Published in final edited form as:

Stat Med. 2014 November 20; 33(26): 4560–4576. doi:10.1002/sim.6242.

Joint Modeling of Longitudinal and Survival Data with Missing and Left-Censored Time-Varying Covariates

Qingxia Chen[a], **Ryan C. May**[b], **Joseph G. Ibrahim**[b,*], **Haitao Chu**[c], and **Stephen R. Cole**[d]

[a]Departments of Biostatistics and Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, 37232, U.S.A

[b]The EMMES Corporation, Rockville, Maryland, 20850, U.S.A

[c]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A

[d]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A

Abstract

We propose a joint model for longitudinal and survival data with time-varying covariates subject to detection limits and intermittent missingness at random (MAR). The model is motivated by data from the Multicenter AIDS Cohort Study (MACS), in which HIV+ subjects have viral load and CD4 cell count measured at repeated visits along with survival data. We model the longitudinal component using a normal linear mixed model, modeling the trajectory of CD4 cell count by regressing on viral load and other covariates. The viral load data are subject to both left-censoring due to detection limits (17%) and intermittent missingness (27%). The survival component of the joint model is a Cox model with time-dependent covariates for death due to AIDS. The longitudinal and survival models are linked using the trajectory function of the linear mixed model. A Bayesian analysis is conducted on the MACS data using the proposed joint model. The proposed method is shown to improve the precision of estimates when compared to alternative methods.

Keywords

Detection Limit; Joint Modeling; Missing Data; Multicenter AIDS Cohort Study

1. Introduction

In many longitudinal studies, time to event data are recorded in addition to longitudinal and baseline covariates. In such studies, interest often lies in understanding the relationships between the longitudinal history of a process and its effect on the risk of an event. For analysis of this type of data, a class of models called *joint models* has been developed, which

[*]Correspondence to: Joseph G. Ibrahim, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A. ibrahim@bios.unc.edu.

jointly model both components simultaneously. Joint models have been used extensively in studies of subjects with Human Immunodeficiency Virus (HIV, [1], [2], etc.) because they can reduce the bias and improve the efficiency of the estimates ([3], [4]). As with any large dataset, and particularly in the case of longitudinal data, it is often the case that a high degree of covariate and response data are missing. Additionally, in an HIV positive individual, the measurement of viral load (the amount of virus in the blood) is only accurate down to a particular limit of detection (LD), which is left-censored. Values below the limit of detection cannot be reliably quantified or distinguished from a “blank” blood sample with no virus. In many cases ([1], etc.) any missing covariate data are omitted from the analysis, and estimation proceeds on the complete data, which is called a complete case analysis (CC). However, when a high degree of covariate data are missing, a great deal of information is lost in a CC analysis; and when the covariates are missing nonignorably and the missingness mechanism depends on the outcome variables, the CC analysis is invalid ([5], Section 3.2). In other words, when the covariate is subject to a detection limit (nonignorable missingness) but the censoring probability does not depend on the response variable, the CC analysis is not subject to bias but efficiency is lost, assuming a correct regression function is specified. However, if the censoring probability depends on the response variable after conditioning on the covariates, the CC estimates of the regression coefficients are biased ([5], [6]).

This article aims to develop a joint modeling strategy that accounts for both intermittently missing and left-censored time-varying covariates. The longitudinal data are intermittently missing if a missing value is followed by an observed value. In other words, the data are non-monotonically missing. This analysis is motivated by data from the Multicenter AIDS Cohort Study (MACS, [7]), a prospective study of disease progression in participants infected with, or at risk for infection with, HIV. The subset of MACS participants who seroconvert with HIV while under observation are followed from the date of HIV seroconversion, with many variables including CD4 cell count and viral load measured at planned study visits every 6 months. Interest lies in the progression of CD4 cell count and viral load from seroconversion with HIV, and their impact on survival. In this paper we are concerned with the effect of calendar period (as a proxy for HIV treatment) with survival. In particular, we assume that HIV treatment (and HIV viral load) affect CD4 cell count, the primary immunologic marker of HIV disease progression, which in turn affects survival. We posit joint models that (a) relate the calendar period and viral load to CD4 cell count, and (b) relate the modeled CD4 cell count to survival. From these joint models we aim to estimate the effect of calendar period on survival mediated through the modeled CD4 cell count. We note that any direct effect of calendar period on survival, not mediated through CD4 cell count, would not be recovered here; but such direct effects are expected to be minor in comparison to the CD4-mediated effect. Of the available viral load data, 27.1% were missing and 16.9% fell below a limit of detection. Using a Bayesian analysis, we model the progression of CD4 cell count over time, while accounting for the missingness and left-censoring on the available viral load data. The Bayesian approach allows us to fully use the observed data and account for the missing data under the MAR assumption, which is superior to the MCAR assumption that is required for a less efficient complete-case analysis.

There is extensive literature on missing data in longitudinal studies ([8], [9], [10], [11], etc.) as well as on joint modeling of longitudinal and survival data ([12], [13], [14], etc.). Wu et al. [15] review joint modeling with comprehensive references. The literature on detection limits in longitudinal studies focuses largely on the scenario with the response variable subject to left censoring, while time-varying covariates subject to left censoring are considered in this paper. Furthermore, the literature on LDs confines attention to the mixed effects model for the longitudinal component alone, but rarely considers joint models for the longitudinal and survival data simultaneously ([16], [17], [18], [19]). Recently, Wu et al. [20] investigated joint modeling in an AIDS clinical trial with informative dropout by incorporating a missing data mechanism into the joint model likelihood, and proposed an EM algorithm for the estimation procedure within the likelihood framework. Thiébaut et al. [21] considered a bivariate linear mixed model for two biomarkers, with one biomarker (plasma HIV RNA) subject to left censoring, and a log-normal survival model for the time to drop out. The model parameters were estimated by a direct maximum likelihood approach. In this paper, we propose a Bayesian approach for a longitudinal study with censored and missing time-varying covariate data within the joint modeling framework. Using data from MACS, the goal of this paper is to jointly model the longitudinal disease progression and failure from the disease in study participants while accounting for both intermittent missingness and a limit of detection on a single covariate. The differences of this paper from the existing literature include that (a) the time-varying covariates are subject to missingness and a detection limit; (b) the method is developed under a joint modeling framework; and (c) the proposed Bayesian approach essentially treats the missing and left-censored covariate values as extra parameters from a computational perspective, and therefore, it is able to account for the missing and left-censored data without resorting to asymptotics or numerical maximization.

The rest of this article is organized as follows. In section 2 we give a review of joint models, and develop notation. In section 3 we develop a Bayesian approach to this problem and apply this approach to the MACS data. We compare our results with those obtained from ad-hoc estimation approaches. We conclude the article in section 4 with a discussion.

2. Preliminaries

2.1. The Longitudinal Model

Of the two submodels included in a joint model, the longitudinal component is less complicated with a model formulation very similar (if not identical) to that of a model fit for the longitudinal data alone. The dataset consists of N subjects with n_i measurements recorded for subject i, ($i = 1, \dots, N$). The response y_{ij} ($j = 1, \dots, n_i$), fixed-effect covariate vector $\mathbf{x}_{ij} = (x_{1ij}, \dots, x_{pij})'$, and random-effect covariate vector $\mathbf{z}_{ij} = (z_{1ij}, \dots, z_{qij})'$ are recorded at times t_{ij}. The longitudinal model is usually specified as a linear mixed effects model [22]

$$y_{ij}=\boldsymbol{\beta}'\mathbf{x}_{ij}+\mathbf{b_i}'\mathbf{z}_{ij}+\varepsilon_{ij}=\psi_{ij}(\boldsymbol{\beta},\mathbf{b}_i)+\varepsilon_{ij}, \quad (1)$$

where β is the $p \times 1$ fixed-effect parameter vector, and $\mathbf{b}_i$ is the $q \times 1$ vector of random effects for subject i with $\mathbf{b}_i \sim N_q(\mathbf{0}, \Sigma_b)$. The error vector $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})'$ is usually specified as $\varepsilon_i \sim N_{n_i}(0, \xi^{-1}\mathbf{I}_{n_i})$, where $\mathbf{I}_{n_i}$ represents the identity matrix of dimension n_i. The trajectory function for the model is defined as $\psi_{ij}(\beta, \mathbf{b}_i) = \beta'\mathbf{x}_{ij} + \mathbf{b_i z}_{ij}$. More generally, (1) can be written in terms of $y_i(t)$, the response at any time t. Taking $\mathbf{x}_i(t) = (x_{i1}(t), \ldots, x_{ip}(t))'$ and $\mathbf{z}_i(t) = (z_{i1}(t), \ldots, z_{iq}(t))$ to represent the fixed-effects and random-effects covariate vectors at time t respectively, the model can be rewritten as

$$y_i(t) = \boldsymbol{\beta}'\mathbf{x}_i(t) + \mathbf{b}_i'\mathbf{z}_i(t) + \varepsilon_i(t) = \psi_i(\boldsymbol{\beta}, \mathbf{b}_i, t) + \varepsilon_i(t), \quad (2)$$

where the error term $\varepsilon_i(t) \sim N(0, \xi^{-1})$, and the trajectory function $\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, t) = \boldsymbol{\beta}'\mathbf{x}_i(t) + \mathbf{b}_i'\mathbf{z}_i(t)$. In many AIDS studies that use joint models, the longitudinal component uses random effects with functions of time only [23]. The form of the random effect covariate vector $\mathbf{z}_i(t)$ is usually simple, including only random intercept and random slope effects, or at most a random quadratic effect of time. In this case, the trajectory can be specified at generic time t, as

$$\psi_i(\boldsymbol{\beta}_{\mathbf{b}}, \mathbf{b}_i, t) = \boldsymbol{\beta}_{\mathbf{b}}{}'\mathbf{x}_{\mathbf{b}_i}(t) + \mathbf{b}_i'\mathbf{z}_i(t), \quad (3)$$

where β_b and $\mathbf{x_b}$ are the parameters and design matrix of the fixed effects that are corresponding to the random effects. It should be noted that many authors have considered a more complex version of (3), involving an additional mean-zero stochastic process that does not depend on $\mathbf{z}_i(t)$ or $\mathbf{b}_i$. This form allows within-subject autocorrelation that accounts for fluctuations from the hypothesized "smooth" trajectory function included in the model ([23]). This extended form is not considered in the proposed modeling approach presented in section 3.

2.2. The Survival Model

The second submodel in a joint model is the survival model. This is usually taken as Cox model with time-dependent covariates [24] with hazard function $\lambda_i(t)$ for subject i at time t. The survival component of the joint model includes a link or connection to the longitudinal submodel, the unique characteristic that makes the model "joint". The link in this case is the inclusion of a portion (or all) of the longitudinal trajectory $\psi_i(\beta, \mathbf{b}_i, t)$ as a covariate within the survival model. The survival component is therefore expressed as

$$\lambda_i(t) = \lambda_0(t)\exp\left\{\theta h(\boldsymbol{\beta}, \mathbf{b}_i, t) + \boldsymbol{\beta}_s'\mathbf{x}_{si}(t)\right\}. \quad (4)$$

Here, $h(\beta, \mathbf{b}_i, t)$ is a function of the fixed effects and random effects in the longitudinal model, with θ a scalar parameter that links the two submodels. The survival covariate vector $\mathbf{x}_{si}(t) = (x_{si1}, \ldots, x_{sir})'$ usually includes baseline covariates for subject i, with β_s representing the $r \times 1$ parameter vector for these baseline covariates. The baseline hazard function is given by $\lambda_0(t)$. The form that $h(\beta, \mathbf{b}_i, t)$ takes determines the type of joint model that is fit. In a trajectory model (TM), the longitudinal trajectory is included in the survival component

([25, 26, 1, 27, 12, 3]). There are different approaches to construct the trajectory function. Some include the mean response composed by the fixed effects model only, and others use the full longitudinal trajectory with both fixed effects and random effects components. For example, we can have $h(\boldsymbol{\beta}, \mathbf{b}_i, t)=\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, t)=\boldsymbol{\beta}'\mathbf{x}_i(t)+\mathbf{b}_i'\mathbf{z}_i(t)$. In a shared parameter model (SPM), only the random effects from the longitudinal model are included instead of the full trajectory ([28, 29, 30]). One example is to take $h(\boldsymbol{\beta}, \mathbf{b}_i, t)=\mathbf{b}_i'\mathbf{z}_i(t)$, such that only the random effects are included in the survival component, or even $h(\beta, \mathbf{b}_i, t) = \mathbf{b}_{ik}$ if only the kth random effect is considered. In general, the difference between TM and SPM is that TM includes at least the fixed effects to represent the mean response (it could include the random effects as well but not necessary) and SPM includes only the random effects. An excellent general review on joint modeling of longitudinal and survival data was given in [23]. The parameter β_s in (4) is a parameter vector for covariates unique to the survival submodel and the $\mathbf{x_{si}}$ are additional covariates that are associated with the survival outcome but not with the longitudinal measurements.

2.3. Likelihood for Joint Model

With both the longitudinal and survival submodels specified, we now combine the two to form the likelihood for the full joint model. In this case, we will specify the joint model using a TM in the survival component, such that the full longitudinal trajectory is included in the hazard function. We take T_i to represent the potential failure time for subject i, and C_i to represent the potential censoring time for subject i. We define $S_i = \min(T_i, C_i)$ as the observed failure/censoring time for subject i, with δ_i taken as an indicator for observing failure, with $\delta_i = 1$ when $T_i < C_i$, and $\delta_i = 0$ otherwise. We define $\psi_i(\beta, \mathbf{b}_i, t)$ as the value of the longitudinal trajectory for subject i at time t and $\psi_{ij}(\beta, \mathbf{b}_i)$ as the value of the longitudinal trajectory for subject i at visit j. Let $f(\cdot)$ represent a generic density function. When the covariates are completely observed without censoring or missingness, the likelihood of the joint distribution of the observed data and random effects from the ith subject can be written as

$$\begin{gathered} L_i \propto f_i(\text{Survival}|\text{Longitudinal}) \times f_i(\text{Longitudinal}) \\ = f(S_i|\theta, \delta_i, \boldsymbol{\beta}_s, \psi_i(\boldsymbol{\beta}, \mathbf{b}_i, t), \mathbf{x}_{si}) \times f(\mathbf{y_i}|\mathbf{x_i}, \mathbf{z_i}, \boldsymbol{\beta}, \mathbf{b}_i) f(\mathbf{b}_i) \\ = \left[\left\{\lambda_0(S_i)\exp\left(\theta\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, S_i)+\boldsymbol{\beta}_s'\mathbf{x}_{si}(S_i)\right)\right\}^{\delta_i} \times \exp\left\{-\int_0^{S_i}\lambda_0(u)\exp(\theta\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, u)+\boldsymbol{\beta}_s'\mathbf{x}_{si}(u))du\right\}\right] \\ \times \left[\left\{\frac{\xi}{(2\pi)}\right\}^{n_i/2}\exp\left\{-\frac{\xi}{2}\sum_{j=1}^{n_i}(y_{ij}-\psi_{ij}(\boldsymbol{\beta}, \mathbf{b}_i))^2\right\}\right] f(\mathbf{b}_i) \end{gathered} \tag{5}$$

and the likelihood for all subjects is $L=\prod_{i=1}^{N} L_i$. When the covariates are missing and/or censored, additional models for those covariates are needed and this will be discussed in Section 3.

2.4. Fitting the Model

Estimation of a joint model may be performed in at least two ways. The first estimation approach is to use the EM algorithm. This approach has been used often in past analyses of AIDS data ([1], [2]). The R package JM [31] was recently released and fits shared parameter models using the EM algorithm. A second approach to estimation uses a Bayesian framework, fitting the model with Markov Chain Monte Carlo (MCMC) methods. This approach is discussed in detail in Ibrahim, Chen, and Sinha ([32], chap.7), and has been used by many authors ([33], [34], [35], etc.). Guo and Carlin [36] provide WinBUGS software for fitting joint models using a Bayesian framework. For the analysis presented in this paper, a Bayesian framework similar in flavor to that of Ibrahim, Chen, and Sinha [32] is used, with all computations being carried out in R [37].

3. MACS Data Analysis

3.1. Background

The motivating data for this paper comes from the Multicenter AIDS Cohort study (MACS), a prospective study of disease progression in participants infected with, or at risk for infection with, HIV. The data collected by the MACS study is of particular interest because participants are followed from the time of seroconversion, when they first develop antibodies to HIV (as a response to contracting the virus). The study population includes participants who contracted HIV during the study follow-up (1986–2005). Participants in the study were seen at semiannual visits, where demographic information was recorded along with laboratory measurements including viral load and CD4 cell count. Survival data for each participant was also recorded, specifically for deaths attributable to AIDS. Of the 470 subjects in the study who seroconverted with HIV during follow-up, 443 were observed at 3 or more visit times, and were included in the study analysis. Of the 443 subjects, 165 (37.2%) died due to AIDS during the study period and had the time of death recorded.

In studies of HIV progression, interest lies in the relationship between CD4 cell count and viral load measurements over time. CD4 cell count is a measure of immune system strength, while viral load is a measure of the amount of circulating virus. These two biomarkers are inversely correlated, as high levels of virus (viral load) indicate low immune system strength (CD4 cell count). In this paper, we are concerned with the effect of calendar period (as a proxy for HIV treatment) with survival. In particular, we assume that HIV treatment (and HIV viral load) affect CD4 cell count, the primary immunologic marker of HIV disease progression, which in turn affects survival. A complication that often arises in HIV studies is that viral load values are subject to a lower limit of detection. Values of viral load falling below this limit are unable to be detected by laboratory tests. In long-term longitudinal studies such as MACS, it is common for the limit of detection to change over time, as newer technology is able to detect even lower levels of viral load. In total, 16.9% of the available viral load data fell below the limit of detection. Additionally, 27.1% of the viral load data were missing intermittently in the dataset. A trajectory plot for CD4 cell count and viral load since seroconversion for a random sample of 50 participants in the study is given in figure 1. In the top panel, the solid lines represent the measured CD4 cell count trajectories of individual subjects and dotted lines connect the CD4 cell count measurements at the latest visit and at the death time assuming the last observation carried forward. Note that avoiding making the last observation carried forward assumption is one of the reasons to implement a joint modeling analysis. A Cox model using CD4 cell count as a time-varying covariate is also not appropriate since a step function or some interpolations would have to be assumed for the CD4 cell count measurement. Another reason of considering joint model for the analysis conducted here is because the CD4 cell count tends to be measured with error and thus modeling the CD4 count by a mixed model accounts for this measurement error. The viral load trajectory is plotted in the bottom panel using similar legends with additional LD information.

One additional limitation to the public-use MACS data was that time of visit for each participant was only available at the year level, no month or day dates are supplied. For a participant with multiple visits in the same year, the available data only lists the year of the visit and the chronological ordering of multiple visits in the same year (i.e., that one visit precedes another). To account for this limitation, exact visit dates were imputed for each subject in the dataset. For a subject with two visits in year X, the time of the first visit was imputed at X + 0.25, with visit 2 at X + 0.75. For a subject with 3 visits, time was imputed as X + 0.17, X + 0.5, and X + 0.83. The time of HIV seroconversion was imputed as the midpoint between times of the first visit where HIV antibodies were detected and the visit immediately preceding this visit. For a particular subject's data to be included in the analysis, baseline covariates for race and age at seroconversion needed to be recorded. For this analysis to be valid, we assume that the probability of missingness for the covariates of race and age does not depend on the longitudinal outcome variables (CD4 and VL) and survival outcome. This assumption is not testable but is likely to be true in the MACS study because intervention was not introduced at baseline. Additionally, at least CD4 cell count or viral load measurement needed to be recorded for a patient to be included in the analysis.

3.2. Joint Model for MACS Analysis

To account for both the longitudinal trajectory of CD4 cell count and survival, a joint model was specified for the analysis. The observed CD4 cell count is subject to measurement errors and hence need to be modeled. We use a joint model rather than a two-stage approach to properly account for the variability yielded from the longitudinal model in the survival model. In particular, the longitudinal component of the model is specified as a mixed-effects model, with a random slope and intercept for each subject. Both CD4 and viral load were $\log_{10}$ transformed, with CD4_{ij} and VL_{ij} representing the $\log_{10}$ transformed values of CD4 and viral load for subject i and visit j, occurring at time Time_{ij}. Additionally, a covariate was included to account for the indirect effect of Highly Active Antiretroviral Treatment therapy (HAART), an HIV treatment that consists of several antiretroviral drugs being taken concurrently. HAART treatment has had a dramatic positive effect on the survival of HIV ([38], [39]). Though records for HAART treatment are available in the MACS public-use data, we instead use HAART calendar period as an instrumental variable [40] for HAART. This approach is similar to those in past HIV studies ([41],[42]), allowing us to circumvent potential bias in results due to residual confounding by indication that could occur if we used the direct HAART variable. We define the HAART calendar period as all visit times occurring after January 1, 1998. We define covariate PD_{ij} as an indicator for the HAART calendar period, such that $\text{PD}_{ij} = 1$ if $\text{Time}_{ij} > 1/1/98$, and $\text{PD}_{ij} = 0$ otherwise. The final longitudinal model is specified as

$$\mathrm{CD4}_{ij}=\beta_1\mathrm{VL}_{ij}+\beta_2\mathrm{PD}_{ij}+\beta_{b0}+\beta_{b1}t_{ij}+b_{i0}+b_{i1}t_{ij}+\varepsilon_{ij}, \quad (6)$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_{b0}, \beta_{b1})'$ is the vector of parameters for the fixed effect covariates and t_{ij} is the time since seroconversion. Viral load is included to better explain the variability in the CD4 cell count. Following standard estimation approaches for a linear mixed-effects model, the joint distribution of the random effects $\mathbf{b}_i = (b_{i0}, b_{i1})'$ was again assumed bivariate normal, with mean $\mathbf{0}$ and covariance matrix Σ_b. The error term ε_{ij} is assumed to have a normal distribution with $\varepsilon_{ij} \sim N\,(0, \xi^{-}1)$.

A TM was chosen for the analysis such that the full longitudinal trajectory was included in the survival model. This trajectory is specified as $\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, t, \mathrm{VL}_i(t)) = \beta_1\mathrm{VL}_i(t) + \beta_2\mathrm{PD}_i(t) + \beta_{b0} + \beta_{b1}t + b_{i0} + b_{i1}t$, where $\mathrm{VL}_i(t)$ and $\mathrm{PD}_i(t)$ represent their respective values at time t. We also denote $\psi_{ij}(\beta, \mathbf{b}_i, \mathrm{VL}_{ij}) = \beta_1\mathrm{VL}_{ij} + \beta_2\mathrm{PD}_{ij} + \beta_{b0} + \beta_{b1}t_{ij} + b_{i0} + b_{i1}t_{ij}$, the value of the longitudinal trajectory for subject i at visit j. Other baseline covariates of interest included the age at which a subject contracted HIV (AGE_i), and race (RACE_i), with $\mathrm{RACE}_i = 1$ if subject i is white, and $\mathrm{RACE}_i = 0$ otherwise. We again define θ as the parameter linking the longitudinal and survival submodels, with $\boldsymbol{\beta}_s = (\beta_{s1}, \beta_{s2})'$ as the parameters corresponding to the baseline covariates. The Cox submodel with time-dependent covariates is specified below, with $\lambda(t)$ and $\lambda_0(t)$ representing the hazard and baseline hazard functions at time t, respectively, and is given by

$$\lambda[t|\psi_i(\boldsymbol{\beta},\mathbf{b}_i,t,\mathrm{VL}_i(t))]=\lambda_0(t)\exp[\theta\psi_i(\boldsymbol{\beta},\mathbf{b}_i,t,\mathrm{VL}_i(t))+\beta_{s1}\mathrm{RACE}_i+\beta_{s2}\mathrm{AGE}_i]. \quad (7)$$

We did not include direct viral load effect in the survival model because while there are almost certainly direct effects of viral load on survival not mediated through CD4 cell count, especially for non-AIDS related mortality, these effects are expected to be small relative to the CD4-mediated effects.

Viral load data were measured longitudinally in the MACS data with potential correlation within the same patients. To account for this correlation, a sequence of conditional densities was used to model the joint density of the viral load data as

$$f(\mathbf{VL}_i|\mathrm{RACE}_i,\mathrm{AGE}_i,t_i)=f(\mathrm{VL}_{i1}|\mathrm{RACE}_i,\mathrm{AGE}_i,t_{i1})\prod_{j=2}^{n_i}f(\mathrm{VL}_{ij}|\mathrm{VL}_{i,j-1},\mathrm{RACE}_i,\mathrm{AGE}_i,t_{ij}), \quad (8)$$

where $\mathbf{VL}_i = (\mathrm{VL}_{i1}, \ldots, \mathrm{VL}_{in_i})$. To do this, a simple linear regression model was specified for the baseline viral load. A multiple linear regression model adjusting for the previous viral load measurement, age, race, and the time since seroconversion was specified for viral load after baseline, where the baseline viral load measurement, VL_{i1}, is defined as the viral measurement at the first visit after seroconversion with HIV. In particular, we have $\mathrm{VL}_{i1} \sim N\,(\mu_v, \tau^{-1})$ and

$$\mathrm{VL}_{ij}=\alpha_0+\alpha_1\mathrm{VL}_{i,j-1}+\alpha_2\mathrm{RACE}_i+\alpha_3\mathrm{AGE}_i+\alpha_4t_{ij}+\varepsilon^*_{ij}, j=2,\ldots,n_i, \quad (9)$$

where $\boldsymbol{\alpha}$ = (α_0, α_1, α_2, α_3, α_4) is the vector of unknown parameters and $\varepsilon^*_{ij} \sim N(0, \eta_v^{-1})$. The VL model considered here is similar to, although not exactly the same as, the random intercept model. We used this model instead of the random intercept and slope model used for CD4 cell count because (a) as shown in Figure 1, compared to CD4 cell count, the VL measure has larger variability in the intercept than in the slope; (b) compared to the current model, which is already computational intensive, the random intercept and slope model requires even more computational effort and the MCMC algorithm tends to be less stable.

For a viral load observation VL$_{ij}$ falling below a limit of detection LD$_{ij}$, the prior distribution is truncated at LD$_{ij}$, taking a nonzero density only below LD$_{ij}$. For viral load observations that are missing, no such truncation is needed. The missing viral load values are assumed to be missing at random, as parameters involving viral load are distinct from the others in the model. Because of this assumption, the complete-data likelihood that now accounts for the missing and left-censored viral load will be appended to (5) to include the viral load covariate distribution. We will again denote **S** = (S_1, …, S_N) as the vector of observed failure/censoring times for each subject, with $\boldsymbol{\delta}$ = (δ_1, …, δ_N) taken as the vector of failure time indicators (with δ_i = 1 if observed failure and 0 otherwise). Therefore, the likelihood of the joint distribution of the observed data, random effects and viral load values VL can be expressed as follows:

$$\begin{aligned} L &= f(\text{Survival}|\text{Longitudinal}) \times f(\text{Longitudinal}) \\ &= f(\mathbf{S}|\boldsymbol{\delta}, \theta, \boldsymbol{\beta_s}, \psi(\boldsymbol{\beta}, \mathbf{b}, t, \mathbf{VL})) \times f(\mathbf{CD4}, \mathbf{b}, \mathbf{VL}|\boldsymbol{\beta}) \\ &= f(\mathbf{S}|\boldsymbol{\delta}, \theta, \boldsymbol{\beta_s}, \psi(\boldsymbol{\beta}, \mathbf{b}, t, \mathbf{VL})) \times [f(\mathbf{CD4}|\boldsymbol{\beta}, \mathbf{b}, \xi, \mathbf{VL}) f(\mathbf{b}|\boldsymbol{\mu}_b, \textstyle\sum_b) f(\mathbf{VL}|\mu_v, \tau, \boldsymbol{\alpha}, \eta_v)]. \end{aligned} \tag{10}$$

Figure 2 shows the underlying diagram for joint modeling strategy. Thus, the full formula for the complete-data likelihood for subject *i* is given by

$$\begin{aligned} L_i \propto & \Big[\{\lambda_0(S_i)\exp(\theta\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, S_i, \text{VL}_i(S_i)) + \beta_{s1}\text{RACE}_i + \beta_{s2}\text{AGE}_i)\}^{\delta_i} \\ & \times \exp\left\{ -\textstyle\int_0^{S_i} \lambda_0(u)\exp(\theta\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, u, \text{VL}_i(u)) + \beta_{s1}\text{RACE}_i + \beta_{s2}\text{AGE}_i) du \right\} \Big] \\ & \times \left[\xi^{n_i/2} \exp\left\{ -\frac{\xi}{2} \sum_{j=1}^{n_i} (\text{CD4}_{ij} - \psi_{ij}(\boldsymbol{\beta}, \mathbf{b}_i, \text{VL}_{ij}))^2 \right\} \right. \\ & \times \left|\textstyle\sum_{\mathbf{b}}^{-\mathbf{1}}\right|^{1/2} \exp\left\{ -\frac{1}{2} \mathbf{b}_i \textstyle\sum_{\mathbf{b}}^{-\mathbf{1}} \mathbf{b}_i \right\} \\ & \times \eta_v^{(n_i-1)/2} \exp\left\{ -\frac{\eta_v}{2} \sum_{j=2}^{n_i} (\text{VL}_{ij} - \alpha_0 - \alpha_1 \text{VL}_{i,j-1} - \alpha_2 \text{RACE}_i - \alpha_3 \text{AGE}_i - \alpha_4 t_{ij})^2 \right\} \\ & \left. \times \tau^{1/2} \exp\left\{ -\frac{\tau}{2} (\text{VL}_{i1} - \mu_v)^2 \right\} \right]. \end{aligned} \tag{11}$$

Note that the $VL_i(t)$ used in the survival function component of the likelihood is the predicted value from model (9) by plugging the covariates and varying *t* values. Since the random effects **b** are not observed and the viral load measurements VL are subject to missingness and left censoring, the observed-data likelihood requires integration of equation (11) over **b** and VL. The proposed fully Bayesian approach will treat **b** and the missing/left-censored VL values as extra “parameters” to be sampled in the Markov Chain Monte Carlo (MCMC) algorithm. Following Ibrahim, Chen, and Sinha ([32], section 7.3), independent uniform improper priors were taken for $\boldsymbol{\beta}$, $\boldsymbol{\beta_s}$, and $\boldsymbol{\alpha}$, with $\pi(\boldsymbol{\beta}) \propto \mathbf{1}$, $\pi(\boldsymbol{\beta_s}) \propto \mathbf{1}$, and $\pi(\boldsymbol{\alpha}) \propto \mathbf{1}$.

Additional priors are specified as follows: $\xi \sim Gamma(10^{-3}, 10^{-3})$, $\mu_v \sim N(0, 10^5)$, $\eta_v \sim Gamma(10^{-3}, 10^{-3})$, $\tau \sim Gamma(10^{-3}, 10^{-3})$, $\sum_b^{-1} \sim \mathrm{Wishart}(3, 10^6\mathbf{I})$. The baseline hazard function $\lambda_0(t)$ was specified as having the form of a piecewise constant hazard, taking the constant value λ_k for each of the $k = 1, \ldots, 10$ time intervals $(s_k, s_{k+1}]$ that span the range of the observed times t_{ij}. Computation of $\exp\left\{-\int_0^{S_i} \lambda_0(u) \exp(\ldots) du\right\}$ was then performed using the approximation given in Ibrahim, Chen, and Sinha ([32], p. 277–278). It can be shown that with this choice of priors, the joint posterior distribution is proper.

Gibbs sampling was performed by sampling from the full conditional distribution for each model parameter. A derivation of the conditionals is given in the Appendix. For parameters with a closed-form full conditional distribution, sampling is straightforward. For parameters with no closed-form full conditional distributions, sampling was performed using the Adaptive Rejection Metropolis Sampling (ARMS) of Gilks, Best, and Tan [43]. Estimation was performed using Gibbs samples from 10,000 iterations, with a burn-in of 1000 iterations. For comparison, several simpler models were also applied to the MACS data. First, a two-stage model was fit, in which each of the two submodels was fit separately. In the first stage, the longitudinal submodel in equation (6) was fit independently of the survival component. The fitted trajectory from the longitudinal component was then fixed, and was included as a covariate in the survival model in equation (7). The second stage fitted this survival model, giving parameter estimates for the survival component only. Such a model is computationally simpler because the likelihood functions for each model are separate, and are not combined as in equation (11). Additionally, a joint model was also fit to only 56% of the total observations with fully observed values of viral load (complete-case analysis). A joint model was also fit in which substituted values of viral load were used for all left-censored viral load values. For a viral load measurement falling below the limit of detection LD_{ij}, the common substitution of $\mathrm{LD}_{ij}/\sqrt{2}$ was used as the "true" viral load value at the specified visit. This substitution analysis was then performed on 72.9% (56% observed + 16.9% substituted) of the total observations after removing 27.1% observations with missing viral load values. For all the analyses, the values of VL and age at seroconversion were normalized to achieve MCMC convergence. In particular, the logarithmic transformed viral load values were normalized as (VL − 3.8)=1.3, and age at seroconversion was normalized as (AGE − 35.3)=8.3. The simulation results from the two-stage, complete-case, substitution, and full joint models are given Table 1. Posterior estimates are taken from the 9000 sampled values. Figure 3 provides trace plots and probability density histograms (with overlaid kernal smoothed density functions) for parameters of interest from the full joint model.

3.3. Results

The results in Table 1 show that decreasing CD4 cell count is associated with an increased risk of death, as expected. Specifically, the full joint model predicts that each 10% decrease in CD4 cell count results in a 15.8% ($= e^{-3.215 \log_{10}(0.9)} - 1$) increase in the risk of death. This estimate ranges from 13.6% in the two-stage model to 20.5% in the substitution model. Additionally, CD4 cell count and viral load levels are shown to be inversely related, with each 10% increase in viral load resulting in a predicted 0.64% ($= 1 - 10^{-0.087 \log_{10}(1.1)/1.3}$)

decrease of CD4. This estimated decrease ranges from 0.42% in the two-stage model to 0.74% in the complete-case model. Neither race nor the age at seroconversion were found to be significantly correlated with risk of death. The calendar period associated with HAART treatment was shown to result in an increase in CD4 cell count values in all models. For the full joint model, a participant during the HAART calendar period is expected to have a CD4 cell count value that is 42.9% (= $10^{0.155} - 1$) higher than a participant in the pre-HAART period. Combining this estimate with θ (the survival model estimate for CD4 cell count) indicates that the HAART calendar period is associated with a 39.2% (= $1 - e^{-3.215*0.155}$) decrease in the risk of death for any particular participant. The results from the two-stage, substitution, and complete-case models show a predicted decrease of 38%, 41%, and 44%, respectively. The results in Table 1 also demonstrate that the full joint model may provide standard error estimates that are smaller than the complete-case model and substitutional model, and is more efficient than these two models. Furthermore, the standard errors of the parameter estimates of the survival model in the two-stage model tend to be smaller than the other three approaches. This is because in the two-stage model, the uncertainty of the estimation in the first stage longitudinal model is not incorporated in the second stage of the survival model estimation.

For the results presented in Table 1, the seroconversion date was defined as the midpoint of the last seronegative and the first seropositive dates. For the MACS data, the median and interquartile range of the seroconversion period are 0.5 and (0.5, 0.75) in years. We also conducted sensitivity analysis on the definition of the seroconversion date by assigning the seroconversion dates as the first seropositive dates or at the last seronegative dates of the individual patient. We found that the parameter estimates of interest were robust to the seroconversion date in the MACS data (Table 2).

4. Discussion

We have proposed a joint model for the analysis of longitudinal and survival data that accounts for both missingness and left-censoring in the longitudinal covariates. The proposed model allows the use of a much greater proportion of available data when longitudinal covariates are missing or left-censored. In many infectious disease studies, measures of biomarkers are subject to a lower limit of detection, resulting in many left-censored cases. The proposed methodology accounts for this left-censoring, and also intermittent missingness that can be considered MAR and conditionally depends on the outcome. Previous analyses on only complete-case data are unable to capture the information contained in the missing and/or left-censored biomarkers. Note that the proposed method can also be used to check whether the whole effect of treatment could be captured through CD4 or if there is a remaining effect on survival by including the calendar period in the $\mathbf{x}_{\mathbf{si}}$ and evaluating its regression coefficient estimate. This is known as the issue of imperfect surrogate markers.

The analysis of the MACS data presented in Table 1 shows that posterior estimates obtained from a joint model can be strongly influenced by the inclusion of observations with covariates that are missing or left-censored. In the available data, only 56% of the viral load measurements were observed, with 27.1% missing and 16.9% falling below the limit of detection. Consequently, a complete-case analysis could only be performed on roughly half of the available data points. Including all cases in the proposed joint model is clearly more desirable, and as shown, can produce results that vary from the complete-case results. However, in the MACS data, we did not see a large difference in the estimated relative hazard for the calendar period associated with HAART. Yet, the precision was notably better for the proposed method compared to a complete-case analysis.

The computing time necessary to fit the proposed model can vary widely depending on the software that is used. The analysis results presented here were conducted in R [37], which was able to run approximately 2000 iterations in 24 hours. A simulation study for such a complicated modeling approach is beyond the scope of this paper due to intensive computational time. This relatively long computing time could likely be lessened by using alternative programming languages, such as C or Fortran.

While the proposed modeling approach can improve estimation with missing data in joint models, the assumptions still specify that the intermittently missing covariates are MAR. In many analyses, this assumption may not be correct, as missing data can often arise from a more complicated mechanism. Future research is needed to develop joint models for more complex missing data mechanisms.

In the analysis of the MACS data, the time of HIV seroconversion was imputed as the midpoint of the last seropositive and the first seronegative dates, which corresponds to assigning all mass at the midpoint of the support region as discussed in the paper by Sweeting et al. [44]. Since the non-parametric maximum likelihood estimate may not be unique [44], an alternative is to make a parametric assumption on the joint distribution of the seroconversion date and the AIDS-related death date. For example, let *X* and *Z* denote the unknown dates of seroconversion and death, respectively. The time to AIDS-related death is then defined as *T* = *Z* − *X*. The likelihood of the joint distribution can be modified by conditioning on *X* and *Z*, and can be expressed as

$$\begin{aligned}
&L_i \propto m(X_i, Z_i;\gamma)\ \Big[\{\lambda_0(Z_i-X_i)\exp(\theta\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, Z_i-X_i, \mathrm{VL}_i(Z_i-X_i))+\beta_{s1}\mathrm{RACE}_i+\beta_{s2}\mathrm{AGE}_i)\}^{\delta_i}\\
&\times\exp\left\{-\textstyle\int_0^{Z_i-X_i}\lambda_0(u)\exp(\theta\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, u, \mathrm{VL}_i(u))+\beta_{s1}\mathrm{RACE}_i+\beta_{s2}\mathrm{AGE}_i)du\right\}\Big]\\
&\qquad\qquad\qquad\qquad\times\left[\xi^{n_i/2}\exp\left\{-\tfrac{\xi}{2}\sum_{j=1}^{n_i}(\mathrm{CD4}_{ij}-\psi_{ij}(\boldsymbol{\beta}, \mathbf{b}_i, \mathrm{VL}_{ij}))^2\right\}\right.\\
&\qquad\qquad\qquad\times\left|\textstyle\sum_{\mathbf{b}}^{-\mathbf{1}}\right|^{1/2}\exp\left\{-\tfrac{1}{2}\mathbf{b}_i'\textstyle\sum_{\mathbf{b}}^{-\mathbf{1}}\mathbf{b}_i\right\}\\
&\quad\times\eta_v^{(n_i-1)/2}\exp\left\{-\tfrac{\eta_v}{2}\sum_{j=2}^{n_i}(\mathrm{VL}_{ij}-\alpha_0-\alpha_1\mathrm{VL}_{i,j-1}-\alpha_2\mathrm{RACE}_i-\alpha_3\mathrm{AGE}_i-\alpha_4 t_{ij}(X_i))^2\right\}\\
&\times\tau^{1/2}\exp\left\{-\tfrac{\tau}{2}(\mathrm{VL}_{i1}-\mu_v)^2\right\}\Big],
\end{aligned} \qquad (12)$$

where $m(X_i, Z_i; \gamma)$ is the joint density of (X, Z) given the parameters γ. For a patient who died from AIDS, the time pair (x, z) are known to lie within the region $R_i=(x_i^L, x_i^U) \times d_i$, and for a patient censored without AIDS-related death, the time pair (x, z) are known to lie within the region $R_i=(x_i^L, x_i^U] \times (c_i, \infty)$, where x_i^L is the last seronegative date, x_i^U is the first seropositive date, d_i is the AIDS-related death date, and c_i is the censoring date. In the full joint modeling framework, the estimation procedure can be conducted by modifying the sampling distributions presented in the Appendix and including additional sampling layers for x_i, z_i and γ. These extensions along with joint models for more complex missing data mechanisms are currently under investigation.

Acknowledgments

The authors wish to thank the editor, the associate editor and two referees for several suggestions and editorial changes which have greatly improved the paper. Dr. Chen's research was supported partially by the US NHLBI R21HL097334 and NCRR 1UL1RR024975. Dr. Ibrahim's research was partially supported by NIH grants #GM 70335 and P01CA142538. Dr. Chu is supported in part by the US NIAID R21AI103012, NCI 1P01CA142538 and NCI 2P30CA077598. Dr. Cole was supported in part by the NIH grants R01AI100654, R24AI067039, U01AI103390, and P30AI50410.

References

1. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. Biometrics. 1997:330–339. [PubMed: 9147598]
2. DeGruttola V, Tu X. Modeling progression of CD-4 lymphocyte count and its relationship to survival time. Biometrics. 1994; 50:1003–1012. [PubMed: 7786983]
3. Ibrahim JG, Chu H, Chen LM. Basic concepts and methods for joint models of longitudinal and survival data. Journal of Clinical Oncology. 2010; 28(16):2796–2801. [PubMed: 20439643]
4. Zhang, D.; Chen, MH.; Ibrahim, JG.; Boye, ME.; Wang, P.; Shen, W. Technical Report. Assessing model fit in joint models of longitudinal and survival data with applications to cancer clinical trials.
5. Little, R.; Rubin, D. Statistical Analysis with Missing Data. 2. John Wiley & Sons; 2002.
6. Wu H, Chen Q, Ware L, Koyama T. A Bayesian Approach for Generalized Linear Models with Explanatory Biomarker Measurement Variables Subject to Detection Limit - an Application to Acute Lung Injury. J Appl Stat. 2012; 39:1733–1747. [PubMed: 23049157]
7. Kaslow R, Ostrow D, Detels R, Phair J, Polk B, Rinaldo C. The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. American Journal of Epidemiology. 1987; 126:310–318. [PubMed: 3300281]
8. Little R. Modeling the dropout mechanism in repeated-measures studies. Journal of the American Statistical Association. 1995; 90:1112–1121.
9. Robins J, Rotnitzky A, Zhao L. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association. 1995; 90:106–121.
10. Daniels, MJ.; Hogan, JW. Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis. Chapman & Hall; 2008.
11. Demirtas H, Schafer J. On the performance of random-coefficient pattern-mixture models for non-ignorable dropout. Statistics in Medicine. 2003; 22:2553–2575. [PubMed: 12898544]
12. Brown E, Ibrahim J. Bayesian approaches to joint cure rate and longitudinal models with applications to cancer vaccine trials. Biometrics. 2003; 59:686–693. [PubMed: 14601770]
13. Hogan J, Laird N. Mixture models for the joint distributions of repeated measures and event times. Statistics in Medicine. 1997; 16:239–257. [PubMed: 9004395]
14. Hogan J, Laird N. Model-based approaches to analysing incomplete longitudinal and failure time data. Statistics in Medicine. 1997; 16:259–272. [PubMed: 9004396]
15. Wu L, Liu W, Yi GY, YH. Analysis of Longitudinal and Survival Data: Joint Modeling, Inference Methods, and Issues. J of Prob and Stat. 2012; 201210.1155/2012/640153
16. Wu H, Wu L. A multiple imputation method for missing covariates in non-linear mixed-effects models with application to HIV dynamics. Statistics in Medicine. 2000; 20:1755–69. [PubMed: 11406839]

17. Wu L. Simultaneous inference for longitudinal data with detection limits and covariates measured with errors, with application to AIDS studies. Statistics in Medicine. 2004; 23:1715–31. [PubMed: 15160404]
18. Jacqmin-Gadda H, Thiébaut R, Chêne G, Commenges D. Analysis of left-censored longitudinal data with application to viral load in HIV infection. Biostatistics. 2000; 1:355–68. [PubMed: 12933561]
19. Lyles RH, Lyles CM, Taylor DJ. Random regression models for human immunodeficiency virus ribonucleic acid data subject to left censoring and informative drop-outs. Journal of the Royal Statistical Society: Series C. 2000; 49:485–97.
20. Wu L, Hu X, Wu H. Joint inference for nonlinear mixed-effects models and time to event at the presence of missing data. Biostatistics. 2008; 9:308–320. [PubMed: 17728318]
21. Thiébaut R, Jacqmin-Gadda H, Babiker A, Commenges D. CASCADE Collaboration. Joint modelling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of CD4+ cell count and HIV RNA viral load in response to treatment of HIV infection. Stat Med. 2005; 24:65–82. [PubMed: 15523706]
22. Laird N, Ware J. Random effects models for longitudinal data. Biometrics. 1982; 38:963–974. [PubMed: 7168798]
23. Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. Statistica Sinica. 2004; 14(3):809–834.
24. Fisher LD, Lin D. Time-dependent covariates in the Cox proportional-hazards regression model. Annual review of public health. 1999; 20(1):145–157.
25. Taylor JM, Cumberland W, Sy J. A stochastic model for analysis of longitudinal AIDS data. Journal of the American Statistical Association. 1994; 89(427):727–736.
26. Tsiatis A, Degruttola V, Wulfsohn M. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. Journal of the American Statistical Association. 1995; 90(429):27–37.
27. Wang Y, Taylor JMG. Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. Journal of the American Statistical Association. 2001; 96(455):895–905.
28. Pawitan Y, Self S. Modeling disease marker processes in AIDS. Journal of the American Statistical Association. 1993; 88(423):719– 726.
29. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. Biostatistics. 2000; 1(4):465– 480. [PubMed: 12933568]
30. Guo X, Carlin BP. Separate and joint modeling of longitudinal and event time data using standard computer packages. The American Statistician. 2004; 58(1):16–24.
31. Rizopoulos D. An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. Journal of Statistical Software. 2010; 35:1–33. [PubMed: 21603108]
32. Ibrahim, J.; Chen, M.; Sinha, D. Bayesian Survival Analysis. New York: Springer; 2001.
33. Xu J, Zeger S. Joint analysis of longitudinal data comprising repeated measures and times to events. Applied Statistics. 2001; 50:375–387.
34. Wang Y, Taylor J. Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. Journal of the American Statistical Association. 2001; 96:895–905.
35. Brown E, Ibrahim J. A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. Biometrics. 2003; 59:221–228. [PubMed: 12926706]
36. Guo X, Carlin B. Separate and Joint Modeling of Longitudinal and Event Time Data Using Standard Computer Packages. The American Statistician. 2004; 58:16–24.
37. R Development Core Team. R: A Language and Environment for Statistical Computing.
38. Hammer S, Squires K, Hughes M, Grimes J, Demeter L, Currier J, Eron J, Feinberg J, Balfour H, Dayton L, et al. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. New England Journal of Medicine. 1997; 337:725–733. [PubMed: 9287227]

39. Cameron D, Heath-Chiozzi M, Danner S, Cohen C, Kravcik S, Maurath C, Sun E, Henry D, Rode R, Potthoff A, et al. Randomised placebo-controlled trial of ritonavir in advanced HIV-1 disease. Lancet. 1998; 351:543–549. [PubMed: 9492772]

40. Angrist J, Imbens G, Rubin D. Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association. 1996; 91:444–455.

41. Detels R, Munoz A, McFarlane G, Kingsley L, Margolick J, Giorgi J, Schrager L, Phair J. Effectiveness of Potent Antiretroviral Therapy on Time to AIDS and Death in Men With Known HIV Infection Duration. Journal of the American Medical Association. 1998; 280:1497–1503. [PubMed: 9809730]

42. Tarwater P, Mellors J, Gore M, Margolick J, Phair J, Detels R, Munoz A. Methods to Assess Population Effectiveness of Therapies in Human Immunodeficiency Virus Incident and Prevalent Cohorts. American Journal of Epidemiology. 2001; 154:675–681. [PubMed: 11581102]

43. Gilks W, Best N, Tan K. Adaptive Rejection Metropolis Sampling. Applied Statistics. 1995; 44:455–472.

44. Sweeting MJ, De Angelis D, Parry J, Suligoi B. Estimating the distribution of the window period for recent HIV infections: A comparison of statistical methods. Statistics in Medicine. 2010:3194–202. [PubMed: 21170913]

Appendix

Sampling Distributions

The following section displays the full conditional distributions for the joint model with likelihood given by (11). We use the same notation given in section 3.2, with slight modifications as detailed here. We define $\mathbf{CD4}_i = (\mathrm{CD4}_{i1}, \ldots, \mathrm{CD4}_{in_i})'$, $\mathbf{VL}_i = (\mathrm{VL}_{i1}, \ldots, \mathrm{VL}_{in_i})'$, $\mathbf{PD}_i = (\mathrm{PD}_{i1}, \ldots, \mathrm{PD}_{in_i})'$, and $\mathbf{t}_i = (t_{i1}, \ldots, t_{in_i})'$ to denote the covariate vectors of CD4, viral load, treatment period, and time for subject *i*. For ease of exposition, we will define $\mathbf{z}_i = (\mathrm{RACE}_i, \mathrm{AGE}_i)'$ and $\boldsymbol{\beta}_s = (\beta_{s1}, \beta_{s2})'$, such that $\mathbf{z}_i'\boldsymbol{\beta}_s = \mathrm{RACE}_i\beta_{s1} + \mathrm{AGE}_i\beta_{s2}$. The baseline hazard function λ_0 is specified as piecewise constant, taking the the value λ_k fore each of the $k = 1, \ldots, K$ time intervals, with $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)'$. For time interval *k*, we use d_k to denote the number of failures that occur within that interval. Computation of $\exp\left\{-\int_0^{S_i} \lambda_0(u)\exp(\ldots)du\right\}$ was performed using the approximation given in Ibrahim, Chen, and Sinha (2001, p. 277–278). The notation for this approximation is as follows:

$$\exp\left[-\sum_{k=1}^{K} \lambda_k B_{ik}\right] \approx \exp\left[-\int_0^{S_i} \lambda_0 \exp(\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, u)\theta + \mathbf{z}_i'\boldsymbol{\beta}_s)du\right]$$

To simplify the notation when writing the full conditionals, we will take $\Omega = (\boldsymbol{\lambda}, \theta, \boldsymbol{\beta}_s, \xi, \Sigma_b, \boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\mu}_b, \boldsymbol{\alpha}, \eta_v, \mu_v, \tau)$ to denote the set of all parameters in the model. We will use the notation $\Omega_{(-\boldsymbol{\beta})}$ to denote the set Ω without the parameter $\boldsymbol{\beta}$ (and similar notation when excluding other parameters). We will use the notation $\mathbf{D}_i$ to denote the set of complete data for subject *i*, that is, $\mathbf{D}_i = (\mathbf{CD4}_i, \mathbf{VL}_i, \mathbf{PD}_i, \mathbf{t}_i, S_i, \delta_i, \mathbf{z}_i)$. We use the shorthand notation $\mathbf{D}_{i(-\mathbf{VL}_i)}$ to denote the set of complete data $\mathbf{D}_i$ not including $\mathbf{VL}_i$. The full set of complete data are denoted by $\mathbf{D} = (\mathbf{D}_1, \ldots, \mathbf{D}_N)$ (with $\mathbf{D}_{(-\mathbf{VL}_i)}$ denoting the set of complete data excluding $\mathbf{VL}_i$).

1. $\boldsymbol{\beta}$: $\pi(\boldsymbol{\beta}) \propto \mathbf{1}$.

$$\begin{aligned}&P(\boldsymbol{\beta}|\Omega_{-\boldsymbol{\beta}},\mathbf{D})\\&\propto \exp\left\{-\frac{\xi}{2}\sum_{i=1}^{N}\sum_{j=1}^{n_i}[\mathrm{CD4}_{ij}-\psi_{ij}(\boldsymbol{\beta},\mathbf{b}_i,\mathrm{VL}_{ij})]^2\right\}\times\exp\left\{\sum_{i=1}^{N}\delta_i[(\psi_i(\boldsymbol{\beta},\mathbf{b}_i,S_i,\mathrm{VL}_i(S_i))\theta+\mathbf{z}_i^{'}\boldsymbol{\beta}_s]\right\}\\&\exp\left(-\sum_{i=1}^{N}\sum_{i=1}^{K}\lambda_k B_{ik}\right)\\&=\text{No closed form.}\end{aligned}$$

2. $\mathbf{b}_i$: $P(\mathbf{b}_i) \sim N_2(\boldsymbol{\mu}_b, \Sigma_b)$

$$\begin{aligned}&P(\mathbf{b}_i|\Omega_{(-\mathbf{b}_i)},\mathbf{D})\\&\propto \exp\left\{-\frac{\xi}{2}\sum_{j=1}^{n_i}[\mathrm{CD4}_{ij}-\psi_{ij}(\boldsymbol{\beta},\mathbf{b}_i,\mathrm{VL}_{ij})]^2\right\}\exp\left[-\frac{1}{2}\mathbf{b}_i^{'}\sum\nolimits_{\mathbf{b}}^{-\mathbf{1}}\mathbf{b}_i\right]\\&\times\exp\left\{\delta_i\left[(\psi_i(\boldsymbol{\beta},\mathbf{b}_i,S_i,\mathrm{VL}_i(S_i))\theta+\mathbf{z}_i^{'}\boldsymbol{\beta}_s\right]\right\}\exp\left(-\sum_{k=1}^{K}\lambda_k B_{ik}\right)\\&=\text{No closed form.}\end{aligned}$$

3. ξ: $P(\xi) \sim$ Gamma(Shape = a_ξ, Rate = b_ξ)

$$\begin{aligned}P(\xi|\Omega_{(-\xi)},\mathbf{D})&\propto\prod_{i=1}^{N}\xi^{n_i/2}\exp\left\{-\tfrac{\xi}{2}\sum_{j=1}^{n_i}[\mathrm{CD4}_{ij}-\psi_{ij}(\boldsymbol{\beta},\mathbf{b}_i,\mathrm{VL}_{ij})]^2\right\}\times\xi^{a_\xi-1}\exp(-b_\xi\xi)\\&\propto\xi^{\frac{1}{2}\sum_{i=1}^{n}n_i+a_\xi-1}\exp\left(-\xi\left\{\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}[\mathrm{CD4}_{ij}-\psi_{ij}(\boldsymbol{\beta},\mathbf{b}_i,\mathrm{VL}_{ij})]^2+b_\xi\right\}\right)\\&\sim\text{Gamma}\left(\text{Shape}=\tfrac{1}{2}\sum_{i=1}^{n}n_i+a_\xi,\text{Rate}=\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}[\mathrm{CD4}_{ij}-\psi_{ij}(\boldsymbol{\beta},\mathbf{b}_i,\mathrm{VL}_{ij})]^2+b_\xi\right).\end{aligned}$$

4. $\mu_v{:}P(\mu_v){\sim}N(\mu_{\mu_v},\eta_{\mu_v}^{-1})$

$$\begin{aligned}P(\mu_v|\Omega_{(-\mu_v)},\mathbf{D})&\propto\exp\left[-\tfrac{\tau}{2}\sum_{i=1}^{N}(\mathrm{VL}_{i1}-\mu_v)^2\right]\times\exp\left[-\tfrac{\eta_{\mu_v}}{2}(\mu_v-\mu_{\mu_v})^2\right]\\&\propto\exp\left\{-\tfrac{A}{2}(\mu_v-C)^2\right\}\\&\sim\text{Normal}\,(C,A^{-1})\end{aligned}$$

Where $A = N\tau + \eta_{\mu_v}$, $C=\frac{N_\tau\overline{VL}_1+\eta_{\mu_v}\mu_{\mu_v}}{A}$, and $\overline{VL}_1=\sum_{i=1}^{N}\mathrm{VL}_{i1}/N$.

5. τ: $P(\tau) \sim$ Gamma(Shape = a_τ, Rate = b_τ)

$$
\begin{aligned}
P(\tau|\Omega_{(-\tau)},\mathbf{D}) &\propto \tau^{N/2}\exp\left[-\tfrac{\tau}{2}\sum_{j=1}^{N}(\mathrm{VL}_{i1}-\mu_v)^2\right]\times\tau^{a_\tau-1}\exp\left(-b_\tau\tau\right)\\
&\propto \tau^{\frac{N}{2}+a_\tau-1}\exp\left\{-\tau\left[\tfrac{1}{2}\sum_{i=1}^{N}(\mathrm{VL}_{i1}-\mu_v)^2+b_\tau\right]\right\}\\
&\sim\mathrm{Gamma}\left(\mathrm{Shape}=\tfrac{N}{2}+a_\tau,\mathrm{Rate}=\tfrac{1}{2}\sum_{i=1}^{N}(\mathrm{VL}_{i1}-\mu_v)^2+b_\tau\right).
\end{aligned}
$$

6. α: $\pi(\alpha)\propto 1$

$$
P(\alpha_j|\Omega_{(-\alpha_j)},\mathbf{D})\propto\exp\left\{-\frac{\eta_v}{2}\sum_{i=1}^{N}\sum_{j=2}^{n_i}(\mathrm{VL}_{ij}-\alpha_0-\alpha_1\mathrm{VL}_{i,j-1}-\alpha_2\mathrm{RACE}_i-\alpha_3\mathrm{AGE}_i-\alpha_4t_{ij})^2\right\}
$$

Therefore

$$
\begin{aligned}
&P(\alpha_0|\Omega_{(-\alpha_0)},\mathbf{D})\\
&\sim\mathrm{Normal}\left(\frac{\sum_{i=1}^{N}\sum_{j=2}^{n_i}(\mathrm{VL}_{ij}-\alpha_1\mathrm{VL}_{i,j-1}-\alpha_2\mathrm{RACE}_i-\alpha_3\mathrm{AGE}_i-\alpha_4t_{ij})}{\sum_{i=1}^{N}(n_i-1)},\frac{1}{\eta_v\sum_{i=1}^{N}(n_i-1)}\right)\\
&P(\alpha_1|\Omega_{(-\alpha_1)},\mathbf{D})\\
&\sim\mathrm{Normal}\left(\frac{\sum_{i=1}^{N}\sum_{j=2}^{n_i}\mathrm{VL}_{i,j-1}(\mathrm{VL}_{ij}-\alpha_0-\alpha_2\mathrm{RACE}_i-\alpha_3\mathrm{AGE}_i-\alpha_4t_{ij})}{\sum_{i=1}^{N}\sum_{j=2}^{n_i}\mathrm{VL}_{i,j-1}^2},\frac{1}{\eta_v\sum_{i=1}^{N}\sum_{j=2}^{n_i}\mathrm{VL}_{i,j-1}^2}\right)\\
&P(\alpha_2|\Omega_{(-\alpha_2)},\mathbf{D})\\
&\sim\mathrm{Normal}\left(\frac{\sum_{i=1}^{N}\sum_{j=2}^{n_i}\mathrm{RACE}_i(\mathrm{VL}_{ij}-\alpha_0-\alpha_1\mathrm{VL}_{i,j-1}-\alpha_3\mathrm{AGE}_i-\alpha_4t_{ij})}{\sum_{i=1}^{N}(n_i-1)\mathrm{RACE}_i^2},\frac{1}{\eta_v\sum_{i=1}^{N}(n_i-1)\mathrm{RACE}_i^2}\right)\\
&P(\alpha_3|\Omega_{-(\alpha_3)},\mathbf{D})\\
&\sim\mathrm{Normal}\left(\frac{\sum_{i=1}^{N}\sum_{j=2}^{n_i}\mathrm{AGE}_i(\mathrm{VL}_{ij}-\alpha_0-\alpha_1\mathrm{VL}_{i,j-1}-\alpha_2\mathrm{RACE}_i-\alpha_4t_{ij})}{\sum_{i=1}^{N}(n_i-1)\mathrm{AGE}_i^2},\frac{1}{\eta_v\sum_{i=1}^{N}(n_i-1)\mathrm{AGE}_i^2}\right)\\
&P(\alpha_4|\Omega_{(-\alpha_4)},\mathbf{D})\\
&\sim\mathrm{Normal}\left(\frac{\sum_{i=1}^{N}\sum_{j=2}^{n_i}t_{ij}(\mathrm{VL}_{ij}-\alpha_0-\alpha_1\mathrm{VL}_{i,j-1}-\alpha_2\mathrm{RACE}_i-\alpha_3\mathrm{AGE}_i)}{\sum_{i=1}^{N}\sum_{j=2}^{n_i}t_{ij}^2},\frac{1}{\eta_v\sum_{i=1}^{N}\sum_{j=2}^{n_i}t_{ij}^2}\right).
\end{aligned}
$$

7. η_v: $P(\eta_v)$ ~ Gamma(Shape = a_η, Rate = b_η)

$$\begin{aligned}
&P(\eta_v|\Omega_{(-\eta_v)},\mathbf{D})\\
&\propto \eta_v^{\sum_{i=1}^{N}(n_i-1)/2}\exp\left[-\frac{\eta_v}{2}\sum_{i=1}^{N}\sum_{j=2}^{n_i}(\mathrm{VL}_{ij}-\alpha_0-\alpha_1\mathrm{VL}_{i,j-1}-\alpha_2\mathrm{RACE}_i-\alpha_3\mathrm{AGE}_i-\alpha_4 t_{ij})^2\right]\\
&\times \eta_v^{a_n-1}\exp\left(-b_\eta\eta_v\right)\\
&\propto \eta_v^{\frac{1}{2}\sum_{i=1}^{n}(n_i-1)+a_\eta-1}\exp\left\{-\eta_v\left[\frac{1}{2}\sum_{i=1}^{N}\sum_{j=2}^{n_i}(\mathrm{VL}_{ij}-\alpha_0-\alpha_1\mathrm{VL}_{i,j-1}-\alpha_2\mathrm{RACE}_i-\alpha_3\mathrm{AGE}_i-\alpha_4 t_{ij})^2+b_\eta\right]\right\}\\
&\sim\mathrm{Gamma}\left(\mathrm{Shape}=\frac{1}{2}\sum_{i=1}^{N}(n_i-1)+a_\eta,\right.\\
&\qquad\left.\mathrm{Rate}=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=2}^{n_i}(\mathrm{VL}_{ij}-\alpha_0-\alpha_1\mathrm{VL}_{i,j-1}-\alpha_2\mathrm{RACE}_i-\alpha_3\mathrm{AGE}_i-\alpha_4 t_{ij})^2+b_\eta\right).
\end{aligned}$$

8. $\sum_b^{-1}:P(\sum_b^{-1})\sim\mathrm{Wishart}(n_0,c_0\mathbf{I})$, (**I** = Identity matrix)

$$\begin{aligned}
P(\textstyle\sum_b^{-1}|\Omega_{(-\sum_b)},\mathbf{D}) &\propto |\textstyle\sum_{\mathbf{b}}^{-\mathbf{1}}|^{N/2}\exp\left[-\frac{1}{2}\sum_{i=1}^{N}\mathbf{b_i}'\textstyle\sum_{\mathbf{b}}^{-\mathbf{1}}\mathbf{b_i}\right]\times|\textstyle\sum_{\mathbf{b}}^{-\mathbf{1}}|^{\frac{1}{2}(n_0-p-1)}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[(c_0\mathbf{I})^{-1}\textstyle\sum_{\mathbf{b}}^{-\mathbf{1}}\right]\right\}\\
&\propto |\textstyle\sum_{\mathbf{b}}^{-\mathbf{1}}|^{\frac{1}{2}(N+n_0-p-1)}\exp\left(-\frac{1}{2}\left\{\sum_{i=1}^{N}\mathbf{b_i}'\textstyle\sum_{\mathbf{b}}^{-\mathbf{1}}\mathbf{b_i}+\mathrm{tr}\left[(c_0\mathbf{I})^{-1}\textstyle\sum_{\mathbf{b}}^{-\mathbf{1}}\right]\right\}\right)\\
&\propto |\textstyle\sum_{\mathbf{b}}^{-\mathbf{1}}|^{\frac{1}{2}(N+n_0-p-1)}\exp\left(-\frac{1}{2}\mathrm{tr}\left\{\left[\sum_{i=1}^{N}\mathbf{b_i}\mathbf{b_i}'+(c_0\mathbf{I})^{-1}\right]\textstyle\sum_{\mathbf{b}}^{-\mathbf{1}}\right\}\right)\\
&\sim\mathrm{Wishart}\left(N+n_0,\left[\sum_{i=1}^{N}\mathbf{b}_i\mathbf{b}_i'+(c_0\mathbf{I})^{-1}\right]^{-1}\right).
\end{aligned}$$

9. θ, $\boldsymbol{\beta}_S$: $\pi(\theta)$, $\pi(\boldsymbol{\beta}_S)\propto 1$.

$$\begin{aligned}
P(\theta|\Omega_{-\theta},\mathbf{D}),P(\boldsymbol{\beta}_s|\Omega_{-\boldsymbol{\beta}_s},\mathbf{D})\propto\exp\left\{\sum_{i=1}^{N}\delta_i\left[(\psi_i(\boldsymbol{\beta},\mathbf{b}_i,S_i)\theta+\mathbf{z}_i'\boldsymbol{\beta}_s\right]\right\}\times\exp\left(-\sum_{i=1}^{N}\sum_{k=1}^{K}\lambda_k B_{ik}\right)\\
=\text{No closed form.}
\end{aligned}$$

10. VL_{ij}: $P(\mathrm{VL}_{ij})\sim N(\mu_v,\mu_v)I(0\le\mathrm{VL}_{ij}\le c_{ij})$, where $c_{ij}=\infty$ when VL_{ij} is missing, and $c_{ij}=L_{ij}$ when VL_{ij} is left-censored at limit of detection L_{ij}. I() denotes the indicator function.

$$\begin{aligned}
&P(\mathrm{VL}_{ij}|\Omega,\mathbf{D}_{-\mathrm{VL}_{ij}},0\le\mathrm{VL}_{ij}\le c_u)\\
&\propto\exp\left\{-\frac{\xi}{2}[\mathrm{CD4}_{ij}-\psi_{ij}(\boldsymbol{\beta},\mathbf{b}_i,\mathrm{VL}_{ij})]^2\right\}\times\exp\left\{\delta_i[\psi_i(\boldsymbol{\beta},\mathbf{b}_i,S_i,\mathrm{VL}_i(S_i))\theta+\mathbf{z}_i'\boldsymbol{\beta}_s]\right\}\times\exp\left(-\sum_{k=1}^{K}\lambda_k B_{ik}\right)\\
&\times F_{ij}\times I(0\le\mathrm{VL}_{ij}<c_{ij})\\
&=\text{No closed form,}
\end{aligned}$$

where

$$F_{ij}=\exp[-\tfrac{\tau}{2}(\mathrm{VL}_{i1}-\mu_v)^2]\exp[-\tfrac{\eta_v}{2}(\mathrm{VL}_{i2}-\alpha_0-\alpha_1\mathrm{VL}_{i1}-\alpha_2\mathrm{RACE}_i-\alpha_3\mathrm{AGE}_i-\alpha_4 t_{ij})^2]$$

, when $j = 1$;

$$F_{ij}=\exp[-\frac{\eta_v}{2}(\mathrm{VL}_{ij}-\alpha_0-\alpha_1\mathrm{VL}_{i,j-1}-\alpha_2\mathrm{RACE}_i-\alpha_3\mathrm{AGE}_i-\alpha_4 t_{ij})^2]\exp[-\frac{\eta_v}{2}(\mathrm{VL}_{i,j+1}-\alpha_0-\alpha_1\mathrm{VL}_{ij}-\alpha_2\mathrm{RACE}_i-\alpha_3\mathrm{AGE}_i-\alpha_4 t_{ij})^2]$$

, when $1 < j < n_i$; $F_{ij}=\exp[-\frac{\eta_v}{2}(\mathrm{VL}_{ij}-\alpha_0-\alpha_1\mathrm{VL}_{i,j-1}-\alpha_2\mathrm{RACE}_i-\alpha_3\mathrm{AGE}_i-\alpha_4 t_{ij})^2]$, when $j = n_i$.

11. λ_k: $P(\lambda_k) \sim \mathrm{Gamma}(\mathrm{Shape} = a_\lambda, \mathrm{Rate} = b_\lambda)$

$$\begin{aligned} P(\lambda_k|\Omega_{(-\lambda_k)},\mathbf{D}) &\propto \lambda_k^{d_k}\exp\left(-\sum_{i=1}^{N}\lambda_k B_{ik}\right)\lambda_k^{a_\lambda-1}\exp(-b_\lambda\lambda_k) \\ &\sim \mathrm{Gamma}\left(\mathrm{Shape}=d_k+a_\lambda, \mathrm{Rate}=\sum_{i=1}^{N}B_{ik}+b_\lambda\right). \end{aligned}$$

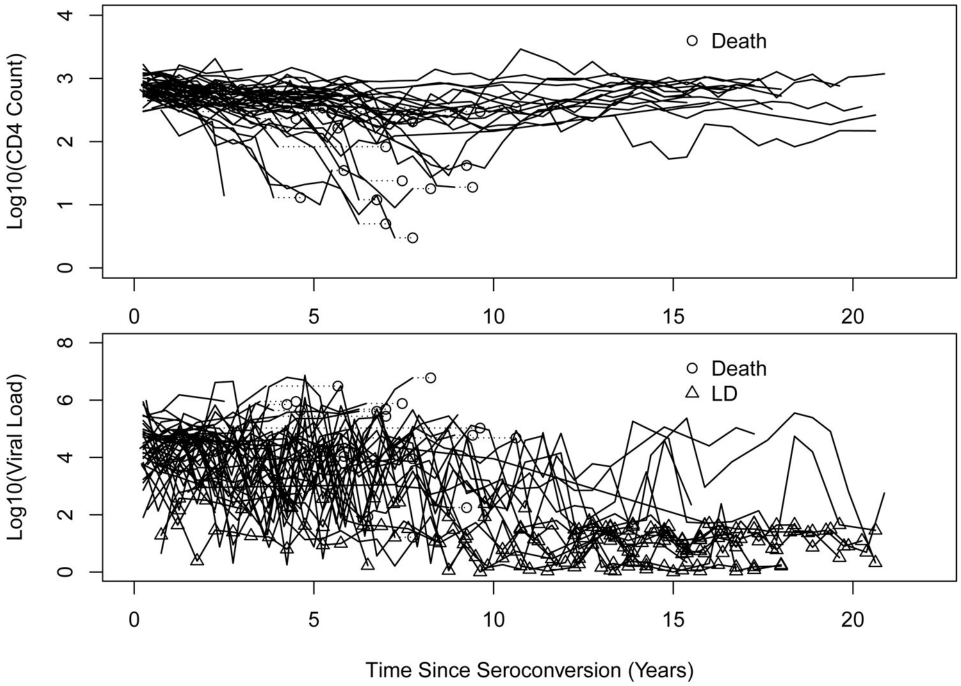

Figure 1.

CD4 Cell Count and Viral Load Trajectories for Random Sample of 50 Participants

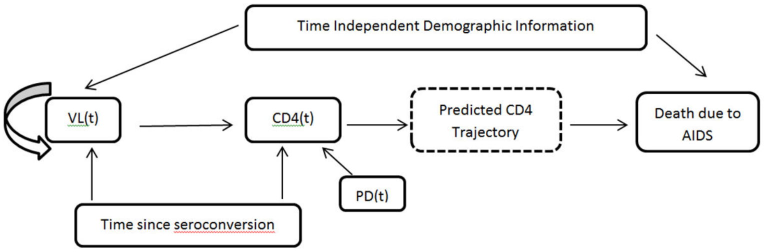
Time Independent Demographic Information
VL(t)
CD4(t)
Predicted CD4 Trajectory
Death due to AIDS
PD(t)
Time since seroconversion

Figure 2.

Joint Modeling Strategy. Solid box: observed; dashed box: predicted; arrow: information used in the modeling of the corresponding variable.

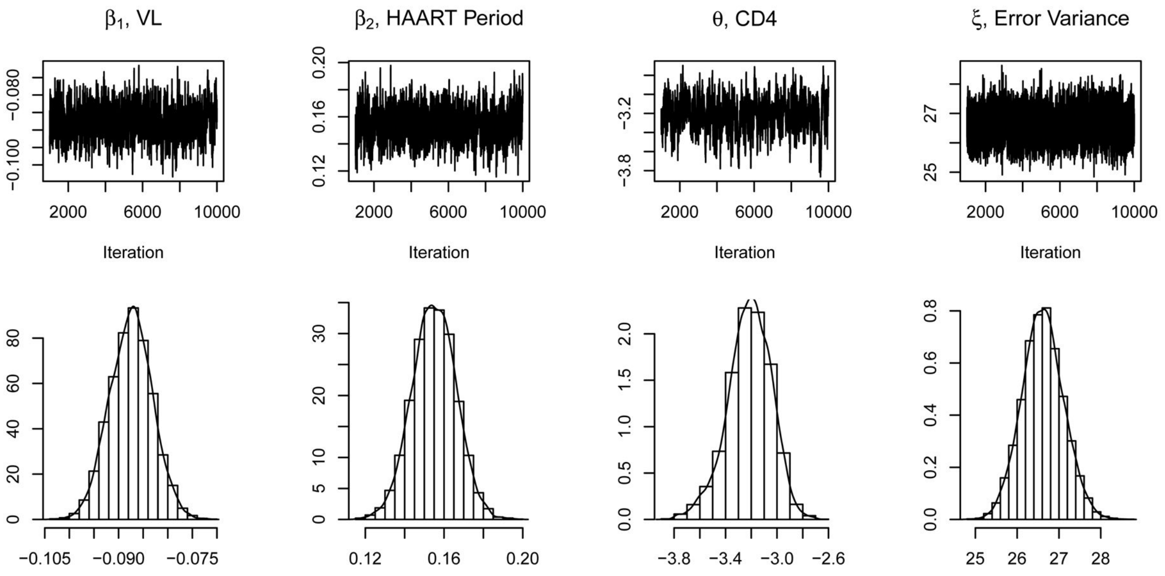

Figure 3.

Trace Plots and Sampled Densities of Selected Parameters from Full Joint Model

Table 1

Parameter estimates for MACS data analysis in all models

	Two-Stage Model[1]				Complete-Case Joint[2]				Substitution Joint[3]				Full Joint[4]			
Parameter	**Mean**	**SD**	**Lower***	**Upper***	**Mean**	**SD**	**Lower**	**Upper**	**Mean**	**SD**	**Lower**	**Upper**	**Mean**	**SD**	**Lower**	**Upper**
Longitudinal Component																
$\beta_1(VL)$	−0.058	0.004	−0.064	−0.052	−0.101	0.007	−0.113	−0.090	−0.071	0.004	−0.078	−0.064	−0.087	0.004	−0.096	−0.079
$\beta_2(PD)$	0.173	0.012	0.153	0.192	0.145	0.014	0.122	0.168	0.128	0.012	0.109	0.148	0.155	0.011	0.133	0.176
β_{b0}	3.071	0.014	3.048	3.094	3.230	0.023	3.194	3.271	3.084	0.016	3.058	3.109	2.908	0.006	2.897	2.919
β_{b1}	−0.105	0.003	−0.110	−0.110	−0.103	0.003	−0.107	−0.099	−0.090	0.002	−0.093	−0.086	−0.101	0.002	−0.105	−0.097
Σ_{b11}	0.092	0.008	0.079	0.105	0.037	0.004	0.031	0.044	0.048	0.005	0.041	0.056	0.049	0.004	0.042	0.058
Σ_{b12}	−0.029	0.004	−0.035	−0.023	−0.008	0.001	−0.011	−0.006	−0.011	0.002	−0.014	−0.009	−0.012	0.002	−0.016	−0.010
Σ_{b22}	0.025	0.003	0.021	0.030	0.009	0.001	0.008	0.011	0.009	0.001	0.008	0.011	0.010	0.001	0.008	0.012
ξ	26.529	0.499	25.710	27.352	26.495	0.662	25.406	27.599	27.430	0.595	26.464	28.414	26.613	0.450	25.649	27.610
Parameters in Biomarker Model																
μ_v	-	-	-	-	-	-	-	-	-	-	-	-	0.357	0.031	0.295	0.419
τ_v	-	-	-	-	-	-	-	-	-	-	-	-	2.614	0.193	2.256	3.007
α_0	-	-	-	-	-	-	-	-	-	-	-	-	0.036	0.056	−0.070	0.150
α_1	-	-	-	-	-	-	-	-	-	-	-	-	0.882	0.029	0.818	0.934
α_2	-	-	-	-	-	-	-	-	-	-	-	-	−0.000	0.000	−0.001	0.001
α_3	-	-	-	-	-	-	-	-	-	-	-	-	−0.000	0.000	−0.001	0.001
α_4	-	-	-	-	-	-	-	-	-	-	-	-	−0.014	0.007	−0.029	0.002
η_v	-	-	-	-	-	-	-	-	-	-	-	-	2.895	0.197	2.333	3.129
Survival Component																
$\theta(CD4)$	−2.777	0.120	−2.980	−2.584	−3.987	0.246	−4.393	−3.578	−4.079	0.219	−4.439	−3.721	−3.215	0.170	−3.590	−2.912
$\beta_{s1}(Race)$	0.183	0.224	−0.178	0.561	0.277	0.292	−0.198	0.768	0.355	0.285	−0.100	0.842	0.252	0.240	−0.207	0.731
$\beta_{s2}(Age)$	0.065	0.076	−0.059	0.192	0.129	0.092	−0.020	0.280	0.125	0.096	−0.035	0.282	0.050	0.088	−0.126	0.218

*Lower and upper 95% bounds from parameter distribution.

[1] Model fitting first longitudinal model, then using longitudinal trajectory as fixed covariate in separate survival model.

[2] Joint model on 56% of observations with observed viral load.

[3] Joint model substituting $\mathrm{LD}/\sqrt{2}$ for 16.9% of observations with viral load below limit of detection and removing 27.1% missing viral load values.

Table 2

MACS data analysis using the full joint model – Sensitivity analysis on the seroconversion date

	Last Seronegative Date[1]				Midpoint of Last Seronegative[2] and First Seropositive Dates				First Seropositive Date[3]			
Parameter	**Mean**	**SD**	**Lower**	**Upper**	**Mean**	**SD**	**Lower**	**Upper**	**Mean**	**SD**	**Lower**	**Upper**
Longitudinal Component												
$\beta_1(VL)$	−0.092	0.005	−0.101	−0.082	−0.087	0.004	−0.096	−0.079	−0.087	0.004	−0.095	−0.078
$\beta_2(PD)$	0.151	0.011	0.129	0.173	0.155	0.011	0.133	0.176	0.154	0.011	0.132	0.175
β_{b0}	2.942	0.007	2.929	2.955	2.908	0.006	2.897	2.919	2.868	0.005	2.858	2.879
β_{b1}	−0.099	0.002	−0.103	−0.095	−0.101	0.002	−0.105	−0.097	−0.103	0.002	−0.107	−0.099
Σ_{b11}	0.058	0.005	0.049	0.068	0.049	0.004	0.042	0.058	0.043	0.004	0.037	0.051
Σ_{b12}	−0.015	0.002	−0.018	−0.012	−0.012	0.002	−0.016	−0.010	−0.010	0.001	−0.008	−0.012
Σ_{b22}	0.009	0.001	0.008	0.011	0.010	0.001	0.008	0.012	0.010	0.001	0.008	0.012
ξ	26.853	0.533	25.824	27.900	26.613	0.450	25.649	27.610	26.603	0.489	25.649	27.549
Parameters in Biomarker Model												
μ_v	0.355	0.033	0.291	0.419	0.357	0.031	0.295	0.419	0.373	0.032	0.310	0.437
τ_v	2.627	0.200	2.242	3.025	2.614	0.193	2.256	3.007	2.625	0.195	2.264	3.021
α_0	0.048	0.131	−0.137	0.250	0.036	0.056	−0.070	0.150	0.030	0.024	−0.016	0.078
α_1	0.875	0.066	0.769	0.963	0.882	0.029	0.818	0.934	0.884	0.013	0.857	0.909
α_2	0.000	0.001	−0.001	0.001	−0.000	0.000	−0.001	0.001	−0.000	0.000	−0.001	0.001
α_3	−0.000	0.000	−0.001	0.001	−0.000	0.000	−0.001	0.001	−0.000	0.000	−0.001	0.001
α_4	−0.015	0.015	−0.039	0.006	−0.014	0.007	−0.029	0.002	−0.014	0.003	−0.020	0.008
η_v	2.323	0.333	1.691	2.979	2.895	0.197	2.333	3.129	3.042	0.074	2.885	3.175
Survival Component												
$\theta(CD4)$	−3.268	0.175	−3.621	−2.928	−3.215	0.170	−3.590	−2.912	−3.254	0.169	−3.588	−2.935
$\beta_{s1}(Race)$	0.299	0.248	−0.168	0.803	0.252	0.240	−0.207	0.731	0.254	0.240	−0.191	0.746
$\beta_{s2}(Age)$	0.050	0.089	−0.124	0.225	0.050	0.088	−0.126	0.218	0.038	0.088	−0.133	0.209

[1] Define the seroconversion date as the last seronegative date.

[2] Define the seroconversion date as the midpoint of the last seronegative and the first seropositive dates.

[3] Define the seroconversion date as the first seropositive date.