**Research Article**

Sihaoyu Gao, Lang Wu*, Tingting Yu, Roger Kouyos, Huldrych F. Günthard and Rui Wang

# Nonlinear mixed-effects models for HIV viral load trajectories before and after antiretroviral therapy interruption, incorporating left censoring

https://doi.org/10.1515/scid-2021-0001

Received January 26, 2021; accepted February 28, 2022; published online April 4, 2022

**Abstract**

**Objectives:** Characterizing features of the viral rebound trajectories and identifying host, virological, and immunological factors that are predictive of the viral rebound trajectories are central to HIV cure research. We investigate if key features of HIV viral decay and CD4 trajectories during antiretroviral therapy (ART) are associated with characteristics of HIV viral rebound following ART interruption.

**Methods:** Nonlinear mixed effect (NLME) models are used to model viral load trajectories before and following ART interruption, incorporating left censoring due to lower detection limits of viral load assays. A stochastic approximation EM (SAEM) algorithm is used for parameter estimation and inference. To circumvent the computational intensity associated with maximizing the joint likelihood, we propose an easy-to-implement three-step method.

**Results:** We evaluate the performance of the proposed method through simulation studies and apply it to data from the Zurich Primary HIV Infection Study. We find that some key features of viral load during ART (e.g., viral decay rate) are significantly associated with important characteristics of viral rebound following ART interruption (e.g., viral set point).

**Conclusions:** The proposed three-step method works well. We have shown that key features of viral decay during ART may be associated with important features of viral rebound following ART interruption.

**Keywords:** censoring; HIV/AIDS studies; longitudinal data; stochastic approximation EM (SAEM) algorithm.

***Corresponding author**: **Lang Wu**, Department of Statistics, University of British Columbia, Vancouver, BC, Canada, E-mail: lang@stat.ubc.ca

**Sihaoyu Gao**, Department of Statistics, University of British Columbia, Vancouver, BC, Canada. https://orcid.org/0000-0002-1319-4188

**Tingting Yu**, Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA, USA

**Roger Kouyos and Huldrych F. Günthard**, Department of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, Zurich, Switzerland; and Institute of Medical Virology, University of Zurich, Zurich, Switzerland. https://orcid.org/0000-0002-1142-6723 (H.F. Günthard)

**Rui Wang**, Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA, USA; and Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. https://orcid.org/0000-0001-5007-193X

# Introduction

Characterizing features of the viral rebound trajectories and identifying host, virological, and immunological factors that are predictive of the viral rebound trajectories are central to HIV cure research (Julg et al. 2019; Li, Smith, and Mellors 2015; Richman et al. 2009). After the initiation of an antiretroviral therapy (ART), viral loads typically decline over time and subsequently drop below the assay's detection limit, i.e., the viral loads may be *left censored*. If the ART is discontinued, viral loads usually rise rapidly to peak values, then decrease and stabilize at a level commonly referred to as a viral set point. There are many barriers to curing HIV. Efforts have been focusing on either a functional cure (lowering viral set points) or a sterilizing cure (eliminating all HIV-infected cells), with the former being a more realistic goal (Rouzioux, Hocqueloux, and Sáez-Cirión 2015).

To date, a number of biomarkers have been found to be predictive of the timing of viral rebound or viral set point after treatment interruption including ART initiation during acute/early HIV infection (Namazi et al. 2018; Von Wyl et al. 2011), pre-ATI (analytic treatment interruption) CA-RNA (cell-associated RNA) levels (Li et al. 2016), and T-cell exhaustion markers measured prior to ART (Hurst et al., 2015). So far, to the best of our knowledge, most studies focused on predictors that reflect values at a single time point (e.g., age at the start of treatment interruption), or provide simple summaries of observed values over history (e.g., nadir CD4 count, changes in numbers of cytotoxic T-lymphocytes) (Bing et al. 2020; Conway, Perelson, and Li 2019; Oxenius et al. 2002; Wang et al. 2020). Here we investigate the effect of longitudinal biomarkers on features of viral rebound, leveraging rich information from the Zurich Primary HIV Infection Study, where viral load and CD4 cell counts were measured longitudinally since seroconversion.

The Zurich Primary HIV Infection Study consists of participants presenting with acute or recent HIV-1 infection between November 2002 and July 2008. This study was described in details in Gianella et al. (2011). In brief, acutely and recently HIV-1 infected individuals were offered immediate standard first line combination ART (cART) according to treatment recommendations of that time (Thompson et al. 2010), and after at least one year of viral suppression below detection limits (<50 HIV-1 RNA copies/mL of plasma), they could elect to stop therapy. Reinitiation of cART was based on CD4 count criteria of that time. Figure 1 shows entire viral load (HIV-1 RNA copies/mL of plasma, in $\log_{10}$-scale) trajectories during ART and following ART interruption for all subjects and for 5 randomly selected subjects respectively (for data following ART interruption, we only show the first 36 weeks of data because viral load levels typically stabilize before then). We see that viral loads decline rapidly during ART and then may rebound quickly following ART interruption, and that viral loads after reaching peak points during rebound exhibit large variations between subjects. Our main objective is to study if key features of viral decay during ART, such as individual-specific viral decay rates, are associated with important characteristics of viral rebound following ART interruption, such as individual-specific viral rebound rates or set points.

Mixed effects models are well-suited to model longitudinal data with large variations between individual viral load trajectories, since random effects in the models can be used to incorporate the between-individual variations, as well as individual-specific inference. To model viral load trajectories during ART, Wu and Ding (1999) proposed nonlinear mixed effects (NLME) models based on reasonable biological arguments, and the proposed exponential decay models have been shown to fit viral load data during ART very well. For viral rebound trajectories following ART interruption, Wang et al. (2020) proposed a different NLME model where the key features of viral rebounds are represented by the model parameters. Thus, here we use these two NLME models to model viral load before and after ART interruptions respectively. Mixed effects models with left censored responses have also been studied in the literature (Hughes 1999; Vaida, Fitzgerald, and DeGruttola 2007; Vaida and Liu 2009; Wu 2002).

Statistical inference for NLME models is typically based on the likelihood method (Wu 2009). Due to unobservable random effects and nonlinearity of the models, exact likelihood estimation based on the numerical integration methods or Monte Carlo expectation-maximization (MCEM) algorithm can be computationally intensive and may suffer from convergence problems (Wu 2009). A widely used and computationally more

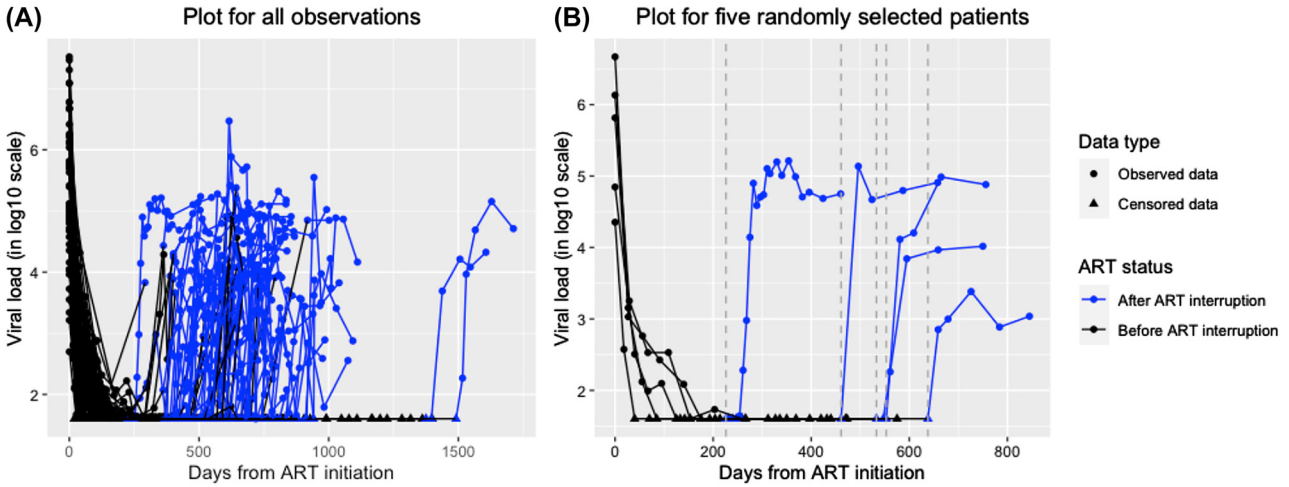


**Figure 1:** Viral load ($log_{10}$-scale) trajectories before and following ART interruption. Left censored values are denoted by triangle dots on the bottom horizontal line with the censored values imputed by the detection limit. Observed values are denoted by circle dots. Data during ART are in black, and data following ART interruption are in blue. The dashed vertical lines in gray indicate times when the ART was interrupted. Figure (A) shows data from all subjects, and Figure (B) shows data from 5 randomly selected subjects.

efficient approximate method is the linearization method of Lindstrom and Bates (Lindstrom and Bates 1990), but its performance can be less satisfactory in some cases (Comets, Lavenu, and Lavielle 2017). Here we consider the stochastic approximation expectation-maximization (SAEM) algorithm for parameter estimation and inference of NLME models (Comets, Lavenu, and Lavielle 2017; Delyon, Lavielle, and Moulines 1999; Samson, Lavielle, and Mentré 2006). The SAEM algorithm is computationally more efficient than the MCEM algorithm and it performs well in the sense of producing reasonable estimates and fast convergence.

In this article, we consider three mixed effects models: two NLME models with left censored responses – one NLME model for viral dynamics during ART and another NLME model for viral rebound following ART interruption, and a linear mixed effects (LME) model for CD4 data during ART. The three models are linked through shared random effects. To reduce the computational burden, we fit the three models separately based on a three-step method, using the SAEM algorithm. We use a parametric bootstrap method to obtain standard errors of all parameter estimates and incorporate estimation uncertainty from separate model fittings. Our contributions are: (i) to our knowledge, our work is the first to study the relationship between viral declines during ART and viral rebound after ART based on NLME models; (ii) the proposed three-step method is simple, and can be implemented with existing software; (iii) the proposed method is based on exact likelihood method, so there is no concern about approximation accuracies as in other approximate methods such as linearization methods, and it is also computationally feasible; and (iv) the proposed method performs well, as shown in simulations, and clearly outperforms a common naive method that uses an imputed value for censored observations and model-based standard errors.

The article is organized as follows. In Section 2, we describe the models motivated by the real dataset, and we propose a three-step method for parameter estimation. Section 3 presents a comprehensive data analysis. The proposed method is evaluated in Section 4 via simulations. We conclude the article with some discussion in Section 5.

# Models for data before and following ART interruption

In this section, we first consider an NLME model for viral decay and an LME model for CD4 trajectories during ART. Then, the random effects in these two models, which summarize individual-specific CD4 and viral load trajectories, are used as "covariates" in the viral rebound NLME model following ART interruption. Our goal is to exam if the individual-specific viral rebound characteristics following ART interruption are associated with individual-specific CD4 and viral load profiles during ART.

## Models for viral load and CD4 during ART

First, we model viral load trajectories during ART. Let $Y_{ij}$ be the ($\log_{10}$-transformed) viral load value (in copies/mL) of individual $i$ measured at time $t_{ij}$ during ART. Let $y_{ij}$ be the observed value of $Y_{ij}$, and let $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{in_i})^{\mathrm{T}}$, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, n_i$. We use similar notation for other variables. The values of $Y$ may be left censored due to the assay's lower detection limit. Based on possible virus elimination and production processes, Wu and Ding (1999) showed that the viral load trajectories during ART typically exhibit exponential decay patterns and may be modelled by NLME models.

A general NLME model can be written as follows:

$$\begin{aligned} &y_{ij} = g(\boldsymbol{x}_{ij}, \boldsymbol{\eta}, \boldsymbol{b}_i) + e_{ij}, \\ &\mathbf{b}_i \sim N(0, B), \quad \mathbf{e}_i \sim N(0, \Sigma_i), \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, n_i, \end{aligned} \tag{1}$$

where $g(\cdot)$ is a known nonlinear function, $\boldsymbol{x}_{ij}$ is a vector containing covariates including time, $\boldsymbol{\eta}$ is a vector containing fixed effect parameters, $\mathbf{b}_i = (b_{1i}, \dots, b_{qi})^{\mathrm{T}}$ contains random effects, $\mathbf{e}_i = (e_{i1}, e_{i2}, \dots, e_{in_i})^{\mathrm{T}}$ contains within-individual random errors, and $B$ and $\Sigma_i$ are covariance matrices. The random effects $\mathbf{b}_i$ and the

random errors $\mathbf{e}_i$ are assumed to be independent. When the function $g(\cdot)$ is a linear function, the NLME model (1) reduces to an LME model. It is common to assume that the within-individual errors are conditionally independent given the random effects, i.e., $\Sigma_i = \sigma^2 I_{n_i}$, where $I_{n_i}$ is the $n_i \times n_i$ identity matrix.

For the viral load data during ART, as shown in Figure 1, we consider the following NLME model (Wu and Ding 1999)

$$\begin{aligned} y_{ij} &= \log_{10}\left(e^{P_{1i}-\lambda_{1i}t_{ij}} + e^{P_{2i}-\lambda_{2i}t_{ij}}\right) + e_{ij}, \quad i = 1, 2, \ldots, n, \ j = 1, 2, \ldots, n_i, \\ P_{1i} &= P_1 + b_{1i}, \quad P_{2i} = P_2 + b_{2i}, \quad \lambda_{1i} = \lambda_1 + b_{3i}, \quad \lambda_{2i} = \lambda_2 + b_{4i}, \end{aligned} \tag{2}$$

where $\mathbf{b}_i = (b_{1i}, b_{2i}, b_{3i}, b_{4i})^{\mathrm{T}}$ are random effects, $\lambda_1$ is the first-phase viral decay rate, which corresponds to the rapid decay phase reflecting decay of productively, long-lived and/or latently infected cells, $\lambda_2$ is the second-phase viral decay rate during ART, which corresponds to the slow decay phase reflecting decay of long-lived and/or latently infected cells and other residual infected cells, $\log_{10}\left(e^{P_1} + e^{P_2}\right)$ is typical viral load value at the start of ART. This NLME model is rooted from a biological compartment model describing the interaction between HIV and its host cells and has been shown to provide a good fit to the viral decay phase after ART initiation (Wu and Ding 1999). Figure 2A shows the viral decline profile for a typical subject based on model (2). Random effects are introduced to each parameter to incorporate large variations between individuals. As shown in Figure 1, some viral load values are *left censored* or below the detection limit. In estimating the parameters in the above NLME model, the censored viral load values must be taken into account to avoid biased results (Hughes 1999; Wu 2002).

In addition to viral loads, CD4 cell count during ART may be also associated with viral rebounds, due to the known association between CD4 and viral load. The observed CD4 values are highly variable, reflecting both short-term biological variation and measurement error. To address measurement errors in the observed CD4 values, we may model the observed CD4 longitudinal data empirically to estimate true CD4 values. Specifically, let $z_{ij}$ be the observed CD4 cell count (in cells/mm$^3$) of individual $i$ measured at time $t_{ij}$, $i = 1, 2, \ldots, n, j = 1, 2, \ldots, m_i$. We may consider the following general LME model for CD4 data

$$z_{ij} = \mathbf{u}_{ij}^{\mathrm{T}}\boldsymbol{\alpha} + \mathbf{v}_{ij}^{\mathrm{T}}\mathbf{a}_i + \varepsilon_{ij} \equiv z_{ij}^* + \varepsilon_{ij}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m_i, \tag{3}$$

where vectors $\mathbf{u}_{ij}$ and $\mathbf{v}_{ij}$ contain covariates including time, vector $\boldsymbol{\alpha}$ contains fixed effect parameters, vector $\mathbf{a}_i$ contains random effects with $\mathbf{a}_i \sim N(0, A)$, $z_{ij}^*$ is the assumed (unobserved) true CD4 value whose corresponding observed error-prone value is $z_{ij}$ based on the classical measurement error model, and $\varepsilon_{ij}$ is the measurement error, with $\varepsilon_{ij}$'s are $\sim N(0, \delta^2)$. We may take appropriate transformations of the observed CD4 values, such as a log-transformation or a $\sqrt{z_{ij}}$-transformation so that the transformed data are more compatible with the normality and constant variance assumptions. We compare different models based on observed/predicted plots, normal QQ-plots, and residual plots (see Appendix), as well as the simplicity of the model and its interpretation. We find that a $\sqrt{z_{ij}}$-transformation of CD4 provides satisfactory results since it produces similar results as the log-transformation and it is widely used in the analysis of datasets from the AIDS Clinical Trials Group network (ACTG) (see, e.g., Noubary and Hughes 2012).

Note that the general LME model (3) includes nonparametric mixed effects models which may be useful if the CD4 trajectories are complicated without clear patterns, since we can use a basis-based approach to approximate the nonparametric mixed model by an LME model (Wu 2009). Thus, the general LME model (3) is quite flexible for modelling complex CD4 longitudinal data.

For the motivating dataset shown in Section 1, we considered several empirical polynomial LME models for CD4 data during ART. We find that the following simple empirical LME model fits the CD4 data reasonably well (see Appendix)

$$z_{ij} = \alpha_{1i} + \alpha_{2i}t_{ij} + \varepsilon_{ij}, \quad \alpha_{1i} = \alpha_1 + a_{1i}, \quad \alpha_{2i} = \alpha_2 + a_{2i}, \tag{4}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^{\mathrm{T}}$ are fixed effects and $\mathbf{a}_i = (a_{1i}, a_{2i})^{\mathrm{T}}$ are random effects. More complex models, such as a quadratic LME model $z_{ij} = \alpha_{1i} + \alpha_{2i}t_{ij} + \alpha_{3i}t_{ij}^2 + \varepsilon_{ij}$ with $\alpha_{3i} = \alpha_3 + a_{3i}$, do not improve the model fit substantially but they are more complex and less stable. Thus, we choose the simpler LME model (4) for the (square root transformed) CD4 data.

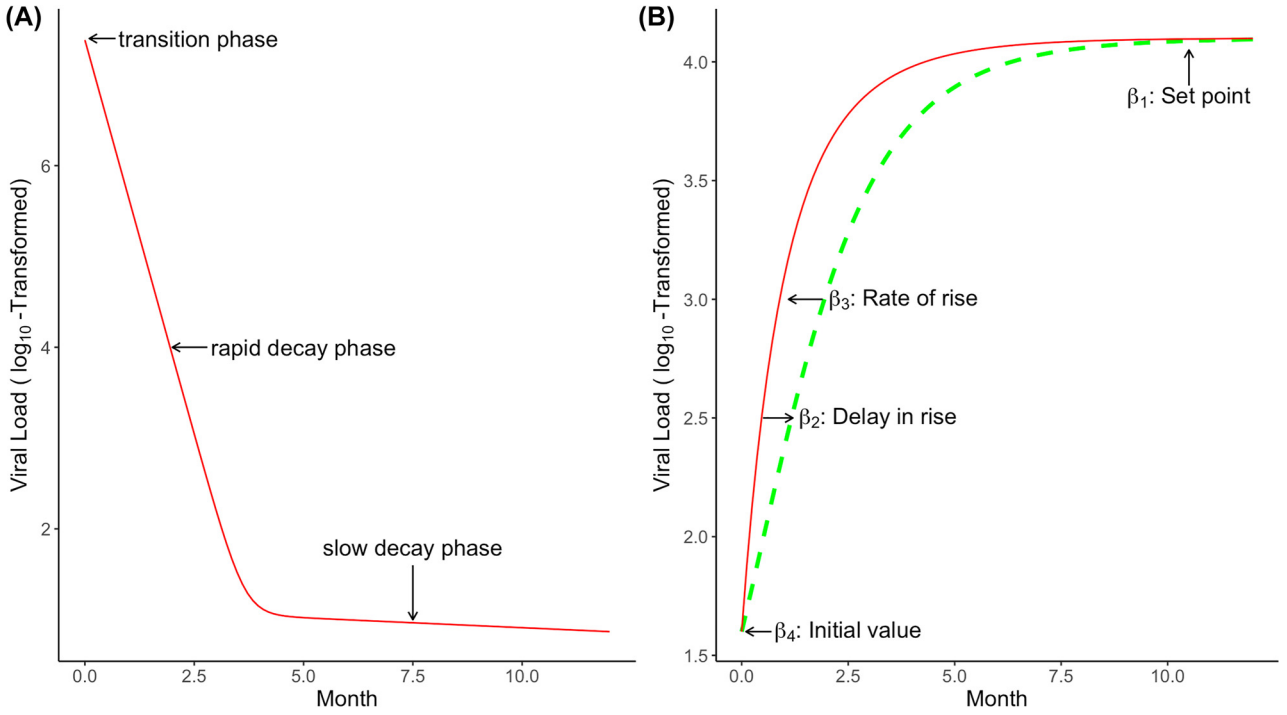


**Figure 2:** Viral load trajectories during and after ART.
(A) A schematic illustration of viral dynamics profiles during ART based on model (2). (B) A schematic illustration of viral rebound profiles following ART interruption based on model (5).

## A viral rebound model following ART interruption

We now model viral rebound data following ART interruption. Let $w_{ij}$ be the ($\log_{10}$-transformed) viral load value of individual $i$ measured at time $t^*_{ij}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n^*_i$, where $t^*_{ij}$ is the time since ART interruption (not since the start of ART). After ART interruption, the viral load trajectories typically rise to a peak value followed by a decrease to a viral load set point. Wang et al. (2020) proposed an NLME model with a flexible functional form to capture this non-linear trajectory and provide biological insights regarding the rebound process. In comparison to non-parametric modeling approaches such as the use of penalized smoothing splines (Zhao et al. 2021), key features of viral rebound trajectories are represented by the parameters in the model, which provides a means to assess the covariate effects on each of these model parameters and allows us to identify critical pre-ATI predictors for these features directly. Comparison of this parametric NLME model to a dynamic viral model (Prague et al. 2019) found that both modeling approaches led to good individual fits and consistent conclusions regarding the features of the viral rebound process (Bing et al. 2020).

Following Wang et al. (2020) (with minor modification), we consider the following NLME model for modelling the viral rebound following ART interruption

$$w_{ij} = \beta_{1i} \frac{t^*_{ij}}{t^*_{ij} + \exp\left(\beta_{2i} - \beta_{3i} t^*_{ij}\right)} + \beta_{4i} + \xi_{ij}, \tag{5}$$

$$\boldsymbol{\beta}_i = R_i \boldsymbol{\beta} + \boldsymbol{\tau}_i, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, n^*_i, \tag{6}$$

where vector $\boldsymbol{\beta}_i = (\beta_{1i}, \ldots, \beta_{4i})^{\mathrm{T}}$ contains individual-specific parameters, vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{q^*})^{\mathrm{T}}$ contains fixed effect parameters, $R_i$ is a $4 \times q^*$ design matrix contains covariates, and $\boldsymbol{\tau}_i = (\tau_{1i}, \ldots, \tau_{4i})^{\mathrm{T}} \sim N(0, G)$ contains random effects with $G$ being a covariance matrix, and $\xi_{ij}$ is within-individual random error. We assume that the random effects $\boldsymbol{\tau}_i$ and the random error $\xi_{ij}$ are independent, and $\xi_{ij}$ are i.i.d. $\sim N(0, \omega^2)$. Here the modification from Wang et al. (2020) is to replace the term that describes the decline of the viral load from peak to set-point by a constant parameter. The reason for this modification is that our viral load data after their peaks during rebound exhibit large between-subject variations without a clear pattern, so the parameter in the original model of Wang et al. (2020) that characterizes the 'dip' after the leak may not be estimated well; including such parameter in the model may also lead to non-convergences issues. In our current setting, we focus on modeling the rate of rise and viral set points. While the simplification with a constant term in the above model limits our ability to estimate the peak and the decline from the peak, it provides a good fit to the data and allows the parameters of primary interest (rate of rise and viral set point) to be estimated well.

Note that the above NLME model (5) and (6) may be viewed as a *two-stage model*: In stage 1, model (5) describes the viral rebound trajectories within an individual; and in stage 2, model (6) assumes that the between-individual variations in the individual-specific parameters in model (5) may be partially explained by covariates in $R_i$ as well as random effects $\boldsymbol{\tau}_i$.

The parameters in NLME model (5) have the following attractive interpretations (Wang et al. 2020): parameter $\beta_{1i}$ represents *set point* after rebound, parameter $\beta_{2i}$ and $\beta_{3i}$ respectively characterize the *timing* and *rate of rise* in viral rebound, and parameter $\beta_{4i}$ denotes *initial viral load value* at the start of rebound. Figure 2B shows the viral load rebound profile for a typical subject based on model (5). Therefore, each of the four parameters denotes an important characteristic of the viral rebound trajectories following ART interruption.

As noted in Section 1, our main objective is to assess if key features of viral load or CD4 trajectories *after ART initiation* may be associated with important characteristics of viral rebounds following ART interruption. We may use the second-stage model (6) to evaluate such possible associations. Note that the random effects $\mathbf{b}_i$ in NLME model (2) for viral load data during ART may be viewed as individual-specific characteristics of the viral load trajectories during ART. Thus, we may use the random effects $\mathbf{b}_i$ as "covariates" in the rebound model (6) to see if these "covariates" may partially explain the large variations in the individual-specific parameters $\boldsymbol{\beta}_i$ during viral rebound. Similarly, we may consider the random effects $\mathbf{a}_i$ in the CD4 model (4) during ART and use them as possible "covariates" in the NLME model (6) for viral load following ART

interruption. Specifically, in the NLME model for viral rebound, we may consider the following second-stage model for (6)

$$\begin{aligned} \beta_{ki} &= \beta_k + \gamma_{k1} b_{1i} + \gamma_{k2} b_{2i} + \gamma_{k3} b_{3i} + \gamma_{k4} b_{4i} + \gamma_{k5} a_{1i} + \gamma_{k6} a_{2i} + \gamma_{k7}^T \mathbf{v}_i^* + \tau_{ki}, \\ \beta_{k'i} &= \beta_{k'} + \tau_{k'i}, \quad k' \neq k, \quad k', k = 1, \ldots, 4, \quad i = 1, \ldots, n, \end{aligned} \tag{7}$$

where $\beta_k$'s are fixed effects parameters, $\gamma_{kl}$'s are fixed effect parameters associated with the corresponding random effects respectively, and $\mathbf{v}_i^*$ denotes other baseline covariates.

For the motivating dataset shown in Figure 1, the sample size is small. In this case, we may simplify the second-stage model (7) to reduce the number of parameters and only focus on the key features of viral decay and viral rebound. For example, to see if the initial viral decay rate $\lambda_{1i} = \lambda_1 + b_{3i}$ in NLME model (2) during ART may be associated with the values of setpoints $\beta_{1i}$ after viral rebound, we may consider the following second-stage model

$$\beta_{1i} = \beta_1 + \gamma_{13} b_{3i} + \tau_{1i}, \quad \beta_{ki} = \beta_k + \tau_{ki}, \quad k = 2, 3, 4. \tag{8}$$

Then, testing $H_0$: $\gamma_{13} = 0$ vs. $H_1$: $\gamma_{13} \neq 0$ allows us to assess possible association between $b_{3i}$ and $\beta_{1i}$. Similarly, we can evaluate other possible associations.

## Parameter estimation and inference

In the previous section, we describe two NLME models and an LME model for viral load and CD4 longitudinal data before and following ART interruption respectively. A challenge in data analysis is that some viral loads are *left censored* in the later period during ART and in the early period after ART interruption. In other words, both NLME models for viral loads must incorporate left censored data. In this section, we consider a likelihood method for parameter estimation and inference, incorporating left censoring. Note that we may consider the (joint) likelihood for all observed data under the three models. However, such a unified approach can be computationally very intensive, as will be discussed later in this section. To circumvent the computational burden, here we consider likelihood methods for the three models separately, but accounting for shared random effects linking these models. Then we propose a simple three-step method for parameter estimation and inference.

We first consider the NLME model (2) for viral dynamics during ART. Suppose that the lower detection limit of viral load is $d_i$ for subject $i$, i.e., viral load values smaller than $d_i$ cannot be observed. The observed value of viral load $y_{ij}$ for individual $i$ at time $t_{ij}$ can then be written as $\left(y_{ij}^o, c_{ij}\right)$, where $c_{ij}$ is the censoring indicator such that $y_{ij} = y_{ij}^o$ is observed if $c_{ij} = 0$ and $y_{ij}$ is left censored if $c_{ij} = 1$, $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, n_i$. Let $\mathbf{c}_i = (c_{i1}, \ldots, c_{in_i})^{\mathrm{T}}$, let $\mathbf{y}_i^o$ denote the observed components of $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^{\mathrm{T}}$, and let $\mathbf{y}_{\mathrm{cen},i}$ denote the censored components of $\mathbf{y}_i$. The observed data of viral load before ART interruption are $\{\left(\mathbf{y}_i^o, \mathbf{c}_i\right),\ i = 1, \ldots, n\}$. Let $f(\cdot)$ denote a generic density function and $F(\cdot)$ denote the corresponding cumulative density function (cdf). Let $\boldsymbol{\theta}_1$ be the collection of all unknown parameters in the NLME model (2). The likelihood for the observed viral load data during ART based on the NLME model (2) can be written as

$$\begin{aligned} L_o(\boldsymbol{\theta}_1) &= \prod_{i=1}^{n} \int \left\{ \prod_{j=1}^{n_i} \left( f(y_{ij} | \mathbf{b}_i, \boldsymbol{\theta}_1) \right)^{1-c_{ij}} \left( F(d_i \mid \mathbf{b}_i, \boldsymbol{\theta}_1) \right)^{c_{ij}} \right\} f(\mathbf{b}_i | B)\, \mathrm{d}\mathbf{b}_i \\ &= \prod_{i=1}^{n} \iint \left\{ \prod_{j=1}^{n_i} f(y_{ij} | \mathbf{b}_i, \boldsymbol{\theta}_1) f(\mathbf{b}_i | B) \right\} \mathrm{d}\mathbf{y}_{\mathrm{cen},i}\, \mathrm{d}\mathbf{b}_i. \end{aligned}$$

We see that the likelihood $L_o(\boldsymbol{\theta}_1)$ involves an intractable integration, since the dimension of $(\mathbf{y}_{\mathrm{cen},i}, \mathbf{b}_i)$ is high and the model is nonlinear.

To evaluate $L_o(\boldsymbol{\theta}_1)$, a commonly used method is the MCEM algorithm (Wei and Tanner 1990), treating the left censored values $\mathbf{y}_{\mathrm{cen},i}$ and random effects $\mathbf{b}_i$ as "missing data" (Hughes 1999; Wu 2002). Specifically, the E-step at the $k$th EM iteration can be written as

$$Q(\theta_1|\theta_1^{(k)}) = \iint \left[\log\ f(\mathbf{y}_i|\mathbf{b}_i, \theta_1^{(k)}) + \log\ f(\mathbf{b}_i|\theta_1^{(k)})\right] f(\mathbf{y}_i, \mathbf{b}_i|\mathbf{y}_i^o, \mathbf{c}_i, \theta_1^{(k)})\ \mathrm{d}\mathbf{y}_{\mathrm{cen},i}\ \mathrm{d}\mathbf{b}_i,$$

where $\theta_1^{(k)}$ is the parameter estimate from previous EM iteration ($k = 1, 2, \ldots$). To evaluate $Q(\theta_1|\theta_1^{(k)})$ in the E-step, we can use Monte Carlo methods to simulate a large number of "missing data" $(\mathbf{y}_{\mathrm{cen},i}, \mathbf{b}_i)$ from the conditional distribution $f(\mathbf{y}_i, \mathbf{b}_i|\mathbf{y}_i^o, \mathbf{c}_i, \theta_1^{(k)})$, and then approximate $Q(\theta_1|\theta_1^{(k)})$ by the empirical mean based on the simulated values. This simulation step can be implemented by Gibbs sampler along with rejection sampling methods or the importance sampling method or other Markov Chain Monte Carlo (MCMC) methods (Wu 2009). The M-step can be based on the Newton-Raphson method. This MCEM algorithm can be computationally intensive since the dimension of the "missing data" $(\mathbf{y}_{\mathrm{cen},i}, \mathbf{b}_i)$ can be high, so simulating large numbers from the conditional distribution $f(\mathbf{y}_i, \mathbf{b}_i|\mathbf{y}_i^o, \mathbf{c}_i, \theta_1^{(k)})$ can be very slow.

Alternatively, a computationally more efficient method than the MCEM algorithm is the stochastic approximation EM (SAEM) algorithm (Delyon, Lavielle, and Moulines 1999). The SAEM algorithm replaces the E-step of the MCEM algorithm by a *single* draw from the conditional distribution $f(\mathbf{y}_i, \mathbf{b}_i|\mathbf{y}_i^o, \mathbf{c}_i, \theta_1^{(k)})$ based on an MCMC method, and then use a stochastic approximation to update $Q(\theta_1|\theta_1^{(k)})$. Delyon, Lavielle, and Moulines (1999) shows theoretically that SAEM converges to a (local) maximum of the likelihood under general conditions. The SAEM algorithm for NLME models has been implemented in the software "Monolix" (Comets, Lavenu, and Lavielle 2017; Kuhn and Lavielle 2005). Samson, Lavielle, and Mentré (2006) extends the SAEM method to NLME models with left censoring, based on simulating the left-censored values $\mathbf{y}_{\mathrm{cen},i}$ from a right-truncated Gaussian distribution $f(\mathbf{y}_{\mathrm{cen},i}|\mathbf{y}_i^o, \mathbf{b}_i, \theta_1^{(k)})$ based on the Gibbs sampling in the E-step of the SAEM algorithm.

Similarly, the foregoing SAEM method can be used for the NLME model (5) and (6) for viral rebound data following ART interruption. In fact, we may consider the SAEM algorithm for all three models simultaneously based on the joint likelihood of all observed data. However, such a joint likelihood method can be difficult to implement and computationally extremely intensive, since the dimension of the "missing data" $(\mathbf{y}_{\mathrm{cen},i}, \mathbf{w}_{\mathrm{cen},i}, \mathbf{b}_i, \boldsymbol{\tau}_i, \mathbf{a}_i)$ is very high so even a single simulation using an MCMC method can be computationally over-whelming, where $\mathbf{w}_{\mathrm{cen},i}$ denotes the censored components of $\mathbf{w}_i = \left(w_{i1}, w_{i2}, \ldots, w_{in_i^*}\right)^{\mathrm{T}}$. Therefore, here we propose to use SAEM for each NLME model with left censoring separately to reduce the computation burden. Specifically, we propose the following *three-step (TS) method*:

- Step 1: For data during ART, fit the NLME model (2) for viral load with censoring using the above SAEM algorithm and fit the LME model for CD4 using the standard method respectively, and then obtain the maximum likelihood estimates (MLEs) of the fixed parameters and the empirical Bayes estimates of the random effects, $\hat{\mathbf{b}}_i$ and $\hat{\mathbf{a}}_i$, respectively;
- Step 2: For viral rebound data following ART interruption, fit the NLME model (5) with left censoring using the above SAEM algorithm, with the random effects $\mathbf{a}_i$ and $\mathbf{b}_i$ in the second-stage model (7) substituted by their empirical Bayes estimates $\hat{\mathbf{b}}_i$ and $\hat{\mathbf{a}}_i$ from Step 1;
- Step 3: Obtain the standard errors of the parameter estimates based on a (parametric) bootstrap method, which incorporates the estimation uncertainty of random effect estimates $\hat{\mathbf{b}}_i$ and $\hat{\mathbf{a}}_i$ in Step 1.

The parametric bootstrap method works as follows:

1. Simulate CD4 and viral load data with left censoring, based on the fitted LME and two NLME models using the above three-step method, where the model parameters are replaced by their estimates.
2. For the simulated CD4 and viral load data, fit all three models again using the above three-step method and obtain all parameter estimates.
3. Repeat the above process $B$ times (say, $B = 100$), we obtain $B$ estimates for each parameter. The sample standard deviation of these $B$ estimates of each parameter is the parametric *bootstrap estimate* of standard error of the corresponding parameter estimate.

The above bootstrap method incorporates the estimation uncertainty of the parameter and random effect estimates in the TS method with separate model fitting, so it should produce more reliable standard errors of the parameter estimates than those from separate model fitting. Note that, in Step 2 of the above TS method,

we use the empirical Bayes estimates of the random effects. An alternative approach is to sample from the posterior distribution of the random effects, but this approach may introduce additional variability from the sampling.

## Results of data analysis

In this section, we analyze the dataset shown in Figure 1 using the proposed TS method and a naive (NV) method, which still uses the SAEM algorithm but the censored values are substituted by half the detection limit and inference is based on model-based standard errors. There are 75 patients in the study. Viral loads and CD4 are repeatedly measured on patients during ART and following ART interruption, with the longest time of 58.4 months and shortest time of 16.73 months. After ART interruption, viral load usually increases to a peak within 6–10 weeks, then decrease to a stable level over a time scale of months. Therefore, we restrict our attention to data within week 36 (9 months). We excluded individuals with 2 or less repeated measurements either during ART or following ART interruption (n = 5) or individuals with un-suppressed viral load values at the first time point following ART interruption (n = 4). For remaining viral load data during ART, the minimum number of repeated measurements is 5, and the maximum number of repeated measurements is 19, with an average number of repeated measurements being 9.32. For viral load data following ART interruption, the minimum number of repeated measurements is 3, and the maximum number of repeated measurements is 23, with an average number of repeated measurements being 6.63. The proportions of viral load measurements that are left censored (below the detection limit) before or after ART interruption are 59.6% or 23.9%, respectively. From Figure 1, we can see that the viral load trajectories during ART exhibit clear patterns of viral decay. Following ART interruption, the viral loads rebound quickly, but their trajectories become complicated after reaching peak points, with substantial between-subject variations.

For viral load data during ART, we fit the NLME model (2) of Wu and Ding (Wu and Ding 1999), with left-censored data addressed by the SAEM method. The bi-exponential decay NLME model (2) fits the viral load data very well. Figure 3A (top four figures) shows the fitted values vs. the corresponding observed values for four randomly selected subjects during ART. For the CD4 data during ART, the CD4 trajectories do not appear to exhibit clear patterns, with large between-subject variations, possibly due to substantial measurement errors. However, there seems an overall upward trend. Thus, we fit the LME model (4) to the CD4 data, which captures a rough upward trend before ART interruption.

For viral load data following ART interruption, we consider the following NLME model

$$\begin{aligned} w_{ij} &= \beta_{1i}\frac{t_{ij}}{t_{ij}+\exp(\beta_{2i}-\beta_{3i}t_{ij})}+\beta_{4i}+\xi_{ij},\\ \beta_{1i} &= \beta_1+\gamma_{11}b_{1i}+\gamma_{12}b_{2i}+\gamma_{13}b_{3i}+\gamma_{14}b_{4i}+\gamma_{15}a_{1i}+\gamma_{16}a_{2i}+\tau_{1i},\\ \beta_{ki} &= \beta_k+\tau_{ki},\ k>1, \end{aligned} \tag{9}$$

where $\beta_{1i}$ is the viral set point during rebound and the random effects $(b_{1i}, b_{2i}, b_{3i}, b_{4i}, a_{1i}, a_{2i})$ are defined in model (7) (e.g., $b_{3i}$ is the random effect associated with the initial viral decay rate $\lambda_{1i}$ during viral decay before ART interruption). We again address left-censored data by the SAEM method. We use the parametric bootstrap method with $B = 100$ to estimate the standard errors of all the fixed effect parameter estimates. This NLME model for viral rebound also fits the data reasonably well. Figure 3B (bottom four figures) shows the fitted values vs. the corresponding observed values for four randomly selected subjects during viral rebound.

Table 1 shows parameter estimation results for models (2) and (9) (the time unit in data analysis is month). We see that parameters $\gamma_{13}$ and $\gamma_{14}$, which link the initial viral decay rates $\lambda_{1i}$ and $\lambda_{2i}$ during ART to viral setpoints $\beta_{1i}$ following ART interruption for individual $i$, suggest an association among these features (p-value = 0.029 and < 0.001 respectively). Specifically, the initial viral decay rates during ART appears to be negatively associated with the viral setpoints following ART interruption: the faster the viral decay after start of ART, the lower the setpoints following ART interruption. In addition, the second-phase viral decay rates during ART appears to be negatively associated with the viral setpoints following ART interruption: the

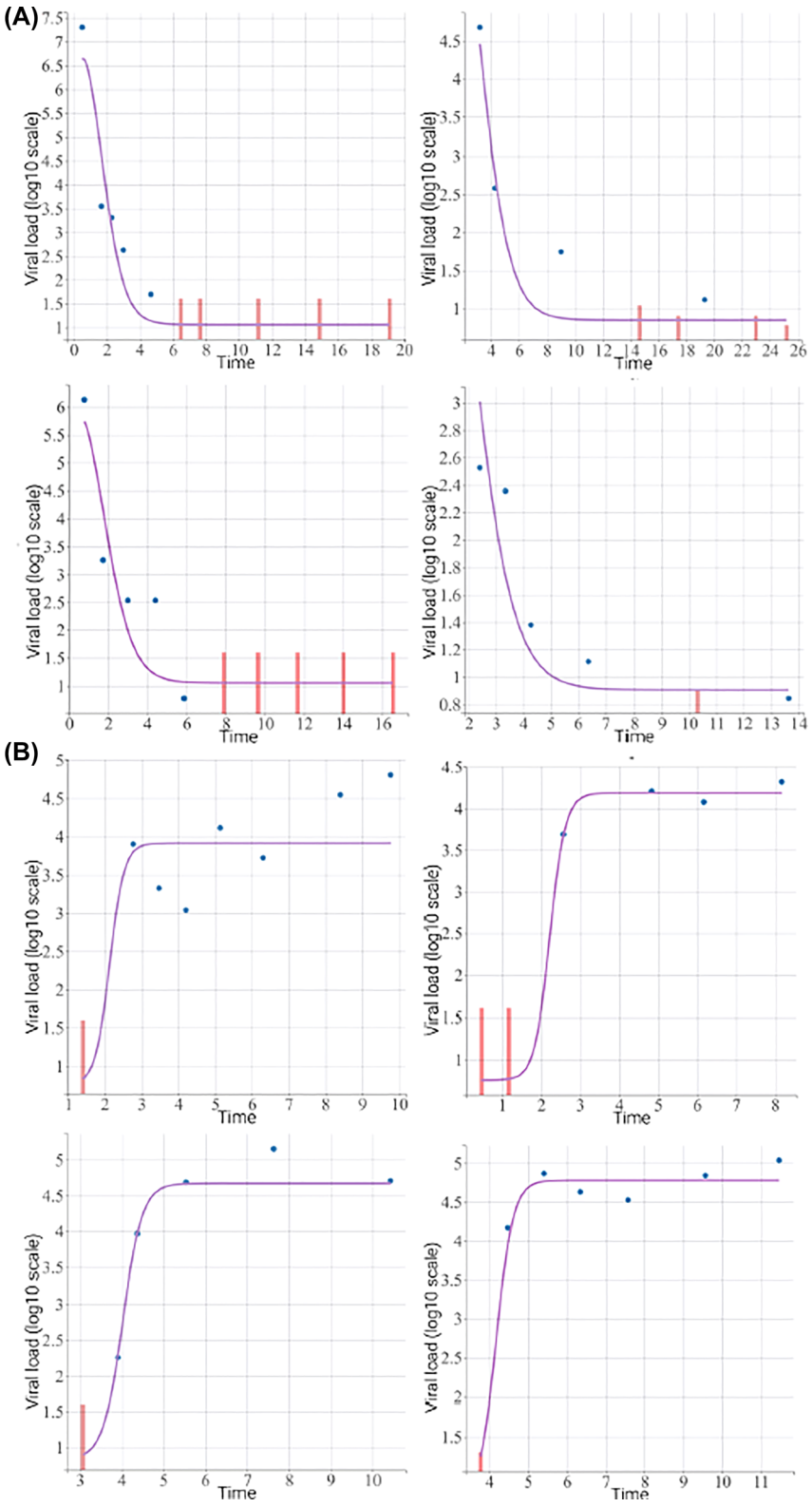


**Figure 3:** Observed and fitted viral load trajectories before (top four figures) and following (bottom four figures) ART interruption for 4 randomly selected subjects respectively. The red vertical bars represent left-censored viral loads.

**Table 1:** Parameter estimates with a second-stage model for setpoint $\beta_{1i}$.

| Parameter | Estimate | Naive SE | Bootstrap SE | *z*-Value | p-Value |
|---|---|---|---|---|---|
| $P_1$ | 11.258 | 0.283 | 0.167 | 67.449 | 0.000 |
| $\lambda_1$ | 4.790 | 0.380 | 0.362 | 13.244 | 0.000 |
| $P_2$ | 3.271 | 0.117 | 0.206 | 15.872 | 0.000 |
| $\lambda_2$ | 0.225 | 0.027 | 0.019 | 11.788 | 0.000 |
| $\alpha_1$ | 0.087 | 0.007 | 0.017 | 5.178 | 0.000 |
| $\alpha_2$ | 24.080 | 0.517 | 0.438 | 54.985 | 0.000 |
| $\beta_1$ | 2.828 | 0.161 | 0.243 | 11.651 | 0.000 |
| $\gamma_{11}$ | 0.144 | 0.073 | 0.075 | 1.911 | 0.056 |
| $\gamma_{12}$ | −0.027 | 0.559 | 0.755 | −0.035 | 0.972 |
| $\gamma_{13}$ | −0.252 | 0.079 | 0.116 | −2.177 | 0.029 |
| $\gamma_{14}$ | −97.574 | 50.944 | 26.294 | −3.711 | 0.000 |
| $\gamma_{15}$ | −0.026 | 0.033 | 0.035 | −0.728 | 0.467 |
| $\gamma_{16}$ | 0.811 | 1.290 | 4.291 | 0.189 | 0.850 |
| $\beta_2$ | 1.588 | 1.308 | 0.694 | 2.289 | 0.022 |
| $\beta_3$ | 3.360 | 1.614 | 0.828 | 4.056 | 0.000 |
| $\beta_4$ | 0.783 | 0.119 | 0.151 | 5.195 | 0.000 |

Naive SE is the standard error based on separate model fitting without bootstrap, *z*-Value is the ratio of estimate/bootstrap SE, and p-Value is based on the *z*-Value and the standard normal tail probability for a two-sided test.

faster the viral decay in the slow decay phase during ART, the lower the setpoints following ART interruption. The naive method produces similar estimates but different standard errors (we only show naive SE in Table 1 since naive estimates are similar). We will evaluate the two methods via simulation in the next section.

We may also consider the following second stage model associated with the NLME model (9) for viral rebound following ART interruption:

$$\begin{aligned} &\beta_{3i} = \beta_3 + \gamma_{31} b_{1i} + \gamma_{32} b_{2i} + \gamma_{33} b_{3i} + \gamma_{34} b_{4i} + \gamma_{35} a_{1i} + \gamma_{36} a_{2i} + \tau_{3i}, \\ &\beta_{ki} = \beta_k + \tau_{ki}, \ k \neq 3. \end{aligned} \tag{10}$$

**Table 2:** Parameter estimates with a second-stage model for rebound rate $\beta_{3i}$.

| Parameter | Estimate | Naive SE | Bootstrap SE | *z*-Value | p-Value |
|---|---|---|---|---|---|
| $P_1$ | 11.258 | 0.283 | 0.204 | 55.176 | 0.000 |
| $\lambda_1$ | 4.790 | 0.380 | 0.236 | 20.321 | 0.000 |
| $P_2$ | 3.271 | 0.117 | 0.167 | 19.575 | 0.000 |
| $\lambda_2$ | 0.225 | 0.027 | 0.019 | 11.627 | 0.000 |
| $\alpha_1$ | 0.087 | 0.007 | 0.011 | 7.765 | 0.000 |
| $\alpha_2$ | 24.080 | 0.517 | 0.518 | 46.456 | 0.000 |
| $\beta_1$ | 2.946 | 0.375 | 0.127 | 23.196 | 0.000 |
| $\beta_2$ | 1.542 | 1.040 | 0.419 | 3.680 | 0.000 |
| $\beta_3$ | 3.280 | – | 0.448 | 7.327 | 0.000 |
| $\gamma_{31}$ | 0.837 | 0.053 | 0.211 | 3.971 | 0.000 |
| $\gamma_{32}$ | −4.589 | 0.177 | 2.691 | −1.705 | 0.088 |
| $\gamma_{33}$ | −1.110 | 0.008 | 0.317 | −3.501 | 0.000 |
| $\gamma_{34}$ | −221.564 | 12.712 | 116.409 | −1.903 | 0.057 |
| $\gamma_{35}$ | −0.106 | 0.023 | 0.109 | −0.969 | 0.333 |
| $\gamma_{36}$ | −0.090 | 0.445 | 8.458 | −0.011 | 0.991 |
| $\beta_4$ | 0.772 | 0.448 | 0.057 | 13.625 | 0.000 |

Naive SE is the standard error based on separate model fitting without bootstrap, *z*-Value is the ratio of estimate/bootstrap SE, and p-Value is based on the *z*-Value and the standard normal tail probability for a two-sided test. A naive SE is unavailable, possibly due to parameters being unidentifiable.

The analysis results are presented in Table 2. We see that the rate of rise during viral rebound following ART interruption $\beta_{3i}$ appears to be negatively correlated with initial viral decay rate ($\gamma_{33}$) and positively correlated with initial viral load values ($\gamma_{31}$). That is, the faster the initial viral decline during ART or the lower the initial viral loads, the slower the viral rising following ART interruption.

In summary, the analysis results show that some key characteristics of the viral load trajectories during ART, especially the initial viral decay rates after the start of ART, appear to be associated with some important features of the viral rebound following ART interruption, such as viral setpoints and rates of viral rising. The CD4 data during ART do not seem to be associated with important features of the viral rebound following ART interruption.

## Simulation study

In this section, we conduct extensive simulations to evaluate the proposed TS method and compare it with the naive method used in data analysis. We choose similar NLME models to those in the data analysis section, but we omit the CD4 model for simplicity. The true values of the model parameters are set to be similar to those estimated in the data analysis section. The sample size is set to be n = 50 individuals. For the within-individual longitudinal measurements, to mimic the real dataset, for half the sample we choose 10 repeated measurements during ART and 9 repeated measurements following ART interruption, while for the remaining half the sample we choose 11 repeated measurements during ART and 11 repeated measurements following ART interruption. The measurement times are chosen to be similar to those in the real dataset. Two sets of measurement times during ART are $t_1 = (0.5, 1.6, 2.3, 3, 4.6, 6.5, 7.6, 11.2, 14.9, 19.1)$ and $t_2 = (0.5, 0.7, 1.7, 3, 4.4, 5.9, 7.9, 9.6, 11.7, 14, 16.5)$. Two sets of measurement times following ART interruption are $t_1^* = (0.1, 1.1, 1.6, 2.4, 2.8, 3.3, 3.7, 4.2, 5.2)$ and $t_2^* = (0.2, 0.6, 1.1, 1.6, 2.1, 2.5, 3, 3.5, 4, 4.4, 4.9)$.

We generate the viral load data during ART based on the following NLME model

$$\begin{aligned} y_{ij} &= \log_{10}\left(e^{P_{1i}-\lambda_{1i}t_{ij}} + e^{P_{2i}-\lambda_{2i}t_{ij}}\right) + e_{ij} \\ P_{1i} &= P_1 + b_{1i}, \quad P_{2i} = P_2 + b_{2i}, \quad \lambda_{1i} = \lambda_1 + b_{3i}, \quad \lambda_{2i} = \lambda_2 + b_{4i}, \end{aligned} \tag{11}$$

where $e_{ij}$ i.i.d. $\sim N\left(0, \sigma_1^2\right)$ and $\mathbf{b}_i = (b_{1i}, b_{2i}, b_{3i}, b_{4i})^{\mathrm{T}} \sim N(0, D)$. The true values are $P_1 = 17.0$, $P_2 = 2.6$, $\lambda_1 = 4$, $\lambda_2 = 0.05$, and $\sigma_1 = 0.5$. The detection limit is set to be $d = 1.60$. For the viral load data following ART interruption, we generate the data based on the following NLME model

$$\begin{aligned} w_{ij} &= \beta_{1i}\frac{t_{ij}}{t_{ij} + \exp(\beta_{2i} - \beta_{3i}t_{ij})} + \beta_{4i} + \xi_{ij}, \\ \beta_{1i} &= \beta_1 + b_{3i}\gamma_3 + \tau_{1i}, \quad \beta_{2i} = \beta_2 + \tau_{2i}, \quad \beta_{3i} = \beta_3 + \tau_{3i}, \quad \beta_{4i} = \beta_4 + \tau_{4i}, \end{aligned}$$

where $b_{3i}$ is the random effect from model (11), $\xi_{ij} \sim N\left(0, \sigma_3^2\right)$, and $\tau_i \sim N(0, G)$. The true parameter values are $\beta_1 = 3.2$, $\beta_2 = 5.6$, $\beta_3 = 10$, $\beta_4 = 1$, $\gamma_3 = 1$, $\sigma_3 = 0.5$,

$$D = \begin{bmatrix} 1.7 & -0.4 & 0.06 & -0.003 \\ -0.4 & 1.5 & -0.1 & 0.005 \\ 0.06 & -0.1 & 0.05 & -0.002 \\ -0.003 & 0.005 & -0.002 & 0.0002 \end{bmatrix}, \quad \text{and} \quad G = \begin{bmatrix} 0.5 & 0.03 & 0.2 & 0.03 \\ 0.03 & 2.3 & -0.3 & -0.04 \\ 0.2 & -0.3 & 11.8 & 0.06 \\ 0.03 & -0.04 & 0.06 & 0.006 \end{bmatrix}.$$

We evaluate the proposed TS method based on bias, mean square error (MSE), and coverage rates of 95% confidence intervals. For a parameter $\beta$ and its estimate $\hat{\beta}$, the bias and root MSE (rMSE) are defined as bias $= E(\hat{\beta}) - \beta$, rMSE $= \sqrt{\text{MSE}}$, and the coverage rate is the proportion of confidence intervals which covers the true value. We compare the TS method to a naive method which replaces censored viral loads by half the

**Table 3:** Simulation results.

| Parameter | True value | Method | Estimate | SE | Bias | rMSE | Coverage |
|---|---|---|---|---|---|---|---|
| $P_1$ | 17.0 | TS | 17.097 | 0.425 | 0.097 | 0.435 | 0.95 |
| | | Naive | 17.088 | 0.257 | 0.088 | 0.272 | 0.91 |
| $\lambda_1$ | 4.0 | TS | 4.092 | 0.240 | 0.092 | 0.257 | 0.91 |
| | | Naive | 4.137 | 0.216 | 0.137 | 0.256 | 0.91 |
| $P_2$ | 2.6 | TS | 2.794 | 0.398 | 0.194 | 0.442 | 0.95 |
| | | Naive | 2.222 | 0.111 | −0.378 | 0.394 | 0.39 |
| $\lambda_2$ | 0.1 | TS | 0.040 | 0.081 | −0.010 | 0.081 | 0.95 |
| | | Naive | 0.029 | 0.009 | −0.021 | 0.023 | 0.40 |
| $\beta_1$ | 3.2 | TS | 3.267 | 0.140 | 0.067 | 0.155 | 0.95 |
| | | Naive | 3.324 | 0.130 | 0.124 | 0.180 | 0.86 |
| $\gamma_3$ | 1.0 | TS | 0.986 | 0.096 | −0.014 | 0.097 | 0.94 |
| | | Naive | 0.999 | 0.083 | −0.001 | 0.083 | 0.92 |
| $\beta_2$ | 5.6 | TS | 5.588 | 1.187 | −0.012 | 1.187 | 0.94 |
| | | Naive | 6.370 | 1.354 | 0.770 | 1.558 | 0.84 |
| $\beta_3$ | 10.0 | TS | 10.079 | 2.062 | 0.079 | 2.064 | 0.94 |
| | | Naive | 11.170 | 2.650 | 1.170 | 2.897 | 0.88 |
| $\beta_4$ | 1.0 | TS | 0.930 | 0.092 | −0.070 | 0.115 | 0.85 |
| | | Naive | 0.861 | 0.072 | −0.139 | 0.157 | 0.62 |

detection limits and uses the model-based standard errors for inference. For the TS method, the number of bootstrap samples is $B = 100$. The simulations are repeated 100 times. While a larger number of repetitions may be more desirable, the computation involving bootstrap is intensive and the 100 repetitions results seem sufficient to allow us to make reasonable conclusions (Morris, White, and Crowther 2019).

The simulation results are shown in Table 3. We see that the proposed TS method performs quite well and clearly outperforms the naive method: estimates based on the TS method are approximately unbiased with estimated coverage probabilities close to the nominal level 0.95, while estimates based on the naive method may sometimes produce biased results with estimated coverage probabilities way below the nominal level 0.95. Despite the simulation results, both methods seem to produce similar results on the real data analysis in the previous section. Note that the SE's based on the naive method maybe underestimate the true variation since the naive method ignores the uncertainties of the censored values and the separate NLME model fitting, therefore the naive method may lead to smaller MSE's but lower coverage probabilities. On the other hand, the proposed TS method incorporates the uncertainty of the censored values by the SAEM algorithm and the separate NLME model fitting by bootstrap, so it may lead to larger MSE's but correct coverage probabilities.

Since the performance of MLEs of mixed effects models depends both on the sample size and the number of repeated measurements, we also conduct another simulation study by choosing more frequent repeated measurements, with other true parameter values remaining the same. The measurement times are chosen to be close to those in the real dataset with additional measurement times in between. Specifically, the new set of measurement times during ART are chosen to be $t_1 = (0.4, 1.2, 1.6, 2.1, 3.2, 4.6, 5.3, 7.8, 10.4, 13.4, 17)$, and the new set of measurement times following ART interruption are chosen to be $t_1^* = (0, 0.6, 0.8, 1.2, 1.4, 1.7, 1.9, 2.1, 2.6, 3.1, 3.8, 4.5, 5.4, 6)$. The simulation results are shown in Table 4. The proposed TS method again outperforms the naive method.

**Table 4:** Simulation results with more frequent repeated measurements.

| Parameter | True value | Method | Estimate | SE | Bias | rMSE | Coverage |
|---|---|---|---|---|---|---|---|
| $P_1$ | 17.0 | TS | 16.820 | 0.733 | −0.180 | 0.755 | 0.93 |
| | | Naive | 16.981 | 0.259 | −0.019 | 0.260 | 0.89 |
| $\lambda_1$ | 4.0 | TS | 4.011 | 0.288 | 0.011 | 0.288 | 0.92 |
| | | Naive | 4.103 | 0.202 | 0.103 | 0.227 | 0.85 |
| $P_2$ | 2.6 | TS | 2.968 | 0.705 | 0.368 | 0.796 | 0.95 |
| | | Naive | 2.273 | 0.142 | −0.327 | 0.357 | 0.51 |
| $\lambda_2$ | 0.1 | TS | 0.096 | 0.186 | 0.046 | 0.191 | 0.95 |
| | | Naive | 0.041 | 0.012 | −0.009 | 0.015 | 0.34 |
| $\beta_1$ | 3.2 | TS | 3.223 | 0.132 | 0.023 | 0.134 | 0.94 |
| | | Naive | 3.304 | 0.121 | 0.104 | 0.159 | 0.87 |
| $\gamma_3$ | 1.0 | TS | 0.986 | 0.095 | −0.014 | 0.096 | 0.97 |
| | | Naive | 0.996 | 0.089 | −0.004 | 0.089 | 0.96 |
| $\beta_2$ | 5.6 | TS | 5.553 | 0.873 | −0.047 | 0.874 | 0.95 |
| | | Naive | 6.258 | 0.676 | 0.658 | 0.943 | 0.79 |
| $\beta_3$ | 10.0 | TS | 9.757 | 1.460 | −0.243 | 1.481 | 0.96 |
| | | Naive | 10.715 | 1.166 | 0.715 | 1.368 | 0.82 |
| $\beta_4$ | 1.0 | TS | 0.980 | 0.079 | −0.020 | 0.081 | 0.88 |
| | | Naive | 0.886 | 0.048 | −0.114 | 0.124 | 0.63 |

# Conclusions and discussion

We have shown that key features of viral decay during ART may be associated with important features of viral rebound following ART interruption. For example, the faster the viral decay after start of ART, the lower the setpoints following ART interruption. Such a finding may provide insights into HIV cure research. Recent findings suggest that HIV-1 latent reservoir is primarily established near the time of ART initiation (Abrahams et al. 2019); interventions in addition to ART to inhibit the formation of latent reservoir may subsequently lead to a lower viral set point – a key goal of the HIV functional cure. A limitation of our dataset is that the sample size is somewhat small.

In the future, if we are able to obtain a larger dataset, we may be able to identify more interesting associations between features of viral decay and viral rebound. Another issue is the frequencies of the repeated measurements within each individual. If the longitudinal data were collected more frequently over time, the mixed effects model parameters might be estimated more accurately in the sense of possibly smaller standard errors, allowing us to identify more interesting associations. It would also be of interest to investigate optimal study design, e.g., how to schedule measurement times, to improve efficiency of data analysis.

We have considered two NLME models with left censoring and an LME model, and we estimate the model parameters separately using an SAEM algorithm and a bootstrap method, called a three-step method. Similar ideas have appeared in the context of measurement error literature (Carroll et al. 2006). A major advantage of the proposed three-step method is that it is easy to implement and is computationally efficient. A disadvantage is that the parameter estimates may not be most efficient if the assumed models hold, since the model parameters are estimated separately. In addition, other possible useful covariates, such as the time from viral suppression to ART interruption, are not included in the models in order to keep the models relatively simple, and the possible association between the random effects in the viral load models and the CD4 model is ignored for simplicity, due to small sample sizes. One may also consider simultaneous likelihood inference based on the joint likelihood of all three models via an Monte Carlo EM algorithm (e.g., Wu 2009), but such a method can be computationally very intensive and may encounter convergence issues, since the dimensions of unobservable random effects and censored values are high. Alternatively, we may consider approximate joint likelihood inferences based on the so-called h-likelihood (Lee, Nelder, and Pawitan 2017) or based on Laplace approximations (Vonesh et al. 2002), but the accuracy of these approximations could be

a potential issue. Another promising approach is to use Bayesian methods (e.g., Dey, Chen, and Chang 1997; Huang et al. 2018), which will be investigated separately.

In this article, we have focused on studying the associations among the individual viral dynamic characteristics during ART and following ART interruption, such as the individual viral decay rates and setpoints. Another important direction is to study the association between the individual viral dynamic characteristics during ART and times to viral rebound or times to setpoints after ART is stopped. For example, are individuals with faster viral declines during ART associated with slower viral rebounds after the therapy is stopped? Such questions may be answered using joint inference for an NLME model for viral dynamics during ART and a time-to-event model such as a Cox proportional hazards model. There is an extensive literature on joint models for longitudinal and survival data and modelling times to viral rebound (e.g., Conway, Perelson, and Li 2019; Hill et al. 2016; Yu, Wu, and Gilbert 2018). We will explore this direction separately.

The models can be extended in different ways. For example, we may consider semiparametric NLME models for viral rebound data since the viral rebound trajectories after reaching peak points may not be easily modelled parametrically due to large between-individual variations without clear patterns. In the article, we have assumed that the left censored viral loads follow the same distribution as the observed viral loads. Such an assumption is not testable based on the observed data. We may consider an approach which does not make such assumption, e.g., treating the censored values as point masses as in Yu, Wu, and Gilbert (2018).

## Appendix: Model selection and diagnostics

In this section, we consider model selection and diagnostics for CD4 models. These viral load models are well studied in the literature (e.g., Wang et al. 2020; Wu and Ding 1999; Wu 2009). For the CD4 data, we considered log-transformation and square root transformation so that the transformed data are more compatible with the normality and constant variance assumptions. We find that these two transformations lead to similar results. More complex transformations, such as those based on the Box-Cox transformations, do not appear to improve the results substantially, and they are harder to interpret. Therefore, in the paper, we choose the square root transformation of CD4 counts since it is widely used in ACTG data analyses.

For CD4 model selections, since the CD4 model is secondary in the paper and CD4 data are measured with errors, we focus on the simplicity and goodness-of-fit of the candidate models. We find that a simple LME model captures the main features of the CD4 trajectories and it also fits the observed CD4 data reasonably well, i.e., in the paper, we choose the CD4 model:

$$\sqrt{\mathrm{CD4}_{ij}} = \alpha_{1i} + \alpha_{2i} t_{ij} + \varepsilon_{ij}, \quad \alpha_{ki} = \alpha_k + a_{ki}, \quad k = 1, 2,$$

where $\mathrm{CD4}_{ij}$ is the original CD4 count for subject $i$ measured at time $t_{ij}$, $\alpha_k$'s are fixed effects, and $a_{ki}$'s are random effects. Note that this simple CD4 model may also be interpreted as a classic measurement error model, where $z^*_{ij} = \alpha_{1i} + \alpha_{2i} t_{ij}$ may be interpreted as the unobserved true (transformed) CD4 value for subject $i$ at time $t_{ij}$. The AIC (BIC) values for the CD4 models with a quadratic term (and a random effect) and without a quadratic term are 2690 (2733) and 2733 (2759), respectively. Thus, adding a quadratic term $t^2_{ij}$ does not appear to improve the model substantially but it may make the model more complicated and less stable.

Figure 4 shows the observed/fitted CD4 values for four randomly selected subjects. We see that the CD4 model captures the main features of the CD4 trajectories. Figure 5 shows the normal QQ-plots for the estimated random effects in the intercepts and slopes of the CD4 model. We see that the normality assumptions are mostly reasonable. Figure 6 shows the overall residual plots of the CD4 model. These model diagnostics indicate that the simple CD4 model is a reasonable choice.

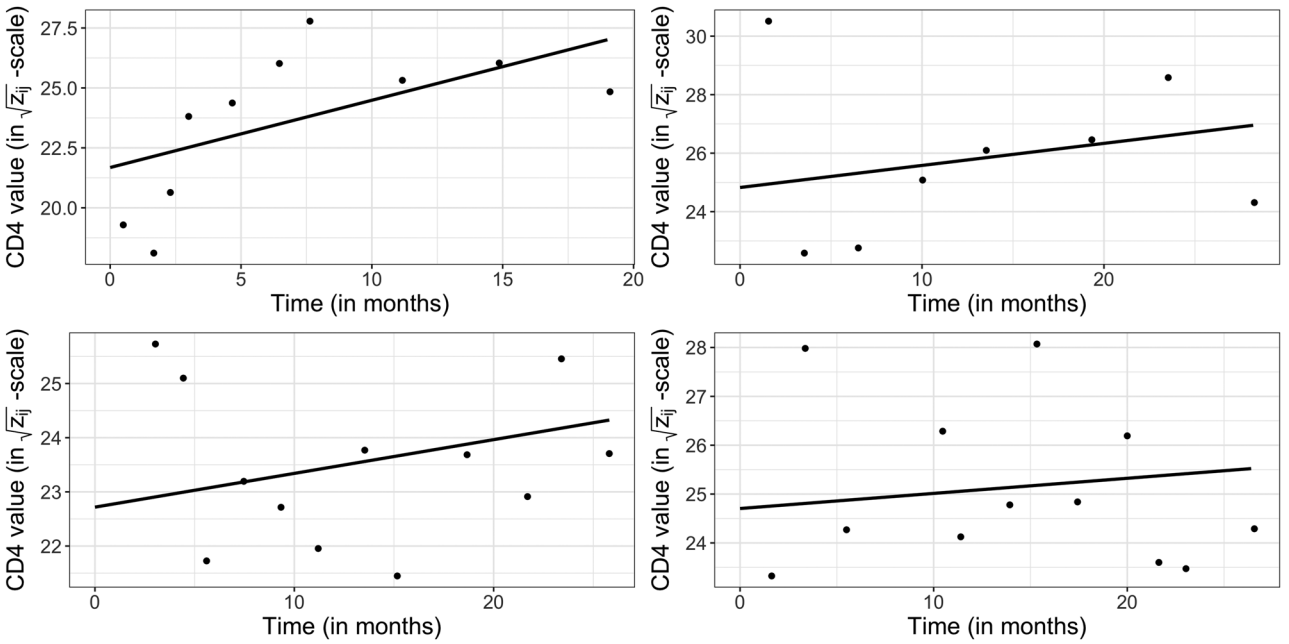


**Figure 4:** Observed/fitted CD4 values for four randomly selected subjects. The solid lines represent individual fitted CD4 models, and the dots represent observed CD4 values.

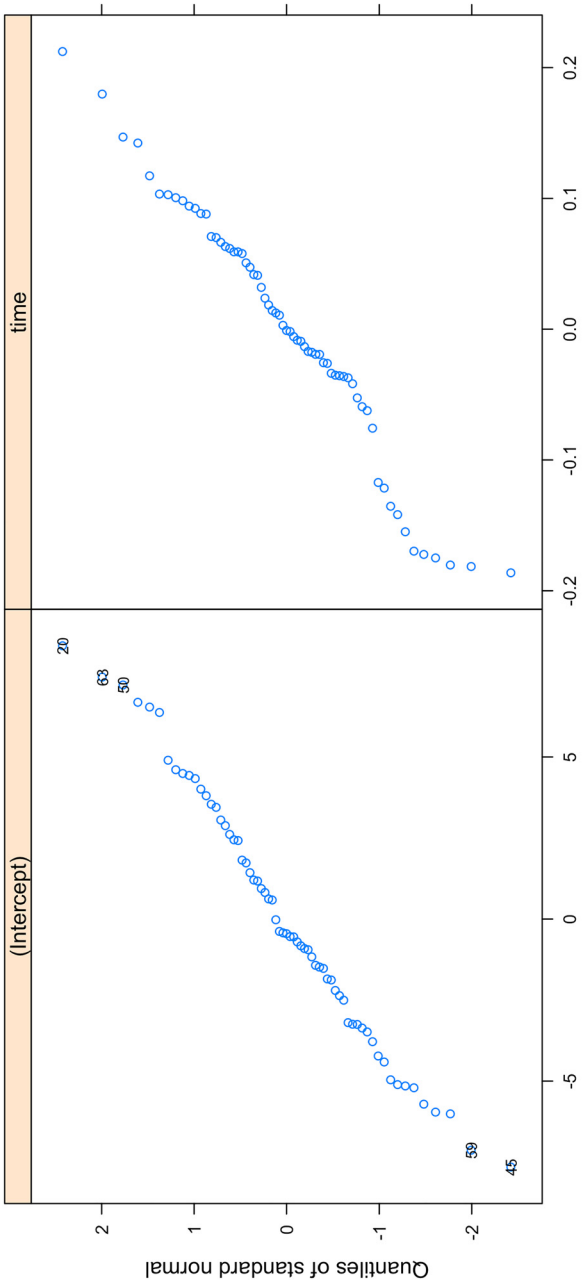


**Figure 5:** Normal QQ-plots for random effects in intercepts and slopes in the CD4 model.

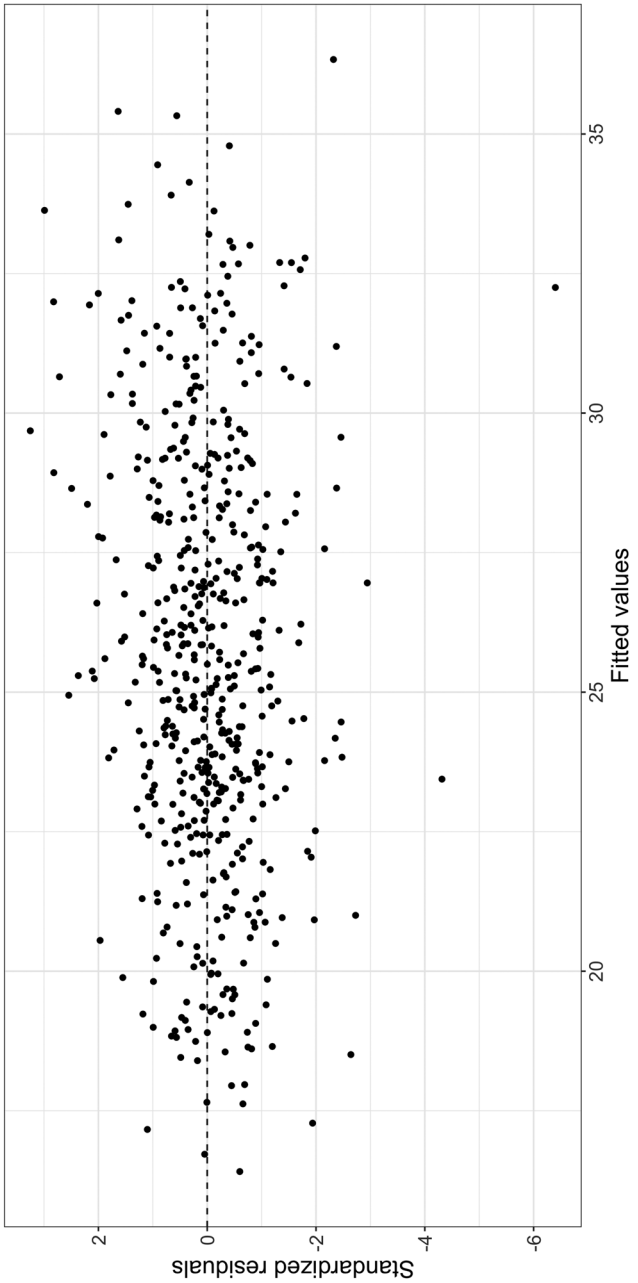


**Figure 6:** Residual plot for the CD4 model.

**Research funding:** We gratefully acknowledge grants from US National Institute of Allergy and Infectious Diseases P01 AI131385, R01 AI136947, an Ebert Career Development Award from Harvard Pilgrim Health Care Institute and Harvard Medical School, and the Natural Sciences and Engineering Research Council of Canada (NSERC) discovery grant 22R80742, and Swiss National Science Foundation, 179571.

## References

Abrahams, M. R., S. B. Joseph, N. Garrett, L. Tyers, M. Moeser, N. Archin, O. D. Council, D. Matten, S. Zhou, D. Doolabh, C. Anthony, N. Goonetilleke, S. Abdool Karim, D. M. Margolis, S. K. Pond, C. Williamson, and R. Swanstrom. 2019. "The Replication-competent HIV-1 Latent Reservoir is Primarily Established Near the Time of Therapy Initiation." *Science Translational Medicine* 11: 1–11.

Bing, A., Y. Hu, M. Prague, A. L. Hill, J. Z. Li, R. J. Bosch, V. DeGruttola, and R. Wang. 2020. "Comparison of Empirical and Dynamic Models for HIV Viral Load Rebound after Treatment Interruption." *Statistical Communications in Infectious Diseases* 12: 20190021

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. 2006. *Measurement Error in Nonlinear Models: a Modern Perspective*, 2nd ed. New York: Chapman and Hall/CRC.

Comets, E., A. Lavenu, and M. Lavielle. 2017. "Parameter Estimation in Nonlinear Mixed Effect Models Using Saemix, an R Implementation of the SAEM Algorithm." *Journal of Statistical Software* 80: i03–41.

Conway, J. M., A. S. Perelson, and J. Z. Li. 2019. "Predictions of Time to HIV Viral Rebound Following ART Suspension that Incorporate Personal Biomarkers." *PLoS Computational Biology* 15: e1007229.

Delyon, B., M. Lavielle, and E. Moulines. 1999. "Convergence of a Stochastic Approximation Version of the EM Algorithm." *Annals of Statistics* 27: 94–128.

Dey, D. K., M.-H. Chen, and H. Chang. 1997. "Bayesian Approach for Nonlinear Random Effects Models." *Biometrics* 53: 1239–52.

Gianella, S., V. von Wyl, M. Fischer, B. Niederoest, M. Battegay, E. Bernasconi, M. Cavassini, A. Rauch, B. Hirschel, P. Vernazza, R. Weber, B. Joos, H. F. Günthard, and The Swiss HIV Cohort Study 2011. "Effect of Early Antiretroviral Therapy during Primary HIV-1 Infection on Cell-Associated HIV-1 DNA and Plasma HIV-1 RNA." *Antiviral Therapy* 16: 535–45.

Hill, A. L., D. I. Rosenbloom, E. Goldstein, E. Hanhauser, D. R. Kuritzkes, R. F. Siliciano, and T. J. Henrich. 2016. "Real-time Predictions of Reservoir Size and Rebound Time during Antiretroviral Therapy Interruption Trials for HIV." *PLoS Pathogens* 12: e1005535.

Huang, Y., X. Lu, J. Chen, J. Liang, and M. Zangmeister. 2018. "Joint Model-based Clustering of Nonlinear Longitudinal Trajectories and Associated Time-to-event Data Analysis, Linked by Latent Class Membership: With Application to AIDS Clinical Studies." *Lifetime Data Analysis* 24: 699–718.

Hughes, J. P. 1999. "Mixed Effects Models with Censored Data with Application to HIV RNA Levels." *Biometrics* 55: 625–9.

Hurst, J., M. Hoffmann, M. Pace, J. P. Williams, J. Thornhill, E. Hamlyn, J. Meyerowitz, C. Willberg, K. K. Koelsch, N. Robinson, H. Brown, M. Fisher, S. Kinloch, D. A. Cooper, M. Schechter, G. Tambussi, S. Fidler, A. Babiker, J. Weber, A. D. Kelleher, R. E. Phillips, and J. Frater 2015. "Immunological Biomarkers Predict HIV-1 Viral Rebound after Treatment Interruption." *Nature Communications* 6: 8495.

Julg, B., L. Dee, J. Ananworanich, D. H. Barouch, K. Bar, M. Caskey, D. J. Colby, L. Dawson, K. L. Dong, K. Dubé, J. Eron, J. Frater, R. T. Gandhi, R. Geleziunas, P. Goulder, G. J. Hanna, R. Jefferys, R. Johnston, D. Kuritzkes, J. Z. Li, U. Likhitwonnawut, J. van Lunzen, J. Martinez-Picado, V. Miller, and L. J. Montaner 2019. "Recommendations for Analytical Antiretroviral Treatment Interruptions in HIV Research Trials—Report of a Consensus Meeting." *The Lancet HIV* 6: e259–e268.

Kuhn, E., and M. Lavielle. 2005. "Maximum Likelihood Estimation in Nonlinear Mixed Effects Models." *Computational Statistics & Data Analysis* 49: 1020–38.

Lee, Y., J. Nelder, and Y. Pawitan. 2017. *Generalized Linear Models with Random Effects: Unified Analysis via h-likelihood*, 2nd ed., vol. 153. Boca Raton: CRC Press/Taylor & Francis Group.

Li, J., B. Etemad, H. Ahmed, E. Aga, R. J. Bosch, J. W. Mellors, D. R. Kuritzkes, M. M. Lederman, M. Para, and R. T. Gandhi. 2016. "The Size of the Expressed HIV Reservoir Predicts Timing of Viral Rebound after Treatment Interruption." *AIDS* 30: 1–353.

Li, J. Z., D. M. Smith, and J. W. Mellors. 2015. "The Critical Roles of Treatment Interruption Studies and Biomarker Identification in the Search for an HIV Cure." *AIDS* 29: 1429.

Lindstrom, M. J., and D. M. Bates. 1990. "Nonlinear Mixed Effects Models for Repeated Measures Data." *Biometrics* 46: 673–87.

Morris, T. P., I. R. White, and M. J. Crowther. 2019. "Using Simulation Studies to Evaluate Statistical Methods." *Statistics in Medicine* 38: 2074–102.

Namazi, G., J. M. Fajnzylber, E. Aga, R. J. Bosch, E. P. Acosta, R. Sharaf, W. Hartogensis, J. M. Jacobson, E. Connick, P. Volberding, D. Skiest, D. Margolis, M. C. Sneller, S. J. Little, S. Gianella, D. M. Smith, D. R. Kuritzkes, R. M. Gulick, J. W. Mellors, V. Mehraj, R. T. Gandhi, R. Mitsuyasu, R. T. Schooley, K. Henry, P. Tebas, and S. G. Deeks 2018. "The Control

of HIV after Antiretroviral Medication Pause (CHAMP) Study: Posttreatment Controllers Identified from 14 Clinical Studies.'' *The Journal of infectious diseases* 218: 1954–63.

Noubary, F., and M. D. Hughes. 2012. ''Factors Affecting Timing of Antiretroviral Treatment Initiation Based on Monitoring CD4 Counts.'' *Journal of acquired immune deficiency syndromes* 61: 326.

Oxenius, A., A. R. McLean, M. Fischer, D. A. Price, S. J. Dawson, R. Hafner, C. Schneider, H. Joller, B. Hirschel, R. E. Phillips, R. Weber, and H. F. Günthard 2002. ''Human Immunodeficiency Virus-Specific $CD8^+$ T-cell Responses Do Not Predict Viral Growth and Clearance Rates during Structured Intermittent Antiretroviral Therapy.'' *Journal of Virology* 76: 10169–76.

Prague, M., J. M. Gerold, I. Balelli, C. Pasin, J. Z. Li, D. H. Barouch, J. B. Whitney, and A. L. Hill. 2019. ''Viral Rebound Kinetics Following Single and Combination Immunotherapy for HIV/SIV,'' BioRxiv: 700401.

Richman, D. D., D. M. Margolis, M. Delaney, W. C. Greene, D. Hazuda, and R. J. Pomerantz. 2009. ''The Challenge of Finding a Cure for HIV Infection.'' *Science* 323: 1304–7.

Rouzioux, C., L. Hocqueloux, and A. Sáez-Cirión. 2015. ''Posttreatment Controllers: What Do They Tell Us?'' *Current Opinion in HIV and AIDS* 10: 29–34.

Samson, A., M. Lavielle, and F. Mentré. 2006. ''Extension of the SAEM Algorithm to Left-Censored Data in Nonlinear Mixed-Effects Model: Application to HIV Dynamics Model.'' *Computational Statistics & Data Analysis* 51: 1562–74.

Thompson, M. A., J. A. Aberg, P. Cahn, J. S. Montaner, G. Rizzardini, A. Telenti, J. M. Gatell, H. F. Günthard, S. M. Hammer, M. S. Hirsch, D. M. Jacobsen, P. Reiss, D. D. Richman, P. A. Volberding, P. Yeni, and R. T. Schooley 2010. ''Antiretroviral Treatment of Adult HIV Infection: 2010 Recommendations of the International AIDS Society–USA Panel.'' *JAMA* 304: 321–33.

Vaida, F., A. P. Fitzgerald, and V. DeGruttola. 2007. ''Efficient Hybrid EM for Linear and Nonlinear Mixed Effects Models with Censored Response.'' *Computational Statistics & Data Analysis* 51: 5718–30.

Vaida, F., and L. Liu. 2009. ''Fast Implementation for Normal Mixed Effects Models with Censored Response.'' *Journal of Computational & Graphical Statistics* 18: 797–817.

Von Wyl, V., S. Gianella, M. Fischer, B. Niederoest, H. Kuster, M. Battegay, E. Bernasconi, M. Cavassini, A. Rauch, B. Hirschel, P. Vernazza, R. Weber, B. Joos, and H. F. Günthard. 2011. ''Early Antiretroviral Therapy During Primary HIV-1 Infection Results in a Transient Reduction of the Viral Setpoint upon Treatment Interruption.'' *PLoS One* 6: e27463.

Vonesh, E. F., H. Wang, L. Nie, and D. Majumdar. 2002. ''Conditional Second-order Generalized Estimating Equations for Generalized Linear and Nonlinear Mixed-effects Models.'' *Journal of the American Statistical Association* 97: 271–83.

Wang, R., A. Bing, C. Wang, Y. Hu, R. J. Bosch, and V. DeGruttola. 2020. ''A Flexible Nonlinear Mixed Effects Model for HIV Viral Load Rebound after Treatment Interruption.'' *Statistics in Medicine* 39: 2051–66.

Wei, G. C., and M. A. Tanner. 1990. ''A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms.'' *Journal of the American Statistical Association* 85: 699–704.

Wu, H., and A. A. Ding. 1999. ''Population HIV-1 Dynamics In Vivo: Applicable Models and Inferential Tools for Virological Data from Aids Clinical Trials.'' *Biometrics* 55: 410–8.

Wu, L. 2002. ''A Joint Model for Nonlinear Mixed-effects Models with Censoring and Covariates Measured with Error, with Application to AIDS Studies.'' *Journal of the American Statistical Association* 97: 955–64.

Wu, L. 2009. *Mixed Effects Models for Complex Data*. Boca Raton: Chapman & Hall/CRC Press.

Yu, T., L. Wu, and P. B. Gilbert. 2018. ''A Joint Model for Mixed and Truncated Longitudinal Data and Survival Data, with Application to HIV Vaccine Studies.'' *Biostatistics* 19: 374–90.

Zhao, L., T. Chen, V. Novitsky, and R. Wang. 2021. ''Joint Penalized Spline Modeling of Multivariate Longitudinal Data, with Application to HIV-1 RNA Load Levels and CD4 Cell Counts.'' *Biometrics* 77: 1061–74.