A framework for analysing longitudinal data involving time-varying covariates

Reza Drikvandi[1,*], **Geert Verbeke**[2,3], **and Geert Molenberghs**[2,3]

[1] Department of Mathematical Sciences, Durham University, Durham, UK

[2] I-BioStat, KU Leuven, Leuven, Belgium

[3] I-BioStat, Universiteit Hasselt, Hasselt, Belgium

[*] Email: reza.drikvandi@durham.ac.uk

Abstract

Standard models for longitudinal data ignore the stochastic nature of time-varying covariates and their stochastic evolution over time by treating them as fixed variables. There have been recent methods for modelling time-varying covariates, however those methods cannot be applied to analyse longitudinal data when the longitudinal response and the time-varying covariates for each subject are measured at different time points. Moreover, it is difficult to study the temporal effects of a time-varying covariate on the longitudinal response and the temporal correlation between them. Motivated by data from an AIDS cohort study conducted over 26 years at the University Hospitals Leuven in which the measurements on the CD4 cell count and viral load for patients are not taken at the same time point, we present a framework to address those challenges by using joint multivariate mixed models to jointly model time-varying covariates and a longitudinal response, instead of including time-varying covariates in the response model. This approach also has the advantage that one can study the association between the covariate at any time point and the response at any other time point, without having to explicitly model the conditional distribution of the response given the covariate. We use penalised spline functions of time to capture the evolutions of both the response and time-varying covariates over time.

Keywords: *AIDS cohort study; Joint mixed model; Longitudinal data; Temporal association; Time-varying covariate.*

1. Introduction

Longitudinal studies are common in medicine, psychology, sociology, economics and other fields, where they allow researchers to study changes over time. They often produce data with both time-invariant (baseline) and time-varying covariates. A time-varying covariate, similar in the design to the response variable, is measured repeatedly over time. As a motivating case study described in Section 2, in AIDS cohort studies, two biomarkers, the CD4 cell count and the viral load, are measured for HIV+ patients at repeated visits before and after receiving treatment. It is difficult to simultaneously analyse them when the measurements on CD4 cell count and viral load are taken at different time points. Common methods in the literature are of an ad hoc nature, such as imputation methods for aligning all measurements temporally, or separate modelling of outcomes. Those methods have their own limitations as shown in our results.

Similar to classical regression models, standard models for longitudinal data ignore the stochastic nature of covariates and treat them as fixed variables. But, like the response variable, a time-varying covariate also changes over time and its stochastic evolution, which provides important information, should be modelled

as well. Ignoring the covariate process (i.e., the stochastic nature/measurement error) for time-varying covariates is perilous and could lead to incorrect inference, which in turn may lead to wrong decisions.

There is a large literature on modelling the covariate process for time-varying covariates. The literature can roughly be classified into two broad categories. The first category includes papers that develop models for time-varying covariates while still including the time-varying covariates in the response model. We briefly review some of the major papers in this category. To analyse multilevel binary longitudinal data, Miglioretti and Heagerty (2004) proposed to model time-varying covariates and incorporate them in the response model. Roy et al. (2006) suggested a model for time-varying covariates subject to missingness to analyse incomplete count data. Roy and Lin (2005) and Ghosh and Tu (2009) used transition models to take into account the covariate process to handle missing observations in time-varying covariates. Chen et al. (2014) modelled the covariate process to account for censored time-varying covariates. Despite the advantages of these methods, they cannot handle the situation where the longitudinal response and the time-varying covariates for each subject are measured at different time points. It is problematic to include a time-varying covariate into the response model while they are measured at different time points. This is because the unequal time points immediately rule out that one can formulate a plausible conditional model (e.g., outcome at time t conditional on covariate at $t-1$), and moreover it becomes more difficult to predict an outcome conditional on a series of longitudinally measured covariate values. Also, endogeneity could be an issue.

The second category includes papers that develop multivariate models to jointly model longitudinal outcomes. Sy et al. (1997) proposed a stochastic model for analysing bivariate longitudinal AIDS data with unequally spaced measurements, which is different than the situation where the longitudinal response and time-varying covariates are measured at different time points. Their joint model mainly includes a single random effect for each submodel and no penalised spline functions. Gueorguieva (2001) developed a multivariate generalised linear mixed model for joint modelling of clustered outcomes. Ferrer and McArdle (2003) suggested structural models for multivariate longitudinal data. Thiébaut et al. (2005) proposed a parametric joint bivariate linear mixed model for two outcomes with a lognormal survival model for drop-out time. Xiang et al. (2013) studied non-parametric models for multivariate longitudinal data. Lin and Wang (2013) developed a multivariate skew-normal linear model for multi-outcome longitudinal study. Kim and Albert (2016) presented a class of joint models for multivariate longitudinal measurements and a binary event. Li et al. (2017) proposed a multivariate joint mixed model for analysing mixed types of responses. Hui et al. (2018) introduced sparse pairwise likelihood estimation for multivariate longitudinal models. Proudfoot et al. (2018) used a joint marginal-conditional approach for modelling longitudinal data. Kürüm et al. (2018) suggested a copula model for joint modelling of longitudinal and time-varying outcomes. Zhao et al. (2021) developed a joint penalised spline model for multivariate longitudinal data.

Despite the large literature, there are still some challenging problems that may not be addressed using those methods. This paper, by advocating the second category of literature, aims to deal with two challenging problems. First, we aim to handle situations where the longitudinal response and the time-varying covariates for each subject are not measured at the same time point, as in our motivating application in Section 2. Second, unlike the current literature, we aim to study the temporal effects of a time-varying covariate on the response variable as well as the temporal correlation between them. This is also important in many applications. For example, in our AIDS case study, it is of interest to understand how the viral load affects the CD4 cell count over time and how the association between both depends on the time-lag between them. We want to avoid the assumption, sometimes encountered in the literature, that the re-

sponse variable measured at each time point depends on only the time-varying covariates measured at the same time point or at an earlier fixed time point. Indeed, the response variable could also depend on a series of previous measurements of time-varying covariates, and it may not be clear how long it takes for a change in the covariate to affect the response, nor how long the effect lasts. Another limitation of the existing methods relates to the fact that conditioning a variable on the other assumes a priori selection of the time points, that is, the outcome at some time point would have to be conditioned on the covariate at another time point, but which one?

To solve those problems, we follow the second category of literature and treat time-varying covariates as outcomes. For this, we use a joint mixed model that includes a submodel for the longitudinal response and a submodel for the time-varying covariate, but we do not include the time-varying covariate in the response submodel. This allows to jointly model the longitudinal response and the time-varying covariate in situations where they are measured at different time points. We link the two submodels through two sets of correlated random effects, where the association between the response and the time-varying covariate is taken into account via the correlation between random effects. We incorporate penalised spline functions of time to capture the evolutions of both the longitudinal response and the time-varying covariate over time. This approach also has the advantage that one can study the correlation between the covariate at any time point and the response at any other time point, without having to explicitly model the conditional distribution of the response conditional on the covariate. It allows, for example, to identify when this correlation is the strongest. More beneficially, we can study the temporal association between the longitudinal response and time-varying covariate. It also allows to predict the outcome at a future time conditional on a vector of measurements or history of covariate. We will illustrate these advantages using our motivating case study.

2. Motivating case study

Regular monitoring of disease, facilitated by technological advances, gives rise to multiple clinical measures whose inter-relation and stochastic evolutions over time could provide important insights into disease progression. A difficulty arises when some of those measures are taken at different time points. As a motivating application, we introduce a real data set obtained from an AIDS cohort study conducted over 26 years at the University Hospitals Leuven, Belgium, where the measurements on the CD4 cell count and the viral load are not taken at the same time point due to medical considerations and risks of successive tests on patients.

The study was performed at the AIDS Reference Center of Leuven in the University Hospitals Leuven, Belgium. In this study, 1257 HIV-infected patients were followed between 07 April 1987 and 16 October 2013. Although the patients were followed from the time HIV was detected, the treatment was started only later. The visits were prescheduled during the study, so there is no implicit assumption about the visit process and whether or not it is informative of the outcomes. The patients received antiretroviral treatments either non-nucleoside reverse transcriptase inhibitors (NNRTI) (about 35.8%) or protease inhibitors (PI) (about 64.2%) as their best first line treatment as approved by the Food and Drug Administration (FDA). The treatment was randomised and no patients switched from NNRTI to PI or vice versa. At each visit, the CD4 cell count as the primary outcome was measured for each patient during the study. Also, copies of viral load (viral RNA copy numbers) were measured repeatedly for each patient. In the study, the CD4 cell counts and the copies of viral load for each patient were not measured at the same time point due to medical considerations and risks of successive tests on patients. The number of repeated measurements was different

among the patients during the study. We present the frequency of the number of repeated measurements in Figure 1a, which shows that some of the patients had even more than 100 repeated measurements. The other variables quantified in the study were gender, the age of each patient at baseline, diagnosis year, best first line therapy used and follow-up time, along with some other demographic variables. We present the proportion of the treatments received by each of the female and male patient groups in Figure 1b. The distribution of the two treatments looks similar for the female and male patients.

There were no observations below the limit of detection (LOD) for CD4 cell count, but there were several for viral load. We replaced the observations below the LOD of viral load using the multiple imputation technique, instead of replacing them with half the LOD or another constant. For this, we used a truncated normal distribution, starting from the detection limit, and then applied multiple imputation to fill in the values below the LOD. The advantage of using multiple imputation over replacing the observations below the LOD with a constant such as half the LOD is that the latter would cause a point mass and hence a mixture distribution on viral load, which is not desirable from a modelling point of view.

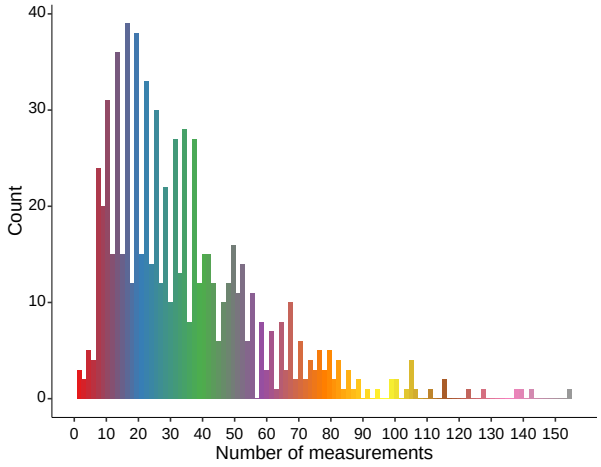
The individual profiles for all patients are presented in Figures 1c and 1d for the CD4 cell count and the viral load respectively, where time 0 (i.e., $t = 0$) is the start time of the therapy. It can be seen that there is a high variability between patients, which makes the analysis more complicated. Also, the plots reveal that both the CD4 cell count and the viral load have complex evolutions over time. For this study, it is of particular interest to answer the following questions:

1. What is the temporal effect of the viral load (a time-varying covariate) on the CD4 cell count (the response variable) at different stages of the disease and particularly before and after the treatment initiation?
2. What is the temporal correlation between the viral load and the CD4 cell count during the study period? How does it change before and after the treatment initiation?
3. How can the CD4 cell count evolutions after the treatment initiation be predicted given the entire viral load curve observed prior to the treatment and given that a particular treatment would be initiated?

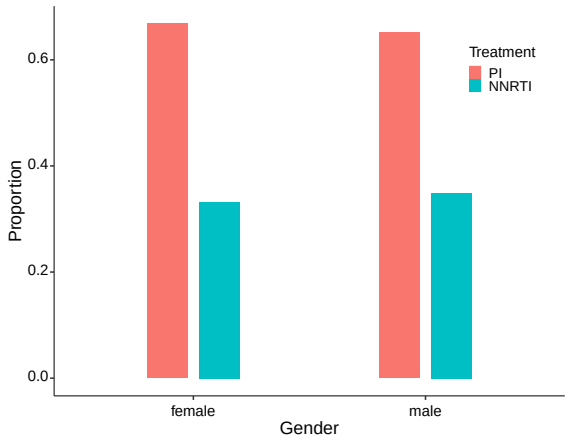
Temporal effects refer to *effects over time* and temporal correlation refers to *correlation over time*. While the two are related, temporal correlation shows the relation between outcomes over time and temporal effect shows the effect of one outcome on another over time. In our analysis of the AIDS data in Section 4.2, we provide answers to the above questions using the framework presented in the paper.

3. Methods

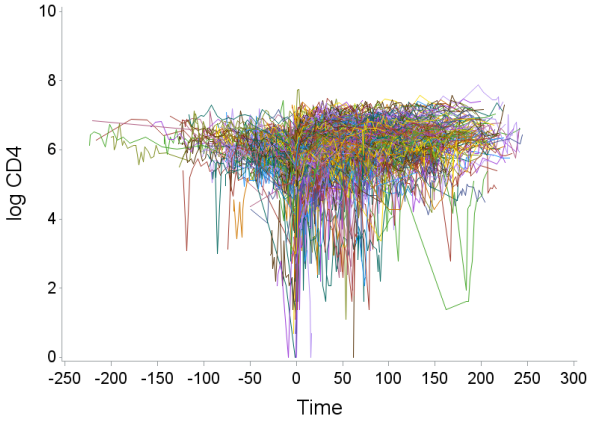
We emphasise that our objective is not to develop novel models but to provide a framework for handling time-varying covariates to address some limitations of the existing methods particularly to handle situations where the response and time-varying covariate are measured at different time points as well as to study their temporal association. Consider a longitudinal study in which N subjects are followed over time. Let Y be the outcome of interest, and further suppose that there are p covariates. For clarity of presentation, in this section we assume that there is one time-varying covariate in the study and denote it by V, while the other $p - 1$ covariates are all time-invariant. The case with more than one time-varying covariate will be discussed in Section 5. Note that in line with our data application, we here assume the assessment process is uninformative of outcomes post-baseline.



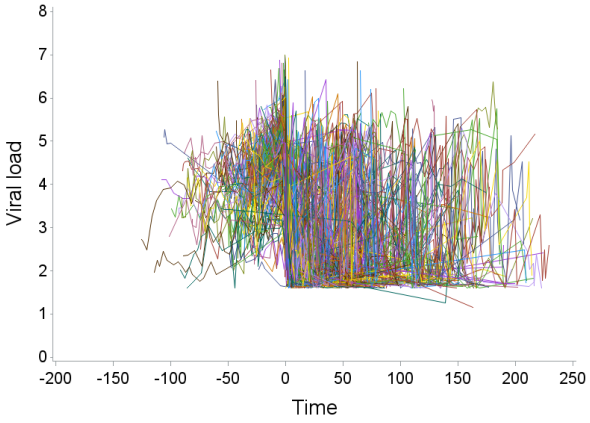
(a) Frequency of number of repeated measurements



(b) Proportion of treatments received by gender group



(c) CD4 cell count (cells/μL)



(d) Viral load (copies/mL)

Figure 1: (a) Frequency of the number of repeated measurements for patients, (b) Proportion of the treatments received by each of the gender groups, (c) Evolution of the log CD4 cell count for the HIV+ patients, (d) Evolution of the log viral load for the HIV+ patients. Note that time 0, say $t = 0$, is the start of the treatment.

Let $Y_i(t_{ij})$ and $V_i(s_{ij})$ denote, respectively, the response for subject i measured at time t_{ij} and the time-varying covariate for subject i measured at time s_{ij}. Since the two processes may be measured at different time points at the j-th follow-up (i.e., $t_{ij} \neq s_{ij}$), we do not include the time-varying covariate $V_i(s_{ij})$ in the model for longitudinal response $Y_i(t_{ij})$. Also in line with the literature, we incorporate non-parametric functions of time to have flexibility in capturing the stochastic evolutions of the response $Y_i(t_{ij})$ and the time-varying covariate $V_i(s_{ij})$. We thus propose to use a joint mixed model as follows:

$$\begin{cases} Y_i(t_{ij}) = \boldsymbol{x}_{1i}^T\boldsymbol{\beta}_1 + \boldsymbol{z}_{1i}^T\boldsymbol{b}_{1i} + m_1(t_{ij}) + \varepsilon_{1i}(t_{ij}) \\ V_i(s_{ij}) = \boldsymbol{x}_{2i}^T\boldsymbol{\beta}_2 + \boldsymbol{z}_{2i}^T\boldsymbol{b}_{2i} + m_2(s_{ij}) + \varepsilon_{2i}(s_{ij}), \end{cases} \tag{1}$$

where $\boldsymbol{x}_{1i}$ and $\boldsymbol{x}_{2i}$ are two $p-1$ and p^*-1 dimensional vectors of time-invariant covariates for subject i, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are two $p-1$ and p^*-1 dimensional vectors of fixed-effects parameters for the response and time-varying covariate submodels respectively, and $\boldsymbol{b}_{1i}$ and $\boldsymbol{b}_{2i}$ are two q and q^* dimensional vectors of subject-specific random effects capturing the between-subject variability for the response and time-varying covariate submodels respectively, with $\boldsymbol{z}_{1i}$ and $\boldsymbol{z}_{2i}$ being their corresponding q and q^* dimensional vectors of random-effects covariates. Also, m_1 and m_2 are two unknown smooth functions of time capturing the evolutions of the response and time-varying covariate over time, respectively. Finally, $\varepsilon_{1i}(t_{ij})$ and $\varepsilon_{2i}(s_{ij})$ are measurement errors associated with the response of subject i at time t_{ij} and the time-varying covariate of subject i at time s_{ij}, respectively.

In the joint mixed model (1), we let the two vectors of random effects $\boldsymbol{b}_{1i}$ and $\boldsymbol{b}_{2i}$ to be correlated. This would link the two submodels, allowing us to account for the association between the response and the time-varying covariate. We assume the combined vector of random effects $\boldsymbol{b}_i = (\boldsymbol{b}_{1i}^T, \boldsymbol{b}_{2i}^T)^T$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\text{Cov}(\boldsymbol{b}_i) = \begin{bmatrix} \boldsymbol{D}_{11} & \boldsymbol{D}_{12} \\ \boldsymbol{D}_{12}^T & \boldsymbol{D}_{22} \end{bmatrix}$.

We assume $(\varepsilon_{1i}(t_{ij}), \varepsilon_{2i}(s_{ij}))^T$ follows a bivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\text{Cov}(\varepsilon_{1i}(t_{ij}), \varepsilon_{2i}(s_{ij})) = \begin{bmatrix} \sigma_1^2(t_{ij}) & \sigma_{12}(t_{ij}, s_{ij}) \\ \sigma_{12}(t_{ij}, s_{ij}) & \sigma_2^2(s_{ij}) \end{bmatrix}$. One can also allow potential serial correlation in the residual covariance matrix by setting $\text{Cov}(\varepsilon_{1i}(t_{ij}), \varepsilon_{2i}(s_{ij})) = \begin{bmatrix} \sigma_1^2(t_{ij}) & \sigma_{12}(t_{ij}, s_{ij}) \\ \sigma_{12}(t_{ij}, s_{ij}) & \sigma_2^2(s_{ij}) \end{bmatrix} \otimes \rho^{|t_{ij} - s_{ij}|}$, where $\otimes$ is the Kronecker product and $|\rho| < 1$.

We use d-th degree and r-th degree penalised spline functions to approximate the unknown smooth functions $m_1(t_{ij})$ and $m_2(s_{ij})$, respectively, as follows

$$\begin{aligned} m_1(t_{ij}) &= \alpha_0 + \alpha_1 t_{ij} + ... + \alpha_d t_{ij}^d + \sum_{k=1}^{K} u_k (t_{ij} - \kappa_k)_+^d \\ m_2(s_{ij}) &= \gamma_0 + \gamma_1 s_{ij} + ... + \gamma_r s_{ij}^r + \sum_{l=1}^{L} u_l^* (s_{ij} - \lambda_l)_+^r, \end{aligned} \tag{2}$$

where $\{\kappa_1, ..., \kappa_K\}$ and $\{\lambda_1, ..., \lambda_L\}$ are two sets of fixed knots in the range of t_{ij} and s_{ij} respectively, the u_k and the u_l^* are spline coefficients for the response and time-varying covariate trajectories respectively, and $a_+ = \max(0, a)$. Penalised splines do not use all the time points as knots and usually require a moderate number of knots (e.g., 10 to 30) depending on the number of time points. A common approach is to specify equally-spaced knots, and the smoothness penalty helps to avert overfitting. To further reduce the risk of overfitting, it is common practice to utilise lower degree penalised spline functions, mainly linear or quadratic splines. This is the main advantage of low-degree penalised spline functions over high-degree polynomial functions, as the latter involves a lot of parameters which can increase the risk of overfitting.

Considering (2) and using the connection between penalised spline models and mixed-effects models (see, e.g., Brumback et al., 1999; Currie and Durban, 2002; Wand, 2003; Ruppert et al., 2003), we can equivalently write the spline-based joint mixed model (1) as the following joint parametric mixed model

$$\begin{cases} Y_i(t_{ij}) = \boldsymbol{x}_{1i}^T \boldsymbol{\beta}_1 + \boldsymbol{z}_{1i}^T \boldsymbol{b}_{1i} + \boldsymbol{T}_{ij}^T \boldsymbol{\alpha} + \boldsymbol{K}_{ij}^T \boldsymbol{u} + \varepsilon_{1i}(t_{ij}) \\ V_i(s_{ij}) = \boldsymbol{x}_{2i}^T \boldsymbol{\beta}_2 + \boldsymbol{z}_{2i}^T \boldsymbol{b}_{2i} + \boldsymbol{S}_{ij}^T \boldsymbol{\gamma} + \boldsymbol{\Lambda}_{ij}^T \boldsymbol{u}^* + \varepsilon_{2i}(s_{ij}), \end{cases} \tag{3}$$

where $\boldsymbol{T}_{ij} = (1, t_{ij}, ..., t_{ij}^d)^T$, $\boldsymbol{S}_{ij} = (1, s_{ij}, ..., s_{ij}^r)^T$, $\boldsymbol{\alpha} = (\alpha_0, ..., \alpha_d)^T$, $\boldsymbol{\gamma} = (\gamma_0, ..., \gamma_r)^T$, $\boldsymbol{K}_{ij} = ((t_{ij} - \kappa_1)_+^d, ..., (t_{ij} - \kappa_K)_+^d)^T$, $\boldsymbol{\Lambda}_{ij} = ((s_{ij} - \lambda_1)_+^r, ..., (s_{ij} - \lambda_L)_+^r)^T$, $\boldsymbol{u} = (u_1, ..., u_K)^T$, and $\boldsymbol{u}^* = (u_1^*, ..., u_L^*)^T$. The two vectors of spline coefficients $\boldsymbol{u}$ and $\boldsymbol{u}^*$ in (3), which are treated as random effects, are not correlated with the subject-specific random effects $\boldsymbol{b}_{1i}$ and $\boldsymbol{b}_{2i}$. We assume $\boldsymbol{u} \sim N\left(0, \sigma_u^2 \boldsymbol{I}_K\right)$ and $\boldsymbol{u}^* \sim N\left(0, \sigma_{u^*}^2 \boldsymbol{I}_L\right)$, where σ_u^2 and $\sigma_{u^*}^2$ are variance components of the random spline coefficients. The assumption of finite variance components σ_u^2 and $\sigma_{u^*}^2$ would shrink $\boldsymbol{u}$ and $\boldsymbol{u}^*$, leading to a smooth fit (e.g., Ruppert et al., 2003, p. 63). In fact, σ_u^2 and $\sigma_{u^*}^2$ are smoothing parameters which control the smoothness of the fit. Unlike the subject-specific random effects $\boldsymbol{b}_{1i}$ and $\boldsymbol{b}_{2i}$, the random spline coefficients $\boldsymbol{u}$ and $\boldsymbol{u}^*$ are not subject-specific, so they are constant across subjects.

After transforming the spline-based joint mixed model (1) into the equivalent joint parametric mixed model (3) using penalised splines, we use the restricted maximum likelihood (REML) approach for parameter estimation as we now have a fully parametric mixed model. It is known that the REML method produces less biased estimates for variance parameters compared to the usual maximum likelihood method (e.g., Verbeke and Molenberghs, 2009). For estimation purpose, the joint mixed model (3) can be represented as a unified linear mixed model $\boldsymbol{Y}_i^* = \boldsymbol{X}_i^* \boldsymbol{\beta}_i^* + \boldsymbol{Z}_i^* \boldsymbol{b}_i^* + \boldsymbol{W}_i^* \boldsymbol{U}^* + \boldsymbol{\varepsilon}_i^*$, where $\boldsymbol{Y}_i^* = \begin{bmatrix} \boldsymbol{Y}_i \\ \boldsymbol{V}_i \end{bmatrix}$, $\boldsymbol{\varepsilon}_i^* = \begin{bmatrix} \boldsymbol{\varepsilon}_{1i} \\ \boldsymbol{\varepsilon}_{2i} \end{bmatrix}$,

$$\boldsymbol{X}_i^* = \begin{bmatrix} \boldsymbol{X}_{1i} & \boldsymbol{0} & \boldsymbol{T}_i & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_{2i} & \boldsymbol{0} & \boldsymbol{S}_i \end{bmatrix}, \boldsymbol{Z}_i^* = \begin{bmatrix} \boldsymbol{Z}_{1i} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_{2i} \end{bmatrix}, \boldsymbol{W}_i^* = \begin{bmatrix} \boldsymbol{K}_i & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Lambda}_i \end{bmatrix}, \boldsymbol{\beta}^* = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \boldsymbol{\alpha} \\ \boldsymbol{\gamma} \end{bmatrix}, \boldsymbol{b}_i^* = \begin{bmatrix} \boldsymbol{b}_{1i} \\ \boldsymbol{b}_{2i} \end{bmatrix} \text{ and } \boldsymbol{U}^* = \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{u}^* \end{bmatrix},$$

in which

$$\boldsymbol{Y}_i = \begin{bmatrix} Y_i(t_{i1}) \\ \vdots \\ Y_i(t_{in_i}) \end{bmatrix}, \quad \boldsymbol{V}_i = \begin{bmatrix} V_i(s_{i1}) \\ \vdots \\ V_i(s_{in_i}) \end{bmatrix}, \quad \boldsymbol{\varepsilon}_{1i} = \begin{bmatrix} \varepsilon_{1i}(t_{i1}) \\ \vdots \\ \varepsilon_{1i}(t_{in_i}) \end{bmatrix}, \quad \boldsymbol{\varepsilon}_{2i} = \begin{bmatrix} \varepsilon_{2i}(s_{i1}) \\ \vdots \\ \varepsilon_{2i}(s_{in_i}) \end{bmatrix},$$

$$\boldsymbol{X}_{1i} = \begin{bmatrix} \boldsymbol{x}_{1i}^T \\ \vdots \\ \boldsymbol{x}_{1i}^T \end{bmatrix}, \quad \boldsymbol{X}_{2i} = \begin{bmatrix} \boldsymbol{x}_{2i}^T \\ \vdots \\ \boldsymbol{x}_{2i}^T \end{bmatrix}, \quad \boldsymbol{T}_i = \begin{bmatrix} \boldsymbol{T}_{i1}^T \\ \vdots \\ \boldsymbol{T}_{in_i}^T \end{bmatrix}, \quad \boldsymbol{S}_i = \begin{bmatrix} \boldsymbol{S}_{i1}^T \\ \vdots \\ \boldsymbol{S}_{in_i}^T \end{bmatrix},$$

$$\boldsymbol{Z}_{1i} = \begin{bmatrix} \boldsymbol{z}_{1i}^T \\ \vdots \\ \boldsymbol{z}_{1i}^T \end{bmatrix}, \quad \boldsymbol{Z}_{2i} = \begin{bmatrix} \boldsymbol{z}_{2i}^T \\ \vdots \\ \boldsymbol{z}_{2i}^T \end{bmatrix}, \quad \boldsymbol{K}_i = \begin{bmatrix} \boldsymbol{K}_{i1}^T \\ \vdots \\ \boldsymbol{K}_{in_i}^T \end{bmatrix}, \quad \boldsymbol{\Lambda}_i = \begin{bmatrix} \boldsymbol{\Lambda}_{i1}^T \\ \vdots \\ \boldsymbol{\Lambda}_{in_i}^T \end{bmatrix}.$$

Recall that, unlike the random effects $\boldsymbol{b}_i^*$, the random spline coefficients $\boldsymbol{U}^*$ are not subject-specific. It is straightforward to show that the marginal distribution of $\boldsymbol{Y}_i^* = \begin{bmatrix} \boldsymbol{Y}_i \\ \boldsymbol{V}_i \end{bmatrix}$, after integrating out the random effects $\boldsymbol{b}_i^*$ and $\boldsymbol{U}^*$, is

$$\begin{aligned} &\boldsymbol{Y}_i^* \sim N\left(\boldsymbol{X}_i^* \boldsymbol{\beta}^*, \boldsymbol{\Sigma}_i\right), \\ &\boldsymbol{\Sigma}_i = \boldsymbol{Z}_i^* \mathrm{Cov}(\boldsymbol{b}_i^*) \boldsymbol{Z}_i^{*T} + \boldsymbol{W}_i^* \mathrm{Cov}(\boldsymbol{U}^*) \boldsymbol{W}_i^{*T} + \mathrm{Cov}(\boldsymbol{\varepsilon}_i^*), \end{aligned} \tag{4}$$

in which

$$\mathrm{Cov}(\boldsymbol{b}_i^*) = \begin{bmatrix} \boldsymbol{D}_{11} & \boldsymbol{D}_{12} \\ \boldsymbol{D}_{12}^T & \boldsymbol{D}_{22} \end{bmatrix}, \ \mathrm{Cov}(\boldsymbol{U}^*) = \begin{bmatrix} \sigma_u^2 \boldsymbol{I}_K & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_{u^*}^2 \boldsymbol{I}_L \end{bmatrix}, \ \mathrm{Cov}(\boldsymbol{\varepsilon}_i^*) = \begin{bmatrix} \sigma_1^2(t_{ij}) \boldsymbol{I}_{n_i} & \sigma_{12}(t_{ij}, s_{ij}) \boldsymbol{I}_{n_i} \\ \sigma_{12}(t_{ij}, s_{ij}) \boldsymbol{I}_{n_i} & \sigma_2^2(s_{ij}) \boldsymbol{I}_{n_i} \end{bmatrix}.$$

Hence, the marginal log-likelihood function of the model would be as follows

$$l(\boldsymbol{\beta}^*, \boldsymbol{\phi}) = c - \frac{1}{2} \sum_{i=1}^{N} \log\left(|\boldsymbol{\Sigma}_i|\right) - \frac{1}{2} \sum_{i=1}^{N} (\boldsymbol{Y}_i^* - \boldsymbol{X}_i^* \boldsymbol{\beta}^*)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{Y}_i^* - \boldsymbol{X}_i^* \boldsymbol{\beta}^*), \tag{5}$$

where c is a constant and $\boldsymbol{\phi}$ represents all the variance parameters in the model. We employ an iterative procedure to obtain the restricted maximum likelihood estimates of the parameters, which is explained in the online Appendix A. We provide a SAS programme using **PROC HPMIXED** and **PROC MIXED** to calculate the parameter estimates $\hat{\boldsymbol{\beta}}^*$ and $\hat{\boldsymbol{\phi}}$. Our SAS code can be found in the online Appendix D.

3.1. The temporal correlation between the response variable and the time-varying covariate

The joint mixed model (3) enables us to study and understand the temporal effects of the time-varying covariate on the response variable as well as the temporal correlation between them. These two measures, which we study in this and the next subsections, can be useful in real applications, and we will use them to answer the first two questions related to our motivating application in Section 2.

From the joint mixed model (3), it is straightforward to find that

$$\mathrm{Cov}\big(Y_i(t_{ij}), V_i(s_{ij})\big) = \boldsymbol{z}_{1i}^T \boldsymbol{D}_{12} \boldsymbol{z}_{2i} + \sigma_{12}(t_{ij}, s_{ij}).$$

So, the temporal correlation between the response variable $Y_i(t_{ij})$ and the time-varying covariate $V_i(s_{ij})$ at the j-th follow-up can be obtained as

$$\mathrm{Corr}\big(Y_i(t_{ij}), V_i(s_{ij})\big) = \frac{\boldsymbol{z}_{1i}^T \boldsymbol{D}_{12} \boldsymbol{z}_{2i} + \sigma_{12}(t_{ij}, s_{ij})}{\sqrt{\boldsymbol{z}_{1i}^T \boldsymbol{D}_{11} \boldsymbol{z}_{1i} + \sigma_1^2(t_{ij})} \sqrt{\boldsymbol{z}_{2i}^T \boldsymbol{D}_{22} \boldsymbol{z}_{2i} + \sigma_2^2(s_{ij})}}. \tag{6}$$

Expression (6) can be used to answer practical questions such as when the association between Y and V is the strongest.

As a special case, when the joint mixed model (3) contains only random intercepts and random slopes (i.e., $\boldsymbol{z}_{1i} = (1, t_{ij})^T$ and $\boldsymbol{z}_{2i} = (1, s_{ij})^T$), we get

$$\mathrm{Corr}\big(Y_i(t_{ij}), V_i(s_{ij})\big) = \frac{d_{13} + d_{23}t_{ij} + d_{14}s_{ij} + d_{24}t_{ij}s_{ij} + \sigma_{12}(t_{ij}, s_{ij})}{\sqrt{d_{11} + 2d_{12}t_{ij} + d_{22}t_{ij}^2 + \sigma_1^2(t_{ij})}\sqrt{d_{33} + 2d_{34}s_{ij} + d_{44}s_{ij}^2 + \sigma_2^2(s_{ij})}},$$

where we defined $\boldsymbol{D}_{11} = \begin{bmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{bmatrix}$, $\boldsymbol{D}_{22} = \begin{bmatrix} d_{33} & d_{34} \\ d_{34} & d_{44} \end{bmatrix}$, $\boldsymbol{D}_{12} = \begin{bmatrix} d_{13} & d_{14} \\ d_{23} & d_{24} \end{bmatrix}$.
From the above, it can be seen that the correlation between the response variable $Y_i(t_{ij})$ and the time-varying covariate $V_i(s_{ij})$ is not constant and it depends on the time points t_{ij} and s_{ij}. We will illustrate this in our analysis of the AIDS data in Section 4.2.

3.2. The temporal effects of the time-varying covariate on the response variable

As discussed in the introduction, when the response variable Y and the time-varying covariate V are measured at different times, it would be problematic to regress Y on V by including the time-varying covariate V into the response Y model. The suggested framework avoids such an issue by jointly modelling the response $Y_i(t_{ij})$ and the time-varying covariate $V_i(s_{ij})$, which enables us to find the conditional distribution of $Y_i(t_{ij})$ given $V_i(s_{ij})$ at any points t_{ij} and s_{ij} in time, especially for situations when the response and the time-varying covariate are not measured at the same time point. So we can obtain the conditional expectation of $Y_i(t_{ij})$ given $V_i(s_{ij})$ to assess the temporal effect of the time-varying covariate V on the longitudinal response Y at the j-th follow-up. Since the joint distribution of $Y_i(t_{ij})$ and $V_i(s_{ij})$ is bivariate normal, it is straightforward to show that

$$\mathrm{E}\big(Y_i(t_{ij})|V_i(s_{ij})\big) = \big(\boldsymbol{x}_i^T\boldsymbol{\beta}_1 + \boldsymbol{T}_{ij}^T\boldsymbol{\alpha} + \boldsymbol{K}_{ij}^T\boldsymbol{u}\big) + \theta_{ij}\big(V_i(s_{ij}) - \boldsymbol{x}_i^T\boldsymbol{\beta}_2 - \boldsymbol{S}_{ij}^T\boldsymbol{\gamma} - \boldsymbol{\Lambda}_{ij}^T\boldsymbol{u}^*\big),$$

where

$$\theta_{ij} = \sqrt{\frac{\boldsymbol{z}_{1i}^T\boldsymbol{D}_{11}\boldsymbol{z}_{1i} + \sigma_1^2(t_{ij})}{\boldsymbol{z}_{2i}^T\boldsymbol{D}_{22}\boldsymbol{z}_{2i} + \sigma_2^2(s_{ij})}}\mathrm{Corr}\big(Y_i(t_{ij}), V_i(s_{ij})\big).$$

The above conditional expectation can be represented as

$$\mathrm{E}\big(Y_i(t_{ij})|V_i(s_{ij})\big) = a_{ij} + \theta_{ij}V_i(s_{ij}), \tag{7}$$

where $a_{ij} = \big(\boldsymbol{x}_i^T\boldsymbol{\beta}_1 + \boldsymbol{T}_{ij}^T\boldsymbol{\alpha} + \boldsymbol{K}_{ij}^T\boldsymbol{u}\big) - \theta_{ij}\big(\boldsymbol{x}_i^T\boldsymbol{\beta}_2 + \boldsymbol{S}_{ij}^T\boldsymbol{\gamma} + \boldsymbol{\Lambda}_{ij}^T\boldsymbol{u}^*\big)$. From (7), we can use the estimate of θ_{ij} as the estimated temporal effect of the time-varying covariate V on the longitudinal response Y at the j-th follow-up, which would be as follows

$$\widehat{\theta}_{ij} = \sqrt{\frac{\boldsymbol{z}_{1i}^T\hat{\boldsymbol{D}}_{11}\boldsymbol{z}_{1i} + \hat{\sigma}_1^2(t_{ij})}{\boldsymbol{z}_{2i}^T\hat{\boldsymbol{D}}_{22}\boldsymbol{z}_{2i} + \hat{\sigma}_2^2(s_{ij})}}\widehat{\mathrm{Corr}}\big(Y_i(t_{ij}), V_i(s_{ij})\big). \tag{8}$$

Clearly, when $\widehat{\mathrm{Corr}}\big(Y_i(t_{ij}), V_i(s_{ij})\big) \to 0$, the time-varying covariate V will have no effect on the longitudinal response Y. Reversely, the longitudinal response will be more affected by the time-varying covariate when the temporal correlation between them gets stronger over time. We will investigate this in our analysis of the AIDS data in Section 4.2.

It should be noted that the joint model is symmetrical in the sense that if one is interested in investigating $\mathrm{E}\big(V_i(s_{ij})|Y_i(t_{ij})\big)$, it can be similarly done too.

3.3. Prediction of the response evolutions given the history of the time-varying covariate

In many applications, it is of interest to predict the longitudinal response given the history of the time-varying covariate rather than just conditioning on a single observed value of the time-varying covariate. For

instance, in our case study (the third question), it is very helpful for clinicians to predict the CD4 evolutions after the treatment initiation (i.e., after $t = 0$) given the entire viral load curve observed prior to the treatment as well as the prior CD4 measurements, and given that a particular treatment would be initiated. When the treatment is randomised, such predictions would allow clinicians to pick the treatment with the most favourable outcome profile. Since the joint mixed model (3) can model the longitudinal response and the time-varying covariate observed at different times, it enables us to calculate such predictions.

To predict the longitudinal response Y_i for subject i at the $(j+1)$-th follow-up given the entire history of the time-varying covariate V_i prior to that follow-up, we use the conditional expectation of Y_i at time $t_{i,j+1}$ given the entire history of V_i up to the j-th follow-up, that is, $\mathrm{E}\big(Y_i(t_{i,j+1})|V_i(s_{i1}),\ldots,V_i(s_{ij})\big)$. Similarly, we predict the longitudinal response evolutions in the next $\mathcal{M}$ follow-ups given the entire history of V_i up to the j-th follow-up by

$$\mathrm{E}\big(Y_i(t_{i,j+1}),\ldots,Y_i(t_{i,j+\mathcal{M}})|V_i(s_{i1}),\ldots,V_i(s_{ij})\big). \tag{9}$$

To calculate (9), first define $\boldsymbol{Y}_i^{\mathcal{M}} = \big(Y_i(t_{i,j+1}),\ldots,Y_i(t_{i,j+\mathcal{M}})\big)^T$ and $\boldsymbol{V}_i = \big(V_i(s_{i1}),\ldots,V_i(s_{ij})\big)^T$. Then, since the joint distribution of $\boldsymbol{Y}_i$ and $\boldsymbol{V}_i$ is a multivariate normal distribution as shown in Section 3, we can calculate the conditional expectation (9) as follows

$$\begin{aligned}\mathrm{E}\big(Y_i(t_{i,j+1}),\ldots,Y_i(t_{i,j+\mathcal{M}})|V_i(s_{i1}),\ldots,V_i(s_{ij})\big) &= \mathrm{E}\big(\boldsymbol{Y}_i^{\mathcal{M}}|\boldsymbol{V}_i\big)\\ &= \mathrm{E}(\boldsymbol{Y}_i^{\mathcal{M}}) + \boldsymbol{\Sigma}_{\boldsymbol{Y}_i^{\mathcal{M}}\boldsymbol{V}_i}\boldsymbol{\Sigma}_{\boldsymbol{V}_i}^{-1}\big(\boldsymbol{V}_i - \mathrm{E}(\boldsymbol{V}_i)\big),\end{aligned}$$

in which $\mathrm{E}(\boldsymbol{Y}_i^{\mathcal{M}}) = \boldsymbol{X}_{1i}\boldsymbol{\beta}_1 + \boldsymbol{T}_i\boldsymbol{\alpha}$ and $\mathrm{E}(\boldsymbol{V}_i) = \boldsymbol{X}_{2i}\boldsymbol{\beta}_2 + \boldsymbol{S}_i\boldsymbol{\gamma}$. Also, $\boldsymbol{\Sigma}_{\boldsymbol{V}_i}$ is the covariance matrix of $\boldsymbol{V}_i$ for subject i and $\boldsymbol{\Sigma}_{\boldsymbol{Y}_i\boldsymbol{V}_i}$ is the covariance matrix of $\boldsymbol{Y}_i^{\mathcal{M}}$ and $\boldsymbol{V}_i$ for subject i. More generally, we can predict the response evolutions in the next $\mathcal{M}$ follow-ups by also conditioning on the prior response measurements $\boldsymbol{Y}_i = \big(Y_i(t_{i1}),\ldots,Y_i(t_{ij})\big)^T$. For this, denoting $\boldsymbol{Y}_i^* = \begin{bmatrix}\boldsymbol{Y}_i\\ \boldsymbol{V}_i\end{bmatrix}$ as before, we obtain

$$\begin{aligned}&\mathrm{E}\big(Y_i(t_{i,j+1}),\ldots,Y_i(t_{i,j+\mathcal{M}})|V_i(s_{i1}),\ldots,V_i(s_{ij}),Y_i(t_{i1}),\ldots,Y_i(t_{ij})\big)\\ &= \mathrm{E}\big(\boldsymbol{Y}_i^{\mathcal{M}}|\boldsymbol{V}_i,\boldsymbol{Y}_i\big)\\ &= \mathrm{E}(\boldsymbol{Y}_i^{\mathcal{M}}) + \boldsymbol{\Sigma}_{\boldsymbol{Y}_i^{\mathcal{M}}\boldsymbol{Y}_i^*}\boldsymbol{\Sigma}_{\boldsymbol{Y}_i^*}^{-1}\big(\boldsymbol{Y}_i^* - \mathrm{E}(\boldsymbol{Y}_i^*)\big),\end{aligned} \tag{10}$$

The covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{Y}_i^{\mathcal{M}}\boldsymbol{Y}_i^*}$ and $\boldsymbol{\Sigma}_{\boldsymbol{Y}_i^*}$ are submatrices of the full covariance matrix in (4), which can be directly calculated using **PROC MIXED** (the code is provided in the online Appendix D). Furthermore, $\mathrm{E}(\boldsymbol{Y}_i^{\mathcal{M}})$ and $\mathrm{E}(\boldsymbol{Y}_i^*)$ can be calculated using **PROC MIXED** for each subject (see the online Appendix D). In Section 4.2, we will use formula (10) to predict the CD4 evolutions after the treatment initiation given the entire viral load curve observed prior to the treatment period as well as the prior CD4 measurements.

3.4. Testing for a polynomial fit versus a penalised spline smoother

In the joint mixed model (3), it is useful to develop a test for choosing between a simpler polynomial fit and a general alternative described by penalised splines. This would help us to decide whether we can remove the truncated polynomial functions, resulting in a simpler joint mixed model without random spline coefficients. This is equivalent to testing whether or not all coefficients of the truncated power functions, which account for departures from a polynomial, are identically 0. This testing problem can be expressed as follows:

$$\begin{cases}H_0 : \sigma_u^2 = \sigma_{u^*}^2 = 0\\ H_A : \{\sigma_u^2 > 0\} \cup \{\sigma_{u^*}^2 > 0\}.\end{cases} \tag{11}$$

To perform the above hypothesis test, a main challenge is that the null hypothesis places the variance components on the boundary of parameter space. As a consequence, there is no open set containing the true variance components under the null hypothesis. Therefore, the classical asymptotic chi-squared distribution

of the likelihood ratio (LR) or restricted LR test statistic is not valid (see, for example, Stram and Lee, 1994; Drikvandi et al., 2012, 2013; Drikvandi and Noorian, 2019). For testing zero variance components, it is shown that the correct asymptotic distribution of the LR or restricted LR statistic is a mixture of chi-squared distributions, provided that the response variable can be partitioned into independent subvectors and the number of subvectors tends to infinity (e.g., Stram and Lee, 1994). However, this assumption does not hold under the alternative hypothesis above because the spline coefficients $\boldsymbol{u}$ and $\boldsymbol{u}^*$ are at the population level and not at the subject level. Therefore, the mixture of chi-squared distributions is not applicable either. One alternative approach would be to simulate the null distribution of the LR statistic using bootstrap or permutation; however, this may be unfeasible as it is not clear how bootstrap or permutation samples can be obtained under the null hypothesis in (11).

To avoid the above issues for testing H_0 versus H_A in (11), we use the Bayesian test proposed by Rao et al. (2019), which does not suffer the boundary issues. Their suggested Bayesian test can be used for testing multiple random effects. For this, we first implement the Monte Carlo simulation to calculate their proposed formula for the Bayes factor according to the procedure in Rao et al. (2019) and using their recommended prior distributions for the parameters. We then interpret the simulated Bayes factor using the popular scales in Jeffreys (1961) and Wasserman (2000) to conduct the test. If H_0 is not rejected, one can remove the truncated polynomial functions from m_1 and m_2 in (2), resulting in a simpler joint mixed model without random spline coefficients. Rao et al. (2019) showed that the Bayesian test has a high power for testing multiple random effects. We will apply this test to our case study in the next section.

4. Results

In this section, we first conduct simulations to examine the performance of the joint mixed model (1) on simulated data and compare it with three other methods. We then focus on analysing our motivating AIDS data application using the presented framework.

4.1. Simulations

In the simulations, we generate 100 random simulated data sets according to the joint mixed model (12) used for the AIDS data in next section, where we set the true parameter values based on the estimates obtained in Table 1 for the real data using the joint model. We also use the same number of repeated measurements for subjects as in the AIDS data. We fit the joint mixed model (12) to each simulated data set (using the codes in the online Appendix D) and calculate the estimation bias for all parameters. For comparison purposes, to the best of our knowledge, the recent methods in the literature cannot be applied when the longitudinal response and the time-varying covariate are measured at different time points. However, there are some common ad hoc methods that can be used for this purpose. A sensible approach is to impute the covariate values to temporally align the measurements with the response measurements, and then apply the standard longitudinal analysis. For this, we use the time-varying submodel to estimate those covariate values based on multiple imputation accounting for the uncertainty of estimation. In addition to this method, we also apply a linear mixed model fitted using the REML method to each outcome separately and calculate the bias of parameter estimates from each model accordingly. Additionally, we similarly apply separate Bayesian linear mixed models fitted using the MCMC estimation algorithm (available in R package `MCMCglmm`). Figures 2 and 3 show the bias of fixed-effects parameter estimates from all the

four methods for the CD4 cell count and viral load submodels, respectively. Also, Figures 4 and 5 show the bias of variance parameter estimates from all the four methods for the CD4 cell count and viral load submodels, respectively. The box plots indicate that the estimation bias from the joint mixed model is generally very small for all the parameters. The imputation-based method and the two separate modelling methods show relatively small biases for fixed-effects parameter estimates, while the estimation bias for variance parameters of these methods is substantially larger compared to the joint mixed model. We note that the separate modelling ignores the dependence and association between the response and time-varying covariate, and the imputation-based method tends to improve on this as it conditions on the viral load by incorporating the imputed covariate values. As one expects, the fixed-effects parameter estimates do not show very large biases, but the large bias of the variance parameter estimates would affect the inferences (e.g., hypothesis tests) regarding the fixed-effects parameters, as shown in our data analysis in the next subsection. Unlike the joint mixed model, the separate models do not allow us to investigate the association and temporal correlation between the response and time-varying covariate.

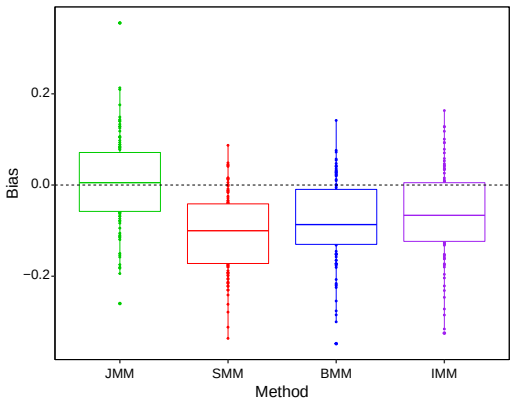
We also evaluate the prediction performance of the methods on 30 new simulated data sets, where the predictions of CD4 cell counts are calculated by conditioning on the previous measurements of the viral load and the CD4 cell count, as in formula (10). The mean squared prediction error of the joint mixed model is 0.09 (the unit is log scale) and it is 0.15 for the imputation-based method, while the two methods with separate mixed models and separate Bayesian mixed models produce relatively larger prediction errors of 0.22 and 0.21, respectively. Additional simulation results, including simulations when the normality assumptions in the joint model do not hold, are deferred to the online Appendix C due to space limitations.

4.2. Analysis of the AIDS data

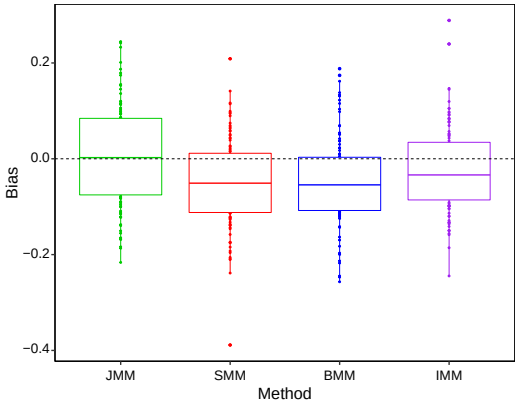
We analyse the AIDS data described in Section 2 using the suggested framework and provide answers to the questions raised in this case study. For this, we jointly model the log CD4 cell count and the log viral load using a spline-based joint mixed model in accordance with the joint mixed model (1). Note that for the viral load there is an immediate drop at time 0 because the start of the treatment immediately suppresses the virus in the body. This suggests that the regression function may not be smooth around 0 in the viral load submodel. For the CD4 cell count there is a V-shaped behaviour in the observed profile (see Figure 6a), which is expected from a clinical perspective, however there seems to be more smoothness around 0 for the CD4 cell count. The qualitative different nature of CD4 jump and viral load jump is biologically understandable, yet it is subtle from a modelling point of view. Therefore, to account for such behaviour, we consider two separate regression functions before and after 0 in each submodel. Our joint mixed model for the AIDS data is then as follows:

$$\begin{cases} CD4_i(t_{ij}) = m_{1,\text{before}}(t_{ij}) + m_{1,\text{after}}(t_{ij}) + \beta_1 Age_i + \beta_2 Gender_i + \beta_3 I_i^{trt} t_{ij} I(t_{ij} > 0) \\ \qquad\qquad + b_{1i,\text{before}} I(t_{ij} \leq 0) + b_{2i,\text{before}} t_{ij} I(t_{ij} \leq 0) + b_{3i,\text{after}} I(t_{ij} > 0) \\ \qquad\qquad + b_{4i,\text{after}} t_{ij} I(t_{ij} > 0) + \varepsilon_{1i}(t_{ij}), \\ VL_i(s_{ij}) = m_{2,\text{before}}(s_{ij}) + m_{2,\text{after}}(s_{ij}) + \beta_4 Age_i + \beta_5 Gender_i + \beta_6 I_i^{trt} s_{ij} I(s_{ij} > 0) \\ \qquad\qquad + b_{5i,\text{before}} I(s_{ij} \leq 0) + b_{6i,\text{before}} s_{ij} I(s_{ij} \leq 0) + b_{7i,\text{after}} I(s_{ij} > 0) \\ \qquad\qquad + b_{8i,\text{after}} s_{ij} I(s_{ij} > 0) + \varepsilon_{2i}(s_{ij}), \end{cases} \tag{12}$$

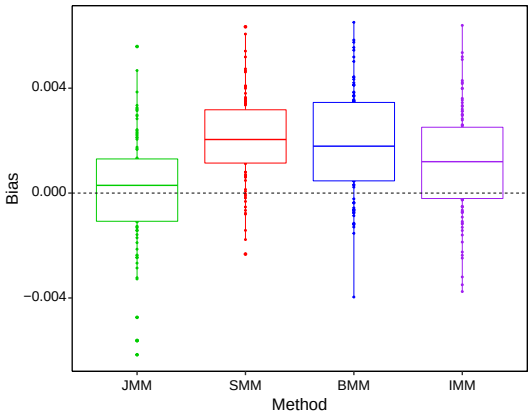
where $CD4_i(t_{ij})$ and $VL_i(s_{ij})$ are the log-transformed values of CD4 for patient i at time t_{ij} and viral load for patient i at time s_{ij} respectively, $m_{1,\text{before}}$ and $m_{1,\text{after}}$ are two separate smooth functions of time capturing the evolution of the CD4 before and after the treatment respectively, $m_{2,\text{before}}$ and $m_{2,\text{after}}$ are two separate smooth functions of time capturing the evolution of the viral load before and



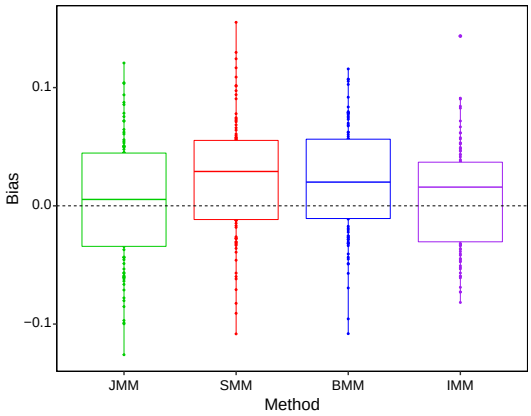
(a) For estimation of $\alpha_{0,\text{before}}$



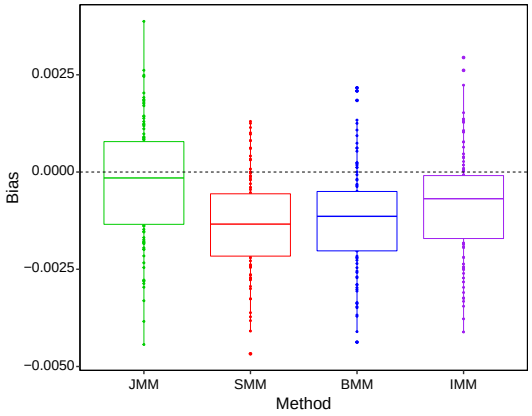
(b) For estimation of $\alpha_{0,\text{after}}$



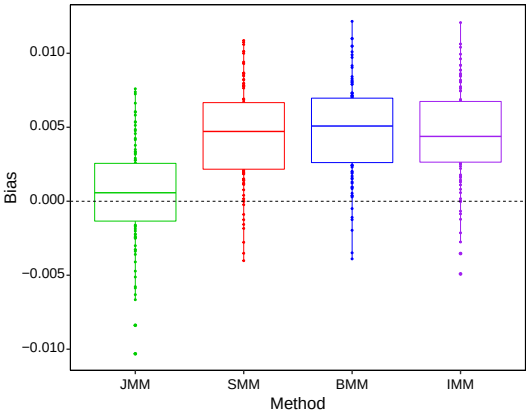
(c) For estimation of β_1



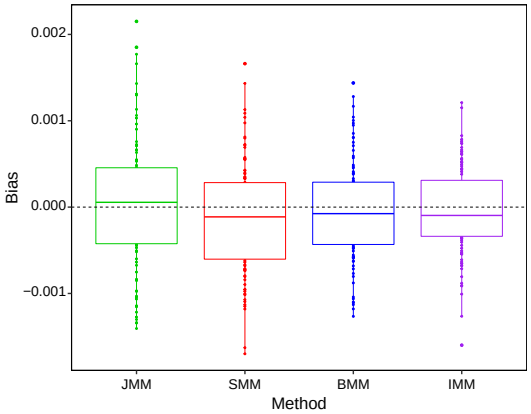
(d) For estimation of β_2



(e) For estimation of $\alpha_{1,\text{before}}$

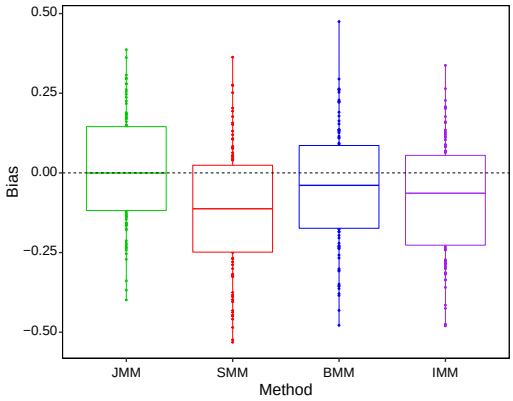


(f) For estimation of $\alpha_{1,\text{after}}$

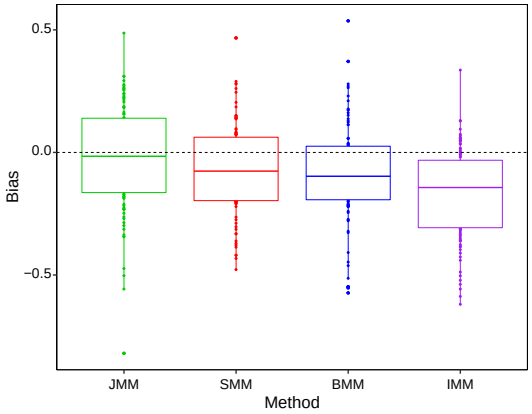


(g) For estimation of β_3

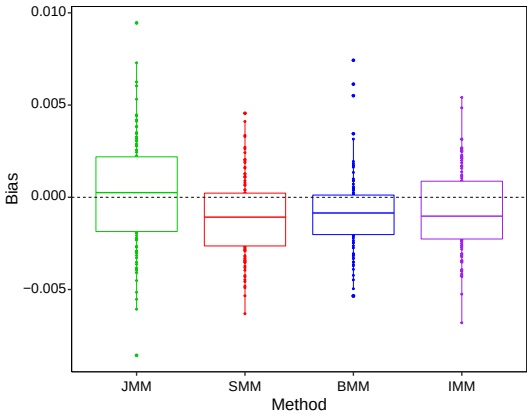
Figure 2: Estimation bias for fixed-effects parameters in the CD4 cell count submodel using the joint mixed model (denoted by JMM), as well as the method with two separate mixed models (denoted by SMM), the method with two separate Bayesian mixed models (denoted by BMM), and the imputation-based mixed model (denoted by IMM) over 100 simulation replications.



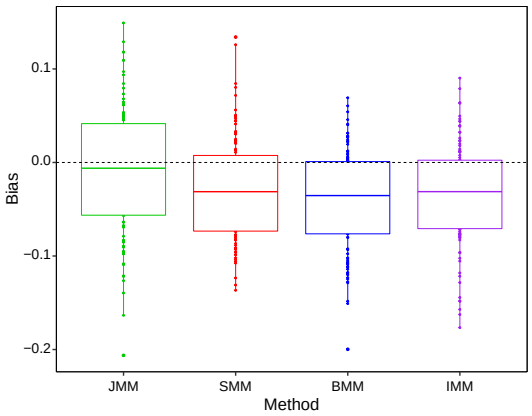
(a) For estimation of $\gamma_{0,\text{before}}$



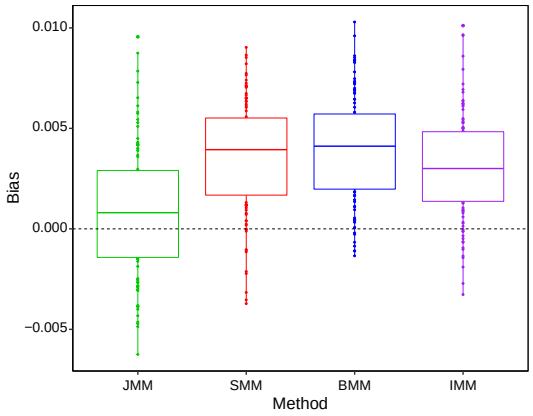
(b) For estimation of $\gamma_{0,\text{after}}$



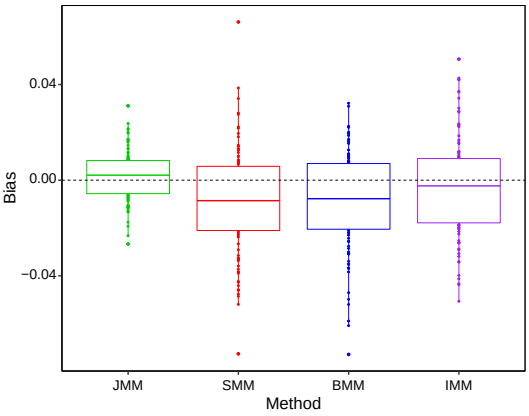
(c) For estimation of β_4



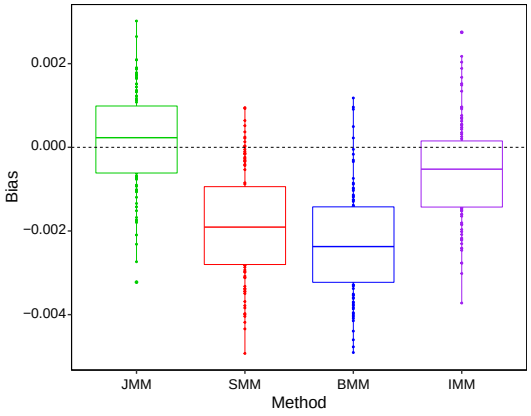
(d) For estimation of β_5



(e) For estimation of $\gamma_{1,\text{before}}$

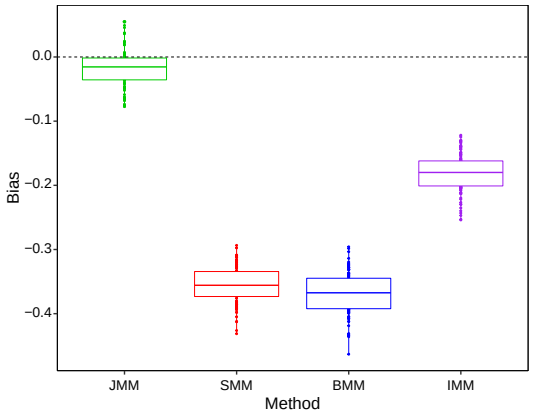


(f) For estimation of $\gamma_{1,\text{after}}$

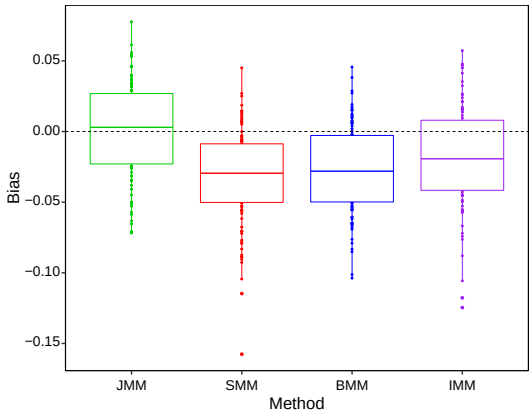


(g) For estimation of β_6

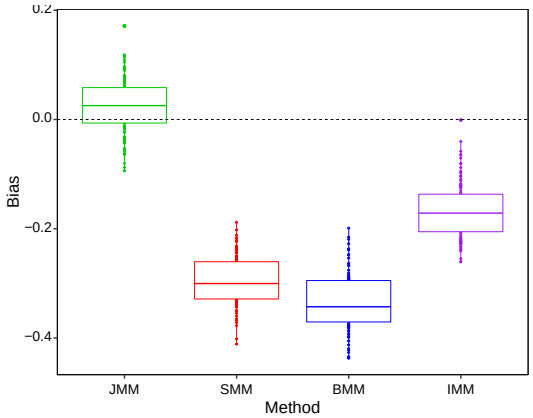
Figure 3: Estimation bias for fixed-effects parameters in the viral load submodel using the joint mixed model (denoted by JMM), as well as the method with two separate mixed models (denoted by SMM), the method with two separate Bayesian mixed models (denoted by BMM), and the imputation-based mixed model (denoted by IMM) over 100 simulation replications.



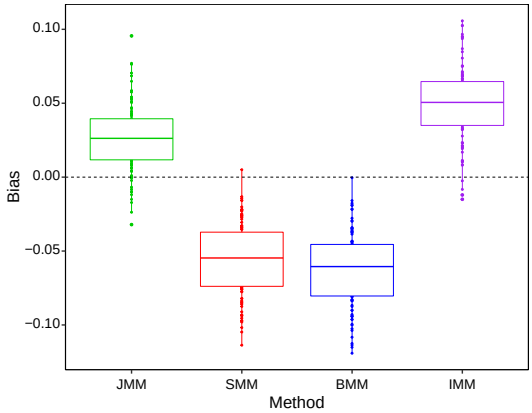
(a) For estimation of $\text{Var}(b_{1i,\text{before}})$



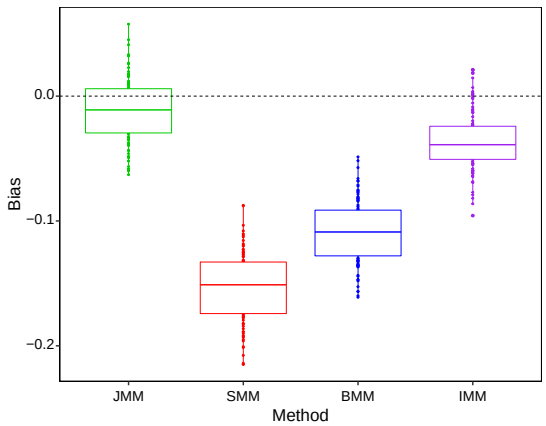
(b) For estimation of $\text{Var}(b_{2i,\text{before}})$



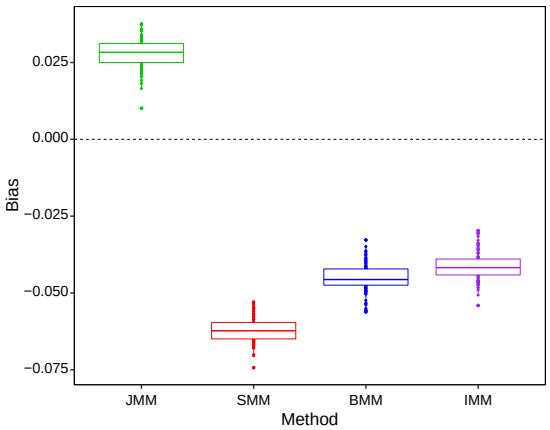
(c) For estimation of $\text{Var}(b_{3i,\text{after}})$



(d) For estimation of $\text{Var}(b_{4i,\text{after}})$

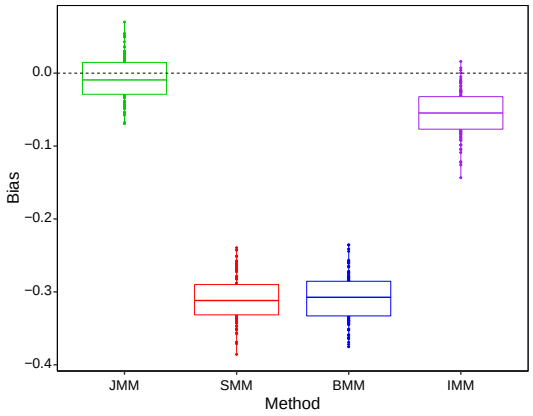


(e) For estimation of $\text{Var}(\varepsilon_{1i})$

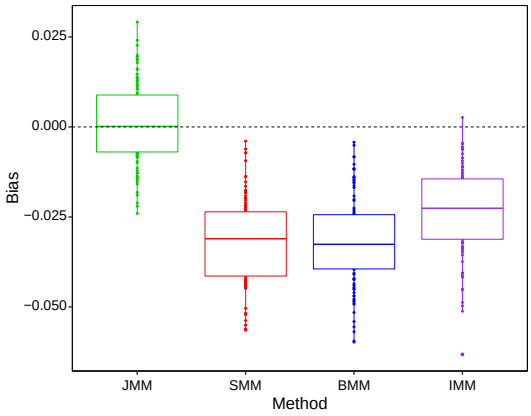


(f) For estimation of $\text{Var}(u_k)$

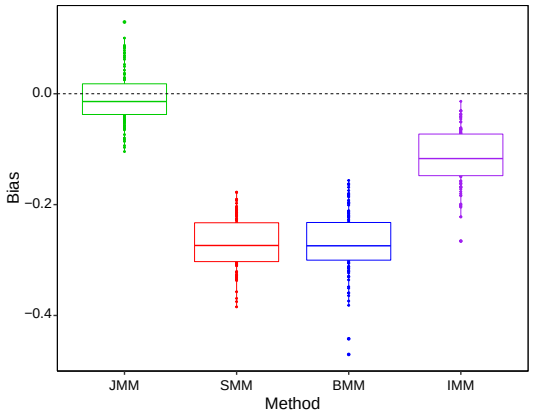
Figure 4: Estimation bias for variance parameters in the CD4 cell count submodel using the joint mixed model (denoted by JMM), as well as the method with two separate mixed models (denoted by SMM), the method with two separate Bayesian mixed models (denoted by BMM), and the imputation-based mixed model (denoted by IMM) over 100 simulation replications.



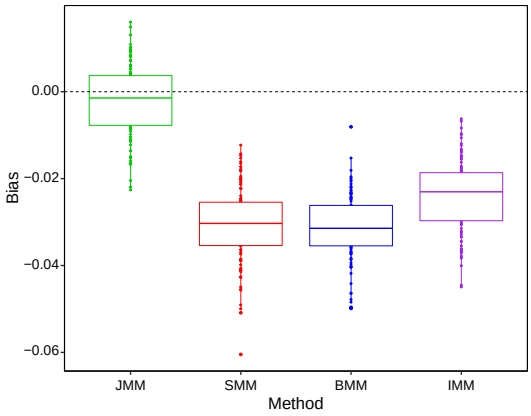
(a) For estimation of $\text{Var}(b_{5i,\text{before}})$



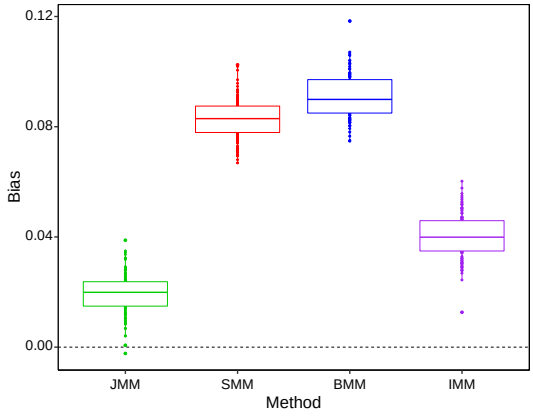
(b) For estimation of $\text{Var}(b_{6i,\text{before}})$



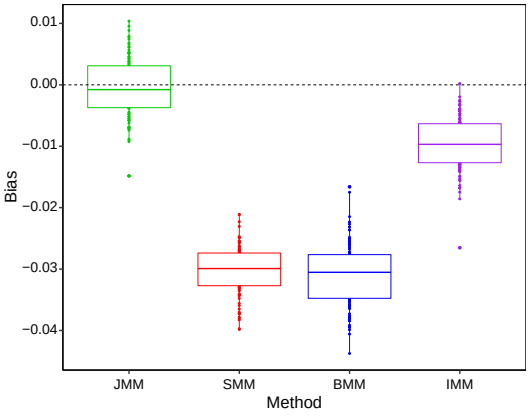
(c) For estimation of $\text{Var}(b_{7i,\text{after}})$



(d) For estimation of $\text{Var}(b_{8i,\text{after}})$



(e) For estimation of $\text{Var}(\varepsilon_{2i})$



(f) For estimation of $\text{Var}(u_l^*)$

Figure 5: Estimation bias for variance parameters in the viral load submodel using the joint mixed model (denoted by JMM), as well as the method with two separate mixed models (denoted by SMM), the method with two separate Bayesian mixed models (denoted by BMM), and the imputation-based mixed model (denoted by IMM) over 100 simulation replications.

after the treatment respectively, Age_i and $Gender_i$ are the age and gender for patient i, and I_i^{trt} is the treatment indicator for patient i which is 1 if the patient received NNTRI and 0 if the patient received PI. Also, $(\beta_1, \beta_2, \beta_3)^T$ and $(\beta_4, \beta_5, \beta_6)'$ are the fixed-effects parameters associated with the baseline covariates for the CD4 and viral load submodels, respectively. Moreover, $(b_{1i,\text{before}}, b_{2i,\text{before}}, b_{3i,\text{after}}, b_{4i,\text{after}})^T$ and $(b_{5i,\text{before}}, b_{6i,\text{before}}, b_{7i,\text{after}}, b_{8i,\text{after}})^T$ are two vectors of subject-specific random effects capturing the between-patient variability for the CD4 and viral load submodels respectively, where we allow different random intercepts and slopes before and after the treatment. Finally, $\varepsilon_{1i}(t_{ij})$ and $\varepsilon_{2i}(s_{ij})$ are, respectively, measurement errors for the CD4 measured at time t_{ij} and the viral load measured at time s_{ij}.

As discussed in Section 3, to minimise the risk of overfitting, we here use the first degree penalised spline functions for the temporal evolutions of the CD4 and the viral load before and after the treatment, as specified below:

$$\begin{cases} m_{1,\text{before}}(t_{ij}) = \alpha_{0,\text{before}} I(t_{ij} \leq 0) + \alpha_{1,\text{before}} t_{ij} I(t_{ij} \leq 0) + \sum_{k=1}^{10} u_k (t_{ij} - \kappa_k)_+ I(t_{ij} \leq 0) \\ m_{1,\text{after}}(t_{ij}) = \alpha_{0,\text{after}} I(t_{ij} > 0) + \alpha_{1,\text{after}} t_{ij} I(t_{ij} > 0) + \sum_{k=1}^{10} u_k (t_{ij} - \kappa_k)_+ I(t_{ij} > 0) \\ m_{2,\text{before}}(s_{ij}) = \gamma_{0,\text{before}} I(s_{ij} \leq 0) + \gamma_{1,\text{before}} s_{ij} I(s_{ij} \leq 0) + \sum_{l=1}^{10} u_l^* (s_{ij} - \lambda_l)_+ I(s_{ij} \leq 0) \\ m_{2,\text{after}}(s_{ij}) = \gamma_{0,\text{after}} I(s_{ij} > 0) + \gamma_{1,\text{after}} s_{ij} I(s_{ij} > 0) + \sum_{l=1}^{10} u_l^* (s_{ij} - \lambda_l)_+ I(s_{ij} > 0), \end{cases}$$

where we recall that the random spline coefficients u_k's and u_l^*'s are not subject-specific, so they are constant across patients.

It is important to note that parameter β_3 in the CD4 submodel is not interpretable as a treatment effect, and one should not conclude causality regarding an effect of treatment from this analysis; see Hernán et al. (2002) for a discussion on this. We emphasise that our aim here is to study how the trend changes once the treatment is initiated and whether or not such changes are different between groups. For this purpose, we add the indicator variable $I(t_{ij} > 0)$ in the CD4 submodel to adjust for time after treatment initiation. Similarly, the indicator variable $I(s_{ij} > 0)$ is added in the viral load submodel for parameter β_6.

As mentioned in Section 3, we use **PROC HPMIXED** and **PROC MIXED** in SAS to fit the joint mixed model (12) to the AIDS data (see the SAS code in the online Appendix D). We consider unstructured covariance matrices for the random effects in the joint mixed model (12). We assume the measurements errors of the CD4 cell count and viral load are uncorrelated (i.e., $\sigma_{12}(t_{ij}, s_{ij}) = 0$) because the repeated measurements of the CD4 cell count and viral load are taken at different time points. Note that this is not because of any computational limitation as the joint model can be similarly fitted allowing a non-zero $\sigma_{12}(t_{ij}, s_{ij})$. We model the error variance as an exponential function of time by setting LOCAL=EXP(time) in the SAS **REPEATED** statement. The estimate of the exponential local effect is 0.0035, which is significant based on a p-value of 0.0015. The estimates of the fixed-effects parameters and their associated standard errors are shown in Table 1, along with those of the imputation-based method and the two methods with separate mixed models and separate Bayesian mixed models. In line with our simulation results, the fixed-effects parameter estimates from the two separate modelling methods and the imputation-based method are not substantially different in comparison with the joint mixed model. The restricted maximum likelihood estimates of the covariances parameters from the joint mixed model are reported in Table 2, along with those of the other methods. As in our simulation results, the variance estimates from the separate modelling methods are different than the joint mixed model, and moreover only the joint mixed model provides covariances between the CD4 cell count and viral load. Despite the fixed-effects parameter estimates are not much different, the two separate models produce different inferences for three of the parameters, which

Table 1: The estimates of the fixed-effects parameters and associated standard errors obtained from fitting the joint mixed model (12), denoted by JMM, to the AIDS data, along with the estimates obtained from fitting the method with two separate mixed models, denoted by SMM, and the method with two separate Bayesian mixed models, denoted by BMM, as well as the imputation-based mixed model, denoted by IMM. Note that the significant coefficients at significance level 5% or 10% are marked by * or **.

Effect	Parameter	Estimate (s.e.) JMM	Estimate (s.e.) SMM	Estimate (pMCMC) BMM	Estimate (s.e.) IMM
	CD4 model				
intercept before treatment	$\alpha_{0,\text{before}}$	5.8231 (0.1035)*	5.7200 (0.0833)*	5.7213 (0.001)*	5.7406 (0.0879)*
intercept after treatment	$\alpha_{0,\text{after}}$	6.0888 (0.1034)*	6.0525 (0.0864)*	6.0535 (0.001)*	6.0528 (0.0867)*
age	β_1	−0.0025 (0.0021)	−0.0005 (0.0017)	−0.0005 (0.754)	−0.0005 (0.0016)
gender (female)	β_2	−0.1136 (0.0529)*	−0.0918 (0.0420)*	−0.0920 (0.028)*	−0.0919 (0.0400)*
time before treatment	$\alpha_{1,\text{before}}$	−0.0062 (0.0014)*	−0.0075 (0.0012)*	−0.0074 (0.001)*	−0.0074 (0.0011)*
time after treatment	$\alpha_{1,\text{after}}$	0.0547 (0.0031)*	0.0595 (0.0029)*	0.0595 (0.001)*	0.0592 (0.0030)*
time∗treatment (PI)	β_3	−0.0002 (0.0007)	−0.0003 (0.0004)	−0.0003 (0.610)	−0.0002 (0.0005)
	Viral load model				
intercept before treatment	$\gamma_{0,\text{before}}$	4.5002 (0.1802)*	4.3931 (0.1664)*	4.4379 (0.001)*	4.4246 (0.0978)*
intercept after treatment	$\gamma_{0,\text{after}}$	2.3058 (0.1861)*	2.2398 (0.1774)*	2.2821 (0.001)*	2.1857 (0.0961)*
age	β_4	−0.0011 (0.0025)	−0.0016 (0.0020)	−0.0021 (0.318)	−0.0024 (0.0022)
gender (female)	β_5	−0.1470 (0.0612)*	−0.1674 (0.0489)*	−0.1772 (0.001)*	−0.2027 (0.0554)*
time before treatment	$\gamma_{1,\text{before}}$	0.0033 (0.0034)	0.0063 (0.0027)*	0.0073 (0.001)*	0.0034 (0.0021)
time after treatment	$\gamma_{1,\text{after}}$	−0.0188 (0.0097)*	−0.0285 (0.0182)	−0.0262 (0.166)	−0.0284 (0.0009)*
time∗treatment (PI)	β_6	0.0036 (0.0019)**	0.0017 (0.0012)	0.0013 (0.258)	0.0059 (0.0014)*
−2 log-likelihood		30721.2	31416.03	·	30115.1

is due to the bias in standard errors of the fixed-effects parameter estimates caused by the large bias of the variance parameter estimates. Also, the imputation-based method does not model the association between the CD4 cell count and viral load, so it does not produce estimates for the covariance parameters between the two outcomes. Consequently, we cannot use this method to investigate the temporal association between the CD4 cell count and viral load.

Furthermore, from the results of the joint mixed model in Table 1, we can see that the time has a significant positive effect on the CD4 cell count after the treatment initiation, while the time effect prior to the treatment is negative. The time effects on the viral load before and after the treatment are reverse, so the time has a significant negative effect on the viral load after the treatment initiation. These results support the effectiveness of the treatments. Also, parameter β_6 in the viral load submodel is significant, which suggests a different trend over time between the treatments. The two methods with separate mixed models do not capture this.

Figure 6a shows the smoothed mean profile of the CD4 cell count and the fitted line obtained from fitting joint model (12) to the AIDS data. Similarly, Figure 6b shows the smoothed mean profile of the viral load and the fitted line obtained from fitting joint model (12) to the AIDS data. Considering the high variability between patients (recall Figure 1), the joint model performs very well in capturing the pattern and evolutions of both the CD4 cell count and the viral load over time. Furthermore, no major issue can be observed with the residuals of the CD4 cell count and viral load presented in Figures 6c and 6d respectively. See further discussion on this in the online Appendix B, where we also provide additional model checking results including checking the normality assumptions in the joint mixed model.

The theory in Sections 3.1 and 3.2 enables us to calculate the estimated temporal effects of the viral load on the CD4 cell count before and after the treatment initiation. The surface plots are shown in Figures 7a and 7b respectively. From these figures, it can be seen that the estimated temporal effects of the viral load on the CD4 cell count is generally negative, and the temporal effects get weaker after treatment compared to before treatment (see the colour bars). In particular, the temporal effects of the viral load on the CD4

Table 2: The estimates of the covariance parameters obtained from fitting the joint mixed model (12), denoted by JMM, to the AIDS data, along with the estimates obtained from fitting the method with two separate mixed models, denoted by SMM, and the method with two separate Bayesian mixed models, denoted by BMM, as well as the imputation-based mixed model, denoted by IMM.

Covariance parameter	Estimate JMM	Estimate SMM	Estimate BMM	Estimate IMM
$\mathrm{Var}(b_{1i,\mathrm{before}})$	0.84450	0.56227	0.56730	0.74985
$\mathrm{Var}(b_{2i,\mathrm{before}})$	0.00022	0.00013	0.00014	0.00014
$\mathrm{Var}(b_{3i,\mathrm{after}})$	0.82370	0.38334	0.38560	0.61917
$\mathrm{Var}(b_{4i,\mathrm{after}})$	0.00012	0.00004	0.00004	0.00005
$\mathrm{Var}(b_{5i,\mathrm{before}})$	0.61600	0.35336	0.35310	0.64571
$\mathrm{Var}(b_{6i,\mathrm{before}})$	0.00075	0.00003	0.00001	0.00064
$\mathrm{Var}(b_{7i,\mathrm{after}})$	0.77060	0.47170	0.47320	0.68430
$\mathrm{Var}(b_{8i,\mathrm{after}})$	0.00025	0.00005	0.00006	0.00023
$\mathrm{Cov}(b_{1i,\mathrm{before}}, b_{2i,\mathrm{before}})$	0.00507	0.00672	0.00671	0.08080
$\mathrm{Cov}(b_{1i,\mathrm{before}}, b_{3i,\mathrm{after}})$	0.26610	0.37509	0.37750	0.07680
$\mathrm{Cov}(b_{1i,\mathrm{before}}, b_{4i,\mathrm{after}})$	−0.00267	−0.00293	−0.00295	−0.06000
$\mathrm{Cov}(b_{1i,\mathrm{before}}, b_{5i,\mathrm{before}})$	−0.06252	·	·	·
$\mathrm{Cov}(b_{1i,\mathrm{before}}, b_{6i,\mathrm{before}})$	0.01669	·	·	·
$\mathrm{Cov}(b_{1i,\mathrm{before}}, b_{7i,\mathrm{after}})$	0.00298	·	·	·
$\mathrm{Cov}(b_{1i,\mathrm{before}}, b_{8i,\mathrm{after}})$	0.00251	·	·	·
$\mathrm{Cov}(b_{2i,\mathrm{before}}, b_{3i,\mathrm{after}})$	0.00514	0.00415	0.00413	0.05740
$\mathrm{Cov}(b_{2i,\mathrm{before}}, b_{4i,\mathrm{after}})$	−0.00009	−0.00005	−0.00004	−0.05800
$\mathrm{Cov}(b_{2i,\mathrm{before}}, b_{5i,\mathrm{before}})$	−0.00040	·	·	·
$\mathrm{Cov}(b_{2i,\mathrm{before}}, b_{6i,\mathrm{before}})$	−0.00009	·	·	·
$\mathrm{Cov}(b_{2i,\mathrm{before}}, b_{7i,\mathrm{after}})$	−0.00033	·	·	·
$\mathrm{Cov}(b_{2i,\mathrm{before}}, b_{8i,\mathrm{after}})$	0.00006	·	·	·
$\mathrm{Cov}(b_{3i,\mathrm{after}}, b_{4i,\mathrm{after}})$	−0.00563	−0.00234	−0.00235	−0.05910
$\mathrm{Cov}(b_{3i,\mathrm{after}}, b_{5i,\mathrm{before}})$	−0.01594	·	·	·
$\mathrm{Cov}(b_{3i,\mathrm{after}}, b_{6i,\mathrm{before}})$	−0.00339	·	·	·
$\mathrm{Cov}(b_{3i,\mathrm{after}}, b_{7i,\mathrm{after}})$	−0.04506	·	·	·
$\mathrm{Cov}(b_{3i,\mathrm{after}}, b_{8i,\mathrm{after}})$	−0.00076	·	·	·
$\mathrm{Cov}(b_{4i,\mathrm{after}}, b_{5i,\mathrm{before}})$	0.00035	·	·	·
$\mathrm{Cov}(b_{4i,\mathrm{after}}, b_{6i,\mathrm{before}})$	0.00001	·	·	·
$\mathrm{Cov}(b_{4i,\mathrm{after}}, b_{7i,\mathrm{after}})$	0.00003	·	·	·
$\mathrm{Cov}(b_{4i,\mathrm{after}}, b_{8i,\mathrm{after}})$	−0.00005	·	·	·
$\mathrm{Cov}(b_{5i,\mathrm{before}}, b_{6i,\mathrm{before}})$	0.00059	0.00085	0.00052	−0.04600
$\mathrm{Cov}(b_{5i,\mathrm{before}}, b_{7i,\mathrm{after}})$	0.01410	−0.00584	−0.01815	−0.06700
$\mathrm{Cov}(b_{5i,\mathrm{before}}, b_{8i,\mathrm{after}})$	−0.00409	−0.00288	0.00369	0.10900
$\mathrm{Cov}(b_{6i,\mathrm{before}}, b_{7i,\mathrm{after}})$	0.00788	0.00151	0.00153	0.05500
$\mathrm{Cov}(b_{6i,\mathrm{before}}, b_{8i,\mathrm{after}})$	0.00001	−0.00001	−0.00001	0.02800
$\mathrm{Cov}(b_{7i,\mathrm{after}}, b_{8i,\mathrm{after}})$	−0.00405	−0.00001	−0.00178	−0.11400
$\mathrm{Var}(u_k)$	0.00013	0.00027	0.00046	0.00028
$\mathrm{Var}(u_l^*)$	0.00006	0.00033	0.00050	0.00013
$\mathrm{Var}(\varepsilon_{1i})$	0.09115	0.09320	0.09325	0.09321
$\mathrm{Var}(\varepsilon_{2i})$	0.50490	0.56943	0.57620	0.49424

cell count is weak (i.e., closer to 0) at the beginning of study (i.e., around $t_{ij} = -250$ and $s_{ij} = -250$), but it then tends to get stronger (more negative) before the treatment is started. It gets weak again once the treatment is initiated (i.e., around $t_{ij} = 0$ and $s_{ij} = 0$) which is because the viral load drops with the start of treatment (recall the mean profile of viral load in Figure 6b). Figures 7c and 7d show the temporal correlation between the CD4 cell count and the viral load before and after the treatment, respectively. From these figures, the temporal correlation between the CD4 cell count and the viral load is negative especially before the treatment period. The correlation is strongest (most negative) prior to the start of treatment, and it gets weaker (i.e., very close to 0) once the treatment is started.

We then used the prediction formula (10) in Section 3.3 to calculate the predictions of the CD4 evolutions after the treatment initiation (i.e., after $t = 0$) given the entire viral load curve observed prior to $t = 0$ and the prior CD4 measurements. As the treatment was randomised, such predictions would allow clinicians to pick the treatment with the most favourable outcome profile as discussed in Section 3.3. We calculated the predictions for six randomly selected patients. The prediction plots, presented in Figure 8, show that the joint mixed model produces good predictions of CD4 cell counts, considering the fact that predictions at the individual level are generally more uncertain compared to predictions at the population level.

We also applied the Bayesian test of Rao et al. (2019) explained in Section 3.4 to test whether a simpler polynomial fit can be used instead of the model with penalised spline functions. For this, we obtained a simulated Bayes factor of 20.83, which is relatively large according to the scales in Jeffreys (1961) and Wasserman (2000). This suggests that the spline functions are significant and therefore needed in the joint mixed model (12) to adequately capture the evolutions of the CD4 cell count and the viral load over time.

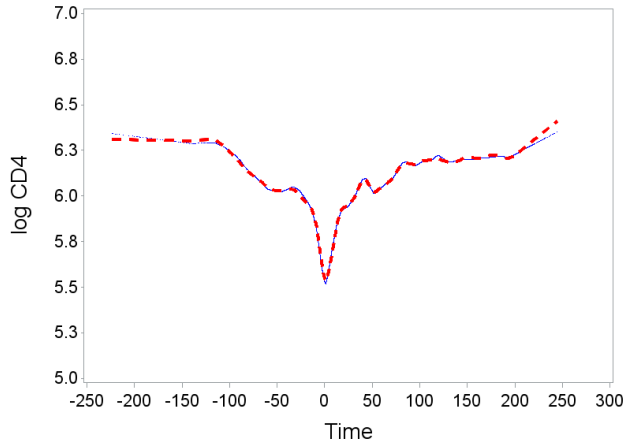
5. Some extensions

5.1. Extension to more than one time-varying covariate

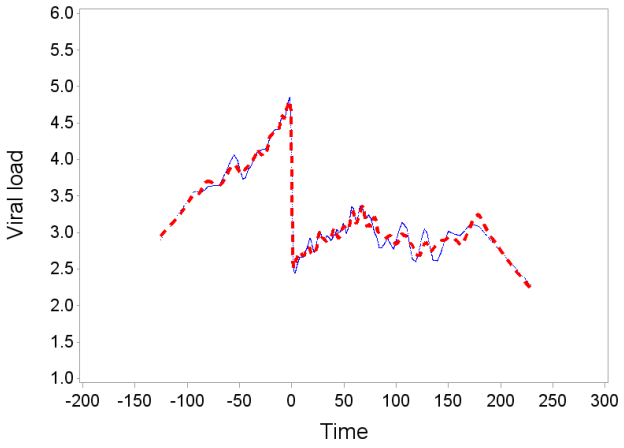
For simplicity of presentation and in line with the AIDS application, the proposed framework is first demonstrated for the case of one time-varying covariate. The joint mixed model can similarly be extended to situations with multiple time-varying covariates. This can be done by including a submodel for each time-varying covariate. We explain this for the case when there are two time-varying covariates in the study, denoted by V_1 and V_2. We write the corresponding joint mixed model as follows

$$\begin{cases} Y_i(t_{ij}) = \boldsymbol{x}_{1i}^T\boldsymbol{\beta}_1 + \boldsymbol{z}_{1i}^T\boldsymbol{b}_{1i} + m_1(t_{ij}) + \varepsilon_{1i}(t_{ij}) \\ V_{1i}(s_{ij}) = \boldsymbol{x}_{2i}^T\boldsymbol{\beta}_2 + \boldsymbol{z}_{2i}^T\boldsymbol{b}_{2i} + m_2(s_{ij}) + \varepsilon_{2i}(s_{ij}) \\ V_{2i}(r_{ij}) = \boldsymbol{x}_{3i}^T\boldsymbol{\beta}_3 + \boldsymbol{z}_{3i}^T\boldsymbol{b}_{3i} + m_3(r_{ij}) + \varepsilon_{3i}(r_{ij}), \end{cases}$$

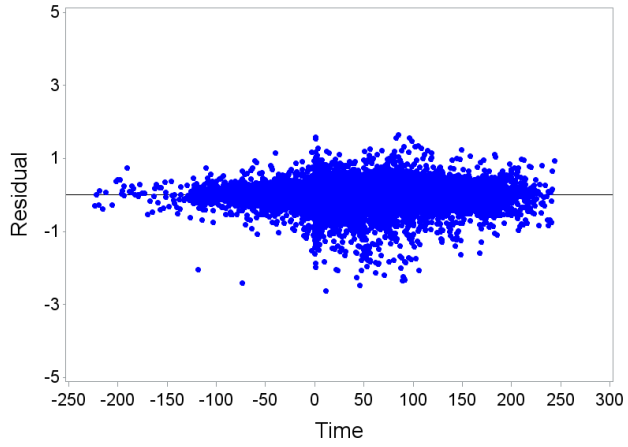
where the second time-varying covariate V_2 is also allowed to be measured at different time points r_{ij}. The calculation for the above model is similarly done using the estimation method we used in Section 3. We note that if there are several time-varying covariates in a data application requiring several possible submodels, we suggest to implement the pairwise model fitting approach of Fieuws and Verbeke (2006) which can calculate the joint likelihood by evaluating each pair of submodels to find the parameter estimates of the joint mixed model. Fieuws and Verbeke (2006) showed the effectiveness of this approach when jointly modelling many variables.



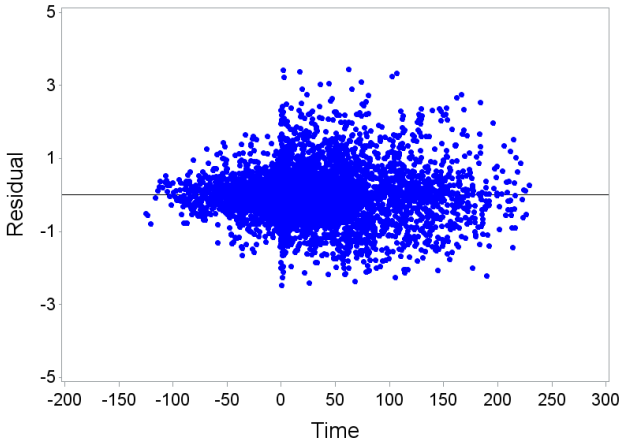
(a) Mean profile of the log CD4 cell count in blue (solid line) and the smoothed fitted line in red (dashed line).



(b) Mean profile of the log viral load in blue (solid line) and the smoothed fitted line in red (dashed line).

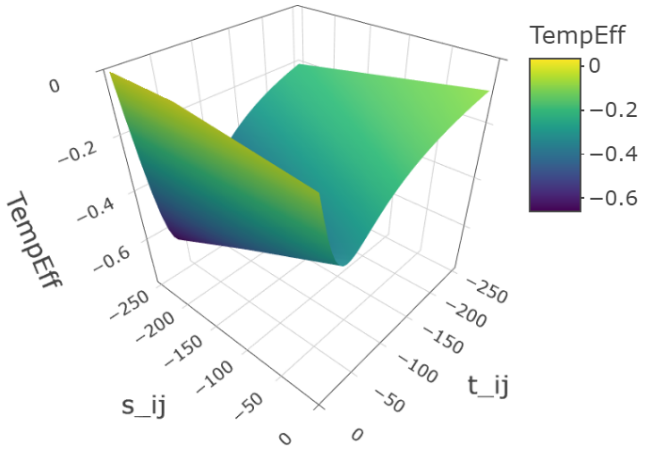


(c) Residuals of the CD4 cell count.

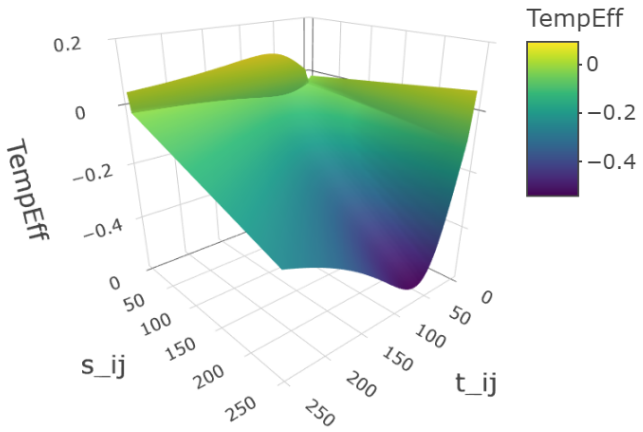


(d) Residuals of the viral load.

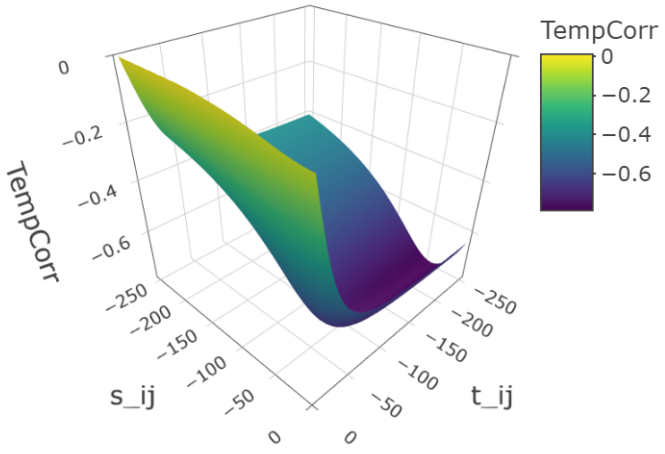
Figure 6: Plots obtained from fitting the joint mixed model (12) to the AIDS data.



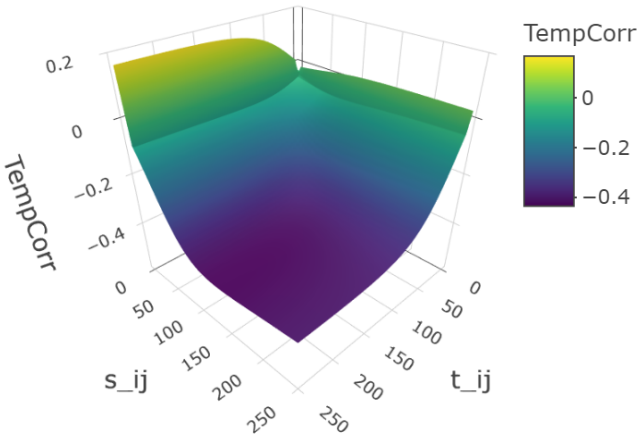
(a) The temporal effects of the viral load on the CD4 cell count before treatment.



(b) The temporal effects of the viral load on the CD4 cell count after treatment.

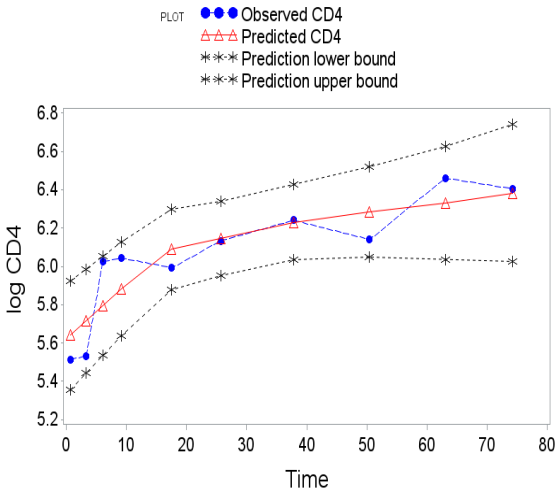


(c) The temporal correlation between the viral load and the CD4 cell count before treatment.

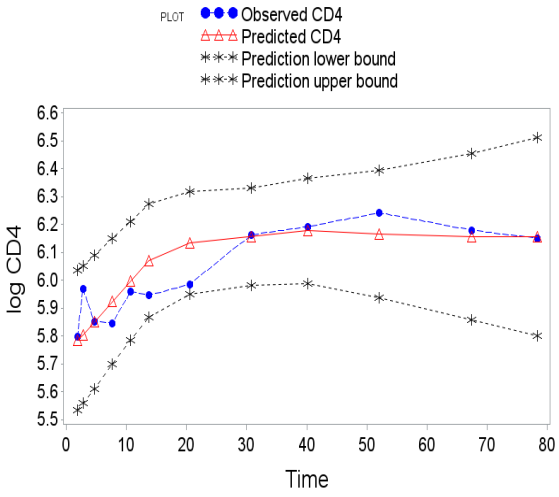


(d) The temporal correlation between the viral load and the CD4 cell count after treatment.

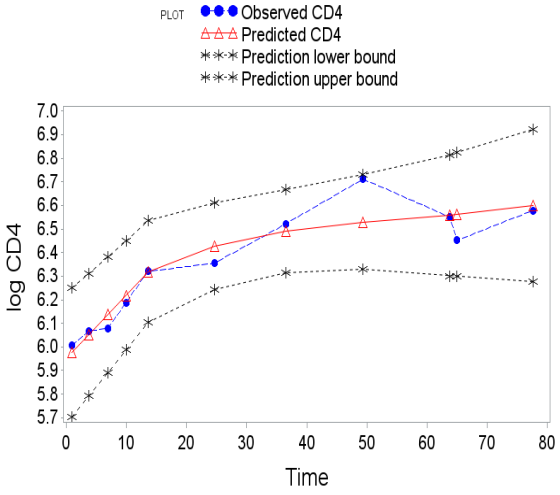
Figure 7: Surface plots of the temporal effects of the viral load on the CD4 cell count as well as the temporal correlation between them before and after treatment, obtained from fitting the joint mixed model (12) to the AIDS data. As the colour bar shows, the darker colour the higher impact the viral load would have on the CD4 cell count at that time.



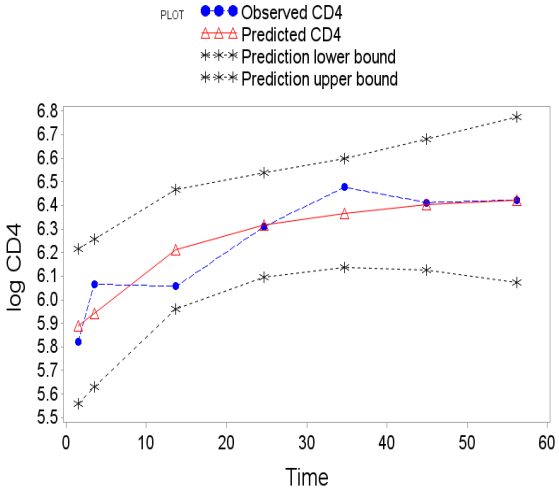
(a) Patient id 98.



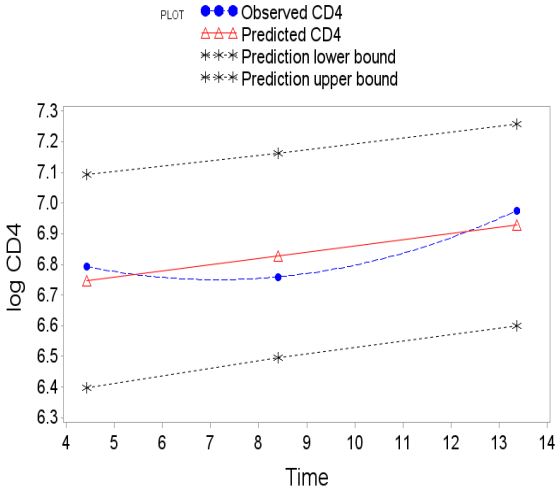
(b) Patient id 437.



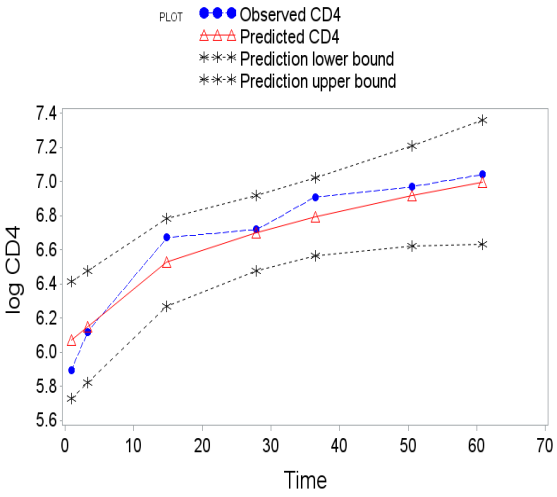
(c) Patient id 492.



(d) Patient id 503.



(e) Patient id 761.



(f) Patient id 884.

Figure 8: The predictions of the CD4 evolutions, along with lower and upper bounds, after the treatment initiation (i.e., after $t = 0$) given the entire viral load curve observed prior to the treatment and the prior CD4 measurements for six randomly selected patients.

5.2. Extension to other types of time-varying covariate

Another extension relates to modelling other types of time-varying covariates which may be binary or categorical etc., such as treatment indicators or whether currently infected with an infectious virus such as COVID-19. Modelling such time-varying covariates will require an extension to the joint mixed model (1) by incorporating a generalised linear mixed model for such time-varying covariate. We consider the setting where there is a binary time-varying covariate, say V_i, taking the value 1 or 0. One may think of a logistic submodel for V_i; however, it can be easier to use a latent variable formulation. For this, suppose that there is a latent continuous variable V_i^* so that $V_i = 1$ if $V_i^* > 0$ and $V_i = 0$ if $V_i^* \leq 0$. Using $V_i^* \sim N\big(\boldsymbol{x}_{2i}^T\boldsymbol{\beta}_2 + \boldsymbol{S}_{ij}^T\boldsymbol{\gamma}, \boldsymbol{z}_{2i}^T\boldsymbol{D}_{22}\boldsymbol{z}_{2i} + \sigma_{u^*}^2\boldsymbol{\Lambda}_{ij}^T\boldsymbol{\Lambda}_{ij} + \sigma_2^2(s_{ij})\big)$, the probabilities for outcomes of the binary covariate are calculated as

$$P(V_i = 1) = P(V_i^* > 0) = 1 - \Phi\big(-\big(\boldsymbol{x}_{2i}^T\boldsymbol{\beta}_2 + \boldsymbol{S}_{ij}^T\boldsymbol{\gamma}\big)/\sqrt{\boldsymbol{z}_{2i}^T\boldsymbol{D}_{22}\boldsymbol{z}_{2i} + \sigma_{u^*}^2\boldsymbol{\Lambda}_{ij}^T\boldsymbol{\Lambda}_{ij} + \sigma_2^2(s_{ij})}\big),$$
$$P(V_i = 0) = P(V_i^* \leq 0) = \Phi\big(-\big(\boldsymbol{x}_{2i}^T\boldsymbol{\beta}_2 + \boldsymbol{S}_{ij}^T\boldsymbol{\gamma}\big)/\sqrt{\boldsymbol{z}_{2i}^T\boldsymbol{D}_{22}\boldsymbol{z}_{2i} + \sigma_{u^*}^2\boldsymbol{\Lambda}_{ij}^T\boldsymbol{\Lambda}_{ij} + \sigma_2^2(s_{ij})}\big),$$

where $\Phi(\cdot)$ denotes the CDF of the standard normal distribution. The corresponding joint mixed model based on the latent continuous variable V_i^* is then as follows

$$\begin{cases} Y_i(t_{ij}) = \boldsymbol{x}_{1i}^T\boldsymbol{\beta}_1 + \boldsymbol{z}_{1i}^T\boldsymbol{b}_{1i} + \boldsymbol{T}_{ij}^T\boldsymbol{\alpha} + \boldsymbol{K}_{ij}^T\boldsymbol{u} + \varepsilon_{1i}(t_{ij}) \\ V_i^*(s_{ij}) = \boldsymbol{x}_{2i}^T\boldsymbol{\beta}_2 + \boldsymbol{z}_{2i}^T\boldsymbol{b}_{2i} + \boldsymbol{S}_{ij}^T\boldsymbol{\gamma} + \boldsymbol{\Lambda}_{ij}^T\boldsymbol{u}^* + \varepsilon_{2i}(s_{ij}). \end{cases}$$

The estimation method in Section 3 can be used to fit this joint mixed model. The latent variable formulation can also be adopted to model a dichotomous time-varying covariate with more than two categories.

6. Conclusions

We have presented a framework for analysing longitudinal data involving time-varying covariates that addresses some limitations of the existing methods. The main advantages of the proposed framework, while capturing the covariate process for time-varying covariates, are its effectiveness in handling the situations where the longitudinal response and time-varying covariates are measured at different time points, as well as its flexibility in both selecting covariance structures and choosing functions for the evolutions of variables over time (simple polynomial or general penalised spline functions). Also, this approach enables us to study and find out the temporal association between a time-varying covariate and the outcome of interest. Furthermore, it allows us to predict the response evolutions given the history of the time-varying covariate rather than just conditioning on a single observed value of the time-varying covariate. We have illustrated these advantages using a motivating data application from an AIDS cohort study conducted in Belgium, where we also saw that the recent methods cannot be applied and moreover the separate modelling of the outcomes would lead to misleading inferences for some parameters. This is because separate modelling ignores the dependence and association between the response and time-varying covariates. Separate modelling also would not easily allow predicting future outcomes conditional on a series of longitudinally measured outcome values. In this paper, we have not discussed the problem of missing data as there were no missing values in the AIDS data. Another advantage of the suggested framework is that the response and the time-varying covariate are treated equally, implying that results are valid under the assumption of missingness at random (MAR). In case this assumption would be believed to be violated, standard techniques such as multiple imputation could be incorporated in this framework. Finally, we have written a SAS programme for fitting the joint mixed model (12) to the AIDS data which can be found in the online Appendix D, along with an R implementation in the online Appendix E.

Acknowledgement

We would like to thank Prof. Kristel Van Laethem and Prof. Anne-Mieke Vandamme from KU Leuven and the University Hospitals Leuven, Belgium, for providing the AIDS data and answering our questions about the data.

References

Brumback, B. A., Ruppert, D. and Wand, M. P. (1999), 'Comment on Variable selection and function estimation in additive nonparametric regression using a data-based prior', *Journal of the American Statistical Association* **94**, 794–797.

Chen, Q., May, R. C., Ibrahim, J. G., Chu, H. and Cole, S. R. (2014), 'Joint modeling of longitudinal and survival data with missing and left-censored time-varying covariates', *Statistics in Medicine* **33**, 4560–4576.

Currie, I. D. and Durban, M. (2002), 'Flexible smoothing with P-splines: a unified approach', *Statistical Modelling* **2**, 333–349.

Drikvandi, R., Khodadadi, A. and Verbeke, G. (2012), 'Testing variance components in balanced linear growth curve models', *Journal of Applied Statistics* **39**, 563–572.

Drikvandi, R. and Noorian, S. (2019), 'Testing random effects in linear mixed-effects models with serially correlated errors', *Biometrical Journal* **61**, 802–812.

Drikvandi, R., Verbeke, G., Khodadadi, A. and PartoviNia, V. (2013), 'Testing multiple variance components in linear mixed-effects models', *Biostatistics* **14**, 144–159.

Ferrer, E. and McArdle, J. (2003), 'Alternative structural models for multivariate longitudinal data analysis', *Structural Equation Modeling* **10**, 493–524.

Fieuws, S. and Verbeke, G. (2006), 'Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles', *Biometrics* **62**, 424–431.

Ghosh, P. and Tu, W. (2009), 'Assessing sexual attitudes and behaviors of young women: a joint model with nonlinear time effects, time varying covariates, and dropouts', *Journal of the American Statistical Association* **104**, 474–485.

Gueorguieva, R. (2001), 'A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family', *Statistical Modelling* **1**, 177–193.

Hernán, M. A., Brumback, B. A. and Robins, J. M. (2002), 'Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures', *Statistics in Medicine* **21**, 1689–1709.

Hui, F. K., Müller, S. and Welsh, A. (2018), 'Sparse pairwise likelihood estimation for multivariate longitudinal mixed models', *Journal of the American Statistical Association* **113**, 1759–1769.

Jeffreys, H. (1961), *The theory of probability*, Oxford University Press.

Kim, S. and Albert, P. S. (2016), 'A class of joint models for multivariate longitudinal measurements and a binary event', *Biometrics* **72**, 917–925.

Kürüm, E., Jeske, D. R., Behrendt, C. E. and Lee, P. (2018), 'A copula model for joint modeling of longitudinal and time-invariant mixed outcomes', *Statistics in Medicine* **37**, 3931–3943.

Li, H., Zhang, Y., Carroll, R. J., Keadle, S. K., Sampson, J. N. and Matthews, C. E. (2017), 'A joint modeling and estimation method for multivariate longitudinal data with mixed types of responses to analyze physical activity data generated by accelerometers', *Statistics in Medicine* **36**, 4028–4040.

Lin, T.-I. and Wang, W.-L. (2013), 'Multivariate skew-normal at linear mixed models for multi-outcome longitudinal data', *Statistical Modelling* **13**, 199–221.

Miglioretti, D. L. and Heagerty, P. J. (2004), 'Marginal modeling of multilevel binary data with time-varying covariates', *Biostatistics* **5**, 381–398.

Proudfoot, J., Faig, W., Natarajan, L. and Xu, R. (2018), 'A joint marginal-conditional model for multivariate longitudinal data', *Statistics in Medicine* **37**, 813–828.

Rao, K., Drikvandi, R. and Saville, B. (2019), 'Permutation and Bayesian tests for testing random effects in linear mixed-effects models', *Statistics in Medicine* **38**, 5034–5047.

Roy, J., Alderson, D., Hogan, J. W. and Tashima, K. T. (2006), 'Conditional inference methods for incomplete poisson data with endogenous time-varying covariates: Emergency department use among hiv-infected women', *Journal of the American Statistical Association* **101**, 424–434.

Roy, J. and Lin, X. (2005), 'Missing covariates in longitudinal data with informative dropouts: Bias analysis and inference', *Biometrics* **61**, 837–846.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge University Press: Cambridge.

Stram, D. O. and Lee, J. W. (1994), 'Variance components testing in the longitudinal mixed effects model', *Biometrics* **50**, 1171–1177.

Sy, J., Taylor, J. and Cumberland, W. (1997), 'A stochastic model for the analysis of bivariate longitudinal AIDS data', *Biometrics* **53**, 542–555.

Thiébaut, R., Jacqmin-Gadda, H., Babiker, A., Commenges, D. and Collaboration, C. (2005), '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', *Statistics in Medicine* **24**, 65–82.

Verbeke, G. and Molenberghs, G. (2009), *Linear mixed models for longitudinal data*, New York: Springer Verlag.

Wand, M. P. (2003), 'Smoothing and mixed models', *Computational Statistics* **18**, 223–249.

Wasserman, L. (2000), 'Bayesian model selection and model averaging', *Journal of Mathematical Psychology* **44**, 92–107.

Xiang, D., Qiu, P. and Pu, X. (2013), 'Nonparametric regression analysis of multivariate longitudinal data', *Statistica Sinica* **23**, 769–789.

Zhao, L., Chen, T., Novitsky, V. and Wang, R. (2021), 'Joint penalized spline modeling of multivariate longitudinal data, with application to HIV-1 RNA load levels and CD4 cell counts', *Biometrics* **77**, 1061–1074.