + \beta_2 \text{Ethnicity}_i + \beta_3 \text{Age}_i + \beta_4 \text{time}_{ij} + u_i + \epsilon_{ij}. \quad (9)$$

The model for the hazard of death is

$$h(t_i) = h_0(t_i) \exp(\gamma_1 \sqrt{BCD4_i} + \gamma_2 \text{Ethnicity}_i + \gamma_3 \text{Age}_i + \nu u_i), \quad (10)$$

We make inference using the SAS procedure NLMIXED with the adaptive Gaussian quadrature estimation method of 50 quadrature points. For baseline hazard $h_0(\cdot)$, we use both Weibull distribution and piecewise constant function (10 intervals with cutpoints at every 1/10th quantiles). Table 5 compares all joint and reduced models under various assumptions of baseline hazard functions using -2LogLik (-2 times the log likelihood), the Akaike information criterion (AIC) (Akaike, 1974) and Bayesian information criterion (BIC) (Schwarz, 1978) as model selection criterion. All joint models perform significantly better than their reduced model counterparts with smaller -2LogLik , AIC and BIC values, suggesting that the joint models are more preferable than their reduced model counterparts. Moreover, the models with piecewise constant baseline hazard functions perform significantly better than their counterparts with Weibull baseline hazard functions. Model JM_{SN} with a piecewise constant baseline hazard function is selected as the final model because it has the best predictive ability with the smallest -2LogLik , AIC and BIC values.

Table 6 provides the means, SE, and 95% confidence interval (CI) of the parameters from the final model JM_{SN} and its reduced model counterpart RM_N with piecewise constant baseline hazard functions. It is observed that both models give different estimates for all parameters, although the same set of parameters are identified for significance by both models. For example, CD4 count has significant effects on log viral load, i.e., the log viral load is expected to decrease by 0.395 (95% CI: $[-0.407, -0.382]$) for 1 units increase of the square root of CD4 count in model JM_{SN} v.s. 0.406 (95% CI: $[-0.418, -0.394]$) in model RM_N . The log viral load is expected to decrease by 0.370 (95% CI: $[-0.497, -0.243]$) for 10 years increase in baseline age in model JM_{SN} v.s. 0.379 (95% CI: $[-0.505, -0.252]$) in model RM_N . The log viral load is expected to decrease by 0.325 (95% CI: $[-0.334, -0.315]$) for 1 year increase in follow-up time in model JM_{SN} v.s. 0.344 (95% CI: $[-0.353, -0.336]$) in model RM_N . Furthermore, the estimate of the skewness parameter is significantly different from zero ($\hat{\lambda} = -0.962$, 95% CI: $[-1.225, -0.698]$), indicating some left skewness in the log-transformed viral load measurements, as manifested in Figure 1.

Moreover, baseline CD4 count and age have significant influence on the hazard of death. The hazard rate of death is 0.861 (i.e., $\exp(-0.149)$; 95% CI: $[0.843, 0.881]$) for every 1 units increase in

the square root of baseline CD4 count in model JM_{SN} v.s. 0.895 (95% CI: [0.877, 0.912]) in model RM_N . The hazard rate of death is 1.163 (i.e., $\exp(0.151)$; 95% CI: [1.060, 1.276]) for every 10 years increase in baseline age in model JM_{SN} v.s. 1.143 (95% CI: [1.048, 1.249]) in model RM_N . We also observe the estimate of γ from model JM_{SN} is significantly different from zero ($\hat{\gamma} = 0.232$, 95% CI: [0.176, 0.289]), indicating that individuals with higher viral load (higher u_i) tend to have higher hazard to death and vice versa, which justifies the need of joint modeling. For completeness, we present the results from joint model JM_N and reduced model RM_{SN} assuming piecewise constant baseline hazard function in the Web Supplement (Web Tables 1 and 2). Similar conclusions can be made from those models, although the parameter estimates are slightly different.

6 Discussion

In this article, we propose a joint modeling framework which consists of a skew-normal Tobit model for the highly skewed longitudinal HIV viral load data subject to censoring and a Cox proportional hazard model for a dependent terminal event (e.g., death). The skew-normal Tobit model and the Cox model are correlated via shared random effects. Our simulation study shows that in the scenario of independent terminal events, the joint models provide results comparable to the reduced models, when the longitudinal data have either normal or skew-normal distributions. In the presence of dependent terminal events, the joint models provide more accurate estimates for all parameters than their reduced model counterparts. If the skewness is not correctly accounted for, the joint model gives severely biased estimate for some parameters. In the analysis of the MACS dataset, the proposed joint models have better fit than their reduced model counterparts. We have identified time-varying CD4 count, age, and time as significant risk factors for the longitudinal viral load, while baseline CD4 count and age are associated with the hazard of death. The proposed joint model can be conveniently fit using adaptive Gaussian quadrature tools implemented in SAS procedure NLMIXED and can be easily accessible to, modified and extended by applied researchers.

The joint modeling strategy has several limitations that can be addressed in future research. We

have modeled the covariate effects as linear. Although this assumption simplifies the models, it may not be supported by the data. In our future research, we would like investigate a class of varying-coefficient models (Sun and Wu, 2005) that incorporate the time-dependent covariate effects via penalized splines with a truncated polynomial basis and a fixed number of knots (Ruppert, 2002). Moreover, the Tobit linear mixed model can be extended to a Tobit nonlinear mixed model to account for the nonlinearity of the viral load measurements. Another issue is that the CD4 count measurements may be subject to measurement errors. How to address covariate measurement errors is an interesting research topic in our joint modeling framework.

We have chosen a normal distribution for the random effects because it is flexible in modeling the covariance structure within subject and between the longitudinal and survival processes and it has meaningful interpretation on correlation. In generalized linear mixed models, misspecification of random effects distribution has little impact on the parameters that are not associated with the random effects (Jacqmin-Gadda et al., 2007; Rizopoulos et al., 2008; McCulloch et al., 2011). The impact of random effects misspecification in the proposed modeling framework warrants further investigation. It is of interest to investigate our joint models performance when the underlying random effects distribution is from the more flexible parametric families of asymmetric distributions (e.g., skew-normal/independent distribution (Lachos et al., 2010) and skew-elliptical distribution (Sahu et al., 2003)), which are analytically tractable, accommodate practical values of skewness and kurtosis, and include the skew-normal distribution as a special case. We will also investigate the effect of random effects misspecification and relax the normality assumption by considering Bayesian non-parametric (BNP) framework based on Dirichlet process mixture (Escobar, 1994).

Acknowledgements

Sheng Luo's research was supported in part by the National Institute of Neurological Disorders and Stroke under Award Number R01NS091307 and by the National Center for Advancing Translational Sciences under Award Number KL2-TR000370. The content is solely the responsibility

ACCEPTED MANUSCRIPT

of the authors and does not necessarily represent the official views of the NIH.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.
- Arellano-Valle, R., Branco, M., and Genton, M. (2006). A unified view on skewed distributions arising from selections. *Canadian Journal of Statistics*, 34(4):581–601.
- Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12(2):171–178.
- Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, 32(2):159–188.
- Bandyopadhyay, D., Lachos, V., Castro, L., and Dey, D. (2012). Skew-normal/independent linear mixed models for censored responses with applications to HIV viral loads. *Biometrical Journal*, 54(3):405–425.
- Dagne, G. and Huang, Y. (2012). Bayesian inference for a nonlinear mixed-effects Tobit model with multivariate skew-t distributions: application to AIDS studies. *The International Journal of Biostatistics*, 8(1):1–37.
- Davidian, M. (1995). *Nonlinear Models for Repeated Measurement Data*, volume 62. CRC Press.
- Escobar, M. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277.
- Feng, S., Wolfe, R., and Port, F. (2005). Frailty survival model analysis of the national deceased donor kidney transplant dataset using poisson variance structures. *Journal of the American Statistical Association*, 100(471):728–735.
- Ghosh, P., Branco, M., and Chakraborty, H. (2007). Bivariate random effect model using skew-normal distribution with application to HIV-RNA. *Statistics in Medicine*, 26(6):1255–1267.

- Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480.
- Huang, Y. and Dagne, G. (2011). A Bayesian approach to joint mixed-effects models with a skew-normal distribution and measurement errors in covariates. *Biometrics*, 67(1):260–269.
- Huang, Y. and Dagne, G. (2012). Bayesian semiparametric nonlinear mixed-effects joint models for data with skewness, missing responses, and measurement errors in covariates. *Biometrics*, 68(3):943–953.
- Hughes, J. (1999). Mixed effects models with censored data with application to HIV RNA levels. *Biometrics*, 55(2):625–629.
- Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J.-M., and Thiébaud, R. (2007). Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics & Data Analysis*, 51(10):5142–5154.
- Lachos, V., Ghosh, P., and Arellano-Valle, R. (2010). Likelihood based inference for skew-normal independent linear mixed models. *Statistica Sinica*, 20(1):303–322.
- Lawless, J. and Zhan, M. (1998). Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *Canadian Journal of Statistics*, 26(4):549–565.
- Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random effects model: an example. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(3):325–335.
- Liu, L., Conaway, M., Knaus, W., and Bergin, J. (2008). A random effects four-part model, with application to correlated medical costs. *Computational Statistics & Data Analysis*, 52(9):4458–4473.

- Liu, L. and Huang, X. (2009). Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(1):65–81.
- Liu, L., Strawderman, R., Cowen, M., and Shih, Y. (2010). A flexible two-part random effects model for correlated medical costs. *Journal of Health Economics*, 29(1):110–123.
- Lynn, H. (2001). Maximum likelihood inference for left-censored HIV RNA data. *Statistics in Medicine*, 20(1):33–45.
- McCulloch, C. E., Neuhaus, J. M., et al. (2011). Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical Science*, 26(3):388–402.
- Mellors, J., Munoz, A., Giorgi, J., Margolick, J., Tassoni, C., Gupta, P., Kingsley, L., Todd, J., Saah, A., Detels, R., et al. (1997). Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. *Annals of Internal Medicine*, 126(12):946–954.
- Owen, D. (1956). Tables for computing bivariate normal probabilities. *The Annals of Mathematical Statistics*, 27(4):1075–1090.
- Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2008). Shared parameter models under random effects misspecification. *Biometrika*, 95(1):63–74.
- Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4):735–757.
- Sahu, S., Dey, D., and Branco, M. (2003). A new class of multivariate skew distributions with applications to bayesian regression models. *Canadian Journal of Statistics*, 31(2):129–150.
- Sattar, A., Weissfeld, L., and Molenberghs, G. (2011). Analysis of non-ignorable missing and left-censored longitudinal data using a weighted random effects tobit model. *Statistics in Medicine*, 30(27):3167–3180.

- Schockmel, G., Yerly, S., and Perrin, L. (1997). Detection of low HIV-1 RNA levels in plasma. *Journal of Acquired Immune Deficiency Syndromes*, 14(2):179–183.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.
- Sun, Y. and Wu, H. (2005). Semiparametric time-varying coefficients regression model for longitudinal data. *Scandinavian Journal of Statistics*, 32(1):21–47.
- Thiébaud, R., Guedj, J., Jacqmin-Gadda, H., Chêne, G., Trimoulet, P., Neau, D., and Commenges, D. (2006). Estimation of dynamical model parameters taking into account undetectable marker values. *BMC Medical Research Methodology*, 6(1):38.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: Journal of the Econometric Society*, 26(1):24–36.
- Tsiatis, A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, 14(3):809–834.
- Wulfsohn, M. and Tsiatis, A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339.
- Yu, M., Law, N., Taylor, J., and Sandler, H. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica*, 14(3):835–862.

Table 1: Results of fitting various joint models and reduced models in Setting I when the terminal event is **independent** on the longitudinal data. The longitudinal data have either normal (upper table) or skew-normal (lower table) distributions.

Parameter	JM_N				RM_N			
	Bias	SE	SEM	CP	Bias	SE	SEM	CP
For longitudinal outcome								
$\beta_0 = 8.77$	0.007	0.168	0.165	0.940	0.001	0.163	0.166	0.950
$\beta_1 = -1.8$	-0.002	0.115	0.111	0.946	-0.002	0.113	0.111	0.943
$\beta_2 = 2.91$	-0.008	0.218	0.213	0.937	-0.002	0.213	0.213	0.944
$\beta_3 = -0.34$	-0.000	0.012	0.012	0.952	0.000	0.011	0.011	0.956
$\sigma_u^2 = 3.87$	-0.038	0.362	0.354	0.933	-0.026	0.364	0.355	0.939
$\sigma_\epsilon = 1.51$	-0.006	0.041	0.041	0.943	-0.005	0.041	0.041	0.945
For survival time								
$\gamma_1 = -0.38$	-0.002	0.055	0.055	0.948	-0.002	0.055	0.055	0.948
$\gamma_2 = 0.36$	0.005	0.107	0.106	0.949	0.002	0.107	0.106	0.950
$\nu = 0$	0.001	0.033	0.033	0.953				
Parameter	JM_{SN}				RM_{SN}			
	Bias	SE	SEM	CP	Bias	SE	SEM	CP
For longitudinal outcome								
$\beta_0 = 10.17$	0.107	0.639	0.679	0.926	0.205	0.574	0.576	0.901
$\beta_1 = -1.8$	0.002	0.118	0.118	0.945	-0.002	0.119	0.117	0.944
$\beta_2 = 2.86$	0.002	0.229	0.228	0.945	-0.003	0.229	0.227	0.942
$\beta_3 = -0.34$	-0.001	0.014	0.015	0.951	0.000	0.014	0.014	0.954
$\sigma_u^2 = 3.81$	0.010	0.392	0.402	0.952	-0.016	0.392	0.399	0.951
$\sigma_\epsilon = 2.69$	0.091	0.319	0.335	0.949	0.138	0.310	0.309	0.934
$\lambda = -0.88$	-0.126	0.428	0.489	0.960	-0.199	0.392	0.437	0.945
For survival time								
$\gamma_1 = -0.38$	-0.005	0.055	0.055	0.952	-0.001	0.055	0.055	0.952
$\gamma_2 = 0.34$	0.010	0.108	0.106	0.944	0.005	0.106	0.106	0.952
$\nu = 0$	-0.006	0.037	0.037	0.951				

Table 2: Results of fitting various joint models and reduced models in Setting II when the terminal event is **dependent** on the longitudinal outcome. The longitudinal data have either normal (upper table) or skew-normal (lower table) distributions.

Parameter	JM _N				RM _N			
	Bias	SE	SEM	CP	Bias	SE	SEM	CP
For longitudinal outcome								
$\beta_0 = 8.77$	0.016	0.169	0.167	0.946	-0.012	0.164	0.165	0.956
$\beta_1 = -1.8$	-0.003	0.110	0.112	0.955	0.047	0.111	0.11	0.927
$\beta_2 = 2.91$	-0.016	0.211	0.215	0.952	-0.045	0.210	0.211	0.940
$\beta_3 = -0.34$	-0.001	0.013	0.013	0.952	-0.023	0.013	0.013	0.567
$\sigma_u^2 = 3.87$	-0.027	0.379	0.373	0.935	-0.233	0.353	0.350	0.867
$\sigma_\epsilon = 1.51$	-0.004	0.042	0.043	0.944	0.002	0.044	0.044	0.952
For survival time								
$\gamma_1 = -0.38$	-0.006	0.063	0.063	0.949	0.048	0.057	0.055	0.843
$\gamma_2 = 0.36$	0.001	0.126	0.123	0.940	-0.043	0.109	0.107	0.930
$\nu = 0.3$	0.003	0.040	0.040	0.949				
Parameter	JM _{SN}				RM _{SN}			
	Bias	SE	SEM	CP	Bias	SE	SEM	CP
For longitudinal outcome								
$\beta_0 = 10.17$	0.117	0.620	0.691	0.924	0.123	0.684	0.758	0.912
$\beta_1 = -1.8$	-0.002	0.119	0.118	0.947	0.060	0.118	0.115	0.916
$\beta_2 = 2.86$	0.006	0.231	0.229	0.942	-0.047	0.225	0.225	0.942
$\beta_3 = -0.34$	-0.000	0.015	0.015	0.951	-0.032	0.015	0.015	0.425
$\sigma_u^2 = 3.81$	-0.015	0.405	0.421	0.952	-0.281	0.391	0.394	0.866
$\sigma_\epsilon = 2.69$	0.092	0.325	0.342	0.944	0.107	0.335	0.357	0.933
$\lambda = -0.88$	-0.128	0.416	0.495	0.956	-0.122	0.461	0.532	0.951
For survival time								
$\gamma_1 = -0.38$	-0.004	0.063	0.064	0.953	0.047	0.056	0.055	0.856
$\gamma_2 = 0.34$	0.006	0.125	0.124	0.951	-0.040	0.110	0.107	0.925
$\nu = 0.3$	0.006	0.046	0.047	0.958				

Table 3: Results of fitting joint models JM_{SN} and JM_N in Setting III when the longitudinal data have skew-normal distribution with skewness $\lambda = -1$ and $\sigma_\epsilon = 1$.

Parameter	JM_{SN}				JM_N			
	Bias	SE	SEM	CP	Bias	SE	SEM	CP
For longitudinal outcome								
$\beta_0 = 1$	-0.012	0.264	0.291	0.923	-0.566	0.107	0.108	0.000
$\beta_1 = -0.4$	-0.002	0.072	0.073	0.954	-0.001	0.073	0.072	0.950
$\beta_2 = 2$	-0.002	0.143	0.144	0.948	0.006	0.144	0.144	0.943
$\sigma_u^2 = 2$	-0.009	0.152	0.162	0.955	-0.013	0.153	0.162	0.952
For survival time								
$\gamma_1 = -0.38$	-0.004	0.057	0.059	0.967	-0.003	0.057	0.059	0.964
$\gamma_2 = 0.34$	0.003	0.118	0.115	0.951	0.007	0.120	0.115	0.948
$\nu = 0.3$	0.003	0.047	0.047	0.951	0.004	0.047	0.047	0.949

Table 4: Results of fitting joint models JM_{SN} and JM_N in Setting IV when the longitudinal data have skew-t distribution with $df=5$, skewness $\lambda = 1.5$.

Parameter	JM_{SN}				JM_N			
	Bias	SE	SEM	CP	Bias	SE	SEM	CP
For longitudinal outcome								
$\beta_0 = 1$	-0.983	0.125	0.125	0.000	0.705	0.113	0.114	0.000
$\beta_1 = -2$	-0.019	0.154	0.158	0.954	-0.083	0.163	0.164	0.926
$\beta_2 = 1$	0.006	0.150	0.154	0.953	0.043	0.158	0.161	0.946
$\sigma_u^2 = 2$	0.015	0.190	0.196	0.955	0.155	0.220	0.209	0.896
For survival time								
$\gamma_1 = -0.38$	-0.003	0.118	0.118	0.950	-0.006	0.118	0.118	0.952
$\gamma_2 = 0.34$	0.003	0.116	0.117	0.942	0.006	0.114	0.117	0.952
$\nu = 0.3$	-0.002	0.051	0.050	0.946	-0.010	0.051	0.050	0.936

Table 5: Model comparison statistics for the MACS dataset. -2LogLik : -2 times log likelihood; AIC: Akaike information criterion; BIC: Bayesian information criterion. Boldface indicates the preferred model.

Distribution	Baseline hazard	Joint models			Reduced models		
		-2LogLik	AIC	BIC	-2LogLik	AIC	BIC
Normal	Weibull	59046.09	59076.09	59154.08	59129.39	59157.39	59244.46
	Piecewise	58756.66	58802.66	58922.26	58837.70	58881.70	59029.39
Skew normal	Weibull	59037.56	59069.56	59152.75	59111.06	59141.06	59233.33
	Piecewise	58746.58	58794.58	58919.37	58819.37	58865.37	59018.26

Table 6: The means, standard errors (SE), p values, and 95% confidence interval (CI) of the parameters from the proposed joint model JM_{SN} and the reduced model RM_N , both assuming a piecewise constant baseline hazard function.

	JM_{SN}				RM_N			
	MLE	SE	P	95% CI	MLE	SE	P	95% CI
Log Viral Load								
Int	18.538	0.235	< 0.001	18.078, 18.999	17.323	0.204	< 0.001	16.923, 17.723
$\sqrt{CD4}$	-0.395	0.006	< 0.001	-0.407, -0.382	-0.406	0.006	< 0.001	-0.418, -0.394
Ethnicity:white	-0.227	0.166	0.172	-0.552, 0.099	-0.260	0.182	0.153	-0.617, 0.097
Age	-0.370	0.065	< 0.001	-0.497, -0.243	-0.379	0.065	< 0.001	-0.505, -0.252
Time	-0.325	0.005	< 0.001	-0.334, -0.315	-0.344	0.004	< 0.001	-0.353, -0.336
σ_u^2	2.194	0.112	< 0.001	1.976, 2.413	2.220	0.112	< 0.001	2.000, 2.441
σ_ϵ	2.673	0.097	< 0.001	2.483, 2.862	2.209	0.016	< 0.001	2.179, 2.240
λ	-0.962	0.134	< 0.001	-1.225, -0.698				
Time to Death								
$\sqrt{BCD4}$	-0.149	0.011	< 0.001	-0.171, -0.127	-0.111	0.010	< 0.001	-0.131, -0.092
Ethnicity:white	0.039	0.127	0.760	-0.210, 0.288	-0.005	0.120	0.968	-0.240, 0.230
Age	0.151	0.047	0.001	0.058, 0.244	0.134	0.045	0.003	0.047, 0.222
γ	0.232	0.029	< 0.001	0.176, 0.289				

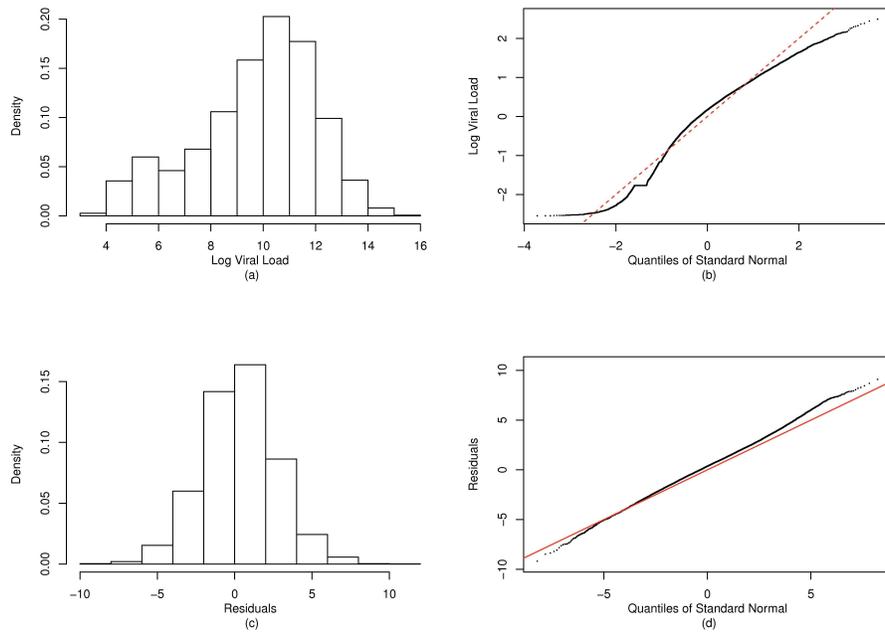


Figure 1: Density histogram and corresponding Q-Q plots for the viral load measurements (in log scale; panels a and b), and the model residuals (panels c and d) obtained after fitting a Tobit model.

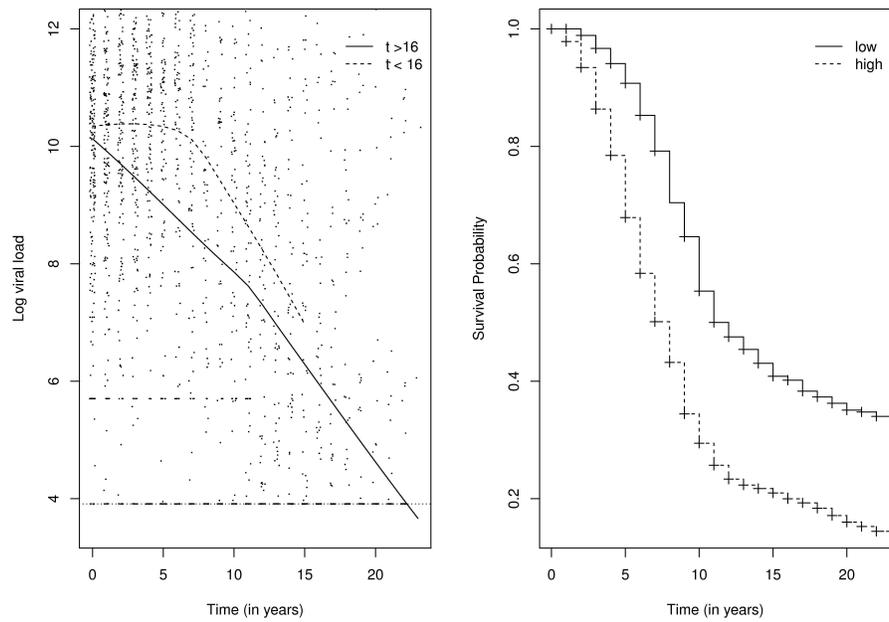


Figure 2: Mean viral load measurements (in log scale, left panel, censored at $\log 50 = 3.9$) and Kaplan-Meier curves displaying difference in time to death for participants with high (median log viral load > 9.38 , dashed line) and low (median log viral load ≤ 9.38 , solid line) viral load measurements.