Bayesian Joint Modeling of Longitudinal and Spatial Survival AIDS Data

Rui Martins[1,5], Giovani L. Silva[2,3] and Valeska Andreozzi[2,4]

Abstract

Joint analysis of longitudinal and survival data has received increasing attention in the recent years, especially for analyzing cancer and AIDS data. As both repeated measurements (longitudinal) and time-to-event (survival) outcomes are observed in an individual, a joint modeling is more appropriate because it takes into account the dependence between the two types of responses, which are often analyzed separately. We propose a Bayesian hierarchical model for jointly modeling longitudinal and survival data considering functional time and spatial frailty effects, respectively. That is, the proposed model deals with nonlinear longitudinal effects and spatial survival effects accounting for the unobserved heterogeneity among individuals living in the same region. This joint approach is applied to a cohort study of patients with HIV/AIDS in Brazil during the years 2002–2006. Our Bayesian joint model presents considerable improvements in the estimation of survival times of the Brazilian HIV/AIDS patients when compared with those ones obtained through a separate survival model and shows that the spatial risk of death is the same across the different Brazilian states.

Keywords: Joint model, Bayesian analysis, Repeated measurements, Time-to-event data, Spatial frailty.

[1] Centro de Investigação Interdisciplinar Egas Moniz (ciiEM), Escola Superior de Saúde Egas Moniz, Quinta da Granja, Monte de Caparica, 2829-511 Caparica, Portugal.

[2] Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL), Bloco C6 - Piso 4, Campo Grande, 1749-016 Lisboa, Portugal.

[3] Departamento de Matemática - Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal.

[4] Faculdade de Ciências Médicas da Universidade Nova de Lisboa, Campo Mártires da Pátria, 130, 1169-056 Lisboa, Portugal.

[5] Corresponding address: Escola Superior de Saúde Egas Moniz, Quinta da Granja, Monte de Caparica, 2829-511 Caparica - Portugal. E-Mail: ruimartins@egasmoniz.edu.pt

1 Introduction

In several biomedical studies, longitudinal and survival data are collected simultaneously but often separately analyzed. A joint analysis of these type of data has some advantages compared to the corresponding separate data analysis (Tsiatis and Davidian [1]). Conceptually a joint model assumes that a latent structure links both kinds of data. For instance, in clinical trials to evaluate new treatments in patients with the human immunodeficiency virus (HIV) the number of CD4$^+$ T lymphocyte (CD4 counts for short) has been proposed as a surrogate biomarker (Tsiatis *et al.* [2]). In a blood transfusion safety study involving AIDS-free survival times and longitudinal CD4 counts, Faucett and Thomas [3] concluded that the relative risk of AIDS was larger than those ones obtained by analysis of the component sub-models separately. Most joint models have been applied in AIDS and cancer contexts (see *e.g.* [4, 5, 6, 7]), but also in environmental and health studies, *e.g.*, radiation dose level [8], psychiatric disorder scale [9] and quality-of-life index [10].

Apart from incorporating the repeated measures into the survival model by regarding them as time-dependent covariates measured with error [8] there are other approaches to joint modeling. Namely link longitudinal and time-to-event outcomes via a subject-level [2] or a cluster-level [11] random effects. Another approach is to consider an unknown time-varying latent variable to link the two outcomes [4, 9]. Tsiatis and Davidian [1] is a comprehensive review of these models prior to 2004. Chapter 7 in Ibrahim *et al.* [12] devotes special attention to the subject, summarizing some of the most important joint models, including both Bayesian and frequentist perspectives. Rizopoulos published a book on joint models for longitudinal and time-to-event data with applications in R [13], and two R packages on this matter (JM [14] and JMbayes [15]), which are not able to work with structured spatial components. Recently, Gould *et al.* [16] reviewed currently available methods and software tools for carrying out joint analysis, including issues of implementation and interpretation.

Although many works have been published on joint analysis, it has not yet been routinely applied to the analysis of individuals who share an unobserved heterogeneity within a local health region (spatial frailty), as well as other functional effects of the longitudinal outcome. Usually, survival models with frailties assume independent random effects but we here consider that those effects are spatially correlated representing clusters of individuals living in a given region (see Chapter 9 in Banerjee *et al.* [17]). In this paper we focus on spatial survival analysis jointly modeled with a longitudinal biomarker that can have a functional effect (*e.g.* polynomial) providing more flexibility in the longitudinal trajectories. The model has a fully Bayesian approach being inspired in Henderson *et al.* [9] who proposed a likelihood-based joint model using the EM algorithm, linking the longitudinal and survival responses with a zero-mean latent bivariate Gaussian process and in Guo and Carlin [6] who addressed the problem of joint analysis without a spatial frailty by proposing a Bayesian hierarchical model using Markov chain Monte Carlo (MCMC) methods. Notice that models linked by random effects, which induces correlation between the longitudinal and survival components, are more friendly to implement via MCMC methods than via EM algorithm. However they are time consuming due to the high number of parameters.

The remainder of this paper evolves as follows. In Section 2 we describe the HIV/AIDS data set that motivated our joint modeling approach, whereas Section 3 outlines the spatial joint model with longitudinal and survival components. Section 4 discusses the related Bayesian model assessment by employing Cox-Snell residuals and multiple-imputation-based residuals, random visiting times and prediction of future values. In Section 5 we conduct an analysis of the HIV/AIDS data applying the proposed joint model, including residuals and predictions. Concluding remarks and discussion of important related issues are presented in Section 6. Finally some additional notes on sensitivity and predictive performance analysis as well as additional figures and tables are given in Supplementary Material (SuppMat).

2 The HIV/AIDS data

Brazilian National AIDS Program generated three major electronic databases [18]: (i) SINAN-AIDS (Information System for Notifiable Diseases of AIDS Cases) which is the most important electronic AIDS surveillance database, with all cases reported since 1980; (ii) SISCEL (Laboratory Test Control System) designed to monitor laboratory tests, such as CD4 counts and viral load tests for HIV/AIDS patients followed in the public health sector since 2002; (iii) SICLOM (System for Logistic Control of Drugs) developed to control the logistic for the AIDS treatment deliveries; it shares the patients list with SISCEL since 2002. These three databases have been previously combined in a single database with both HIV and AIDS cases using a process called record linkage, which was adopted by the Surveillance Unit of the Brazilian National AIDS Program [18]. This linkage strategy has been increasingly used in AIDS surveillance and research [19] to verify under-reporting of cases and eliminate the duplicated ones. In Brazil, that procedure has improved the quality of HIV/AIDS data information [18]. Notice that 2002–2006 can be considered as the first period with substantial information on both HIV/AIDS survival and CD4 exams, where 88 laboratories located in all twenty-seven Brazilian states were using SISCEL, covering 90% of all CD4 and viral load exams carried out by the public health sector. Cases diagnosed before 2002 were excluded because personal identifiers were not available in the mortality database for the entire country before that date [18].

For institutional reasons, we had access only to a simple random sample of the combined database, henceforth called HIV/AIDS data. The related information includes $N = 4{,}653$ patients, corresponding to 10% of the total number of diagnosed individuals during the period 2002–2006. The time-to-event after HIV/AIDS diagnosis is defined as the time period, in years, between the date of diagnosis and the date of death (available if death happened before December 31st 2006, and censored otherwise). A longitudinal measure of immunologic and virologic status (CD4 counts) was collected. Apart from those two outcomes, the explanatory variables included were: (i) `age`, coded 0 (15–49 years) and 1 (at least 50 years); (ii) `gender`, coded 0 (female) and 1 (male); (iii) previous opportunistic infection (`PrevOI`) at study entry, coded 0 (without PrevOI) and 1 (with PrevOI); (iv) patient's Brazilian state of residence (state). As referred by Souza-Jr *et al.* [20], the age cut-off was chosen based on the Ministry of Health recommendations, as the group aged over 50 showed a higher proportion of delayed initiation of the therapy when compared to the population group aged 15-49 years.

The CD4 counts distribution by `gender`, `age` and `PrevOI` indicates a high degree of skewness toward high CD4 counts (Figure 1 – SuppMat), suggesting a power transformation for that outcome to achieve the normality (see Taylor and Law [21] for a discussion about the power transformation of CD4 counts). There were about 7% of dead patients, 88% were between 15 and 49 years, 60% of patients were males of whom 8% died, and 61% had no previous infection, whereas 6.7% lived in the Central-West region, 11.5% in the Northeast, 4.8% in the North, 60% in the Southeast region and 16.7% in the South (Table 5 – SuppMat). The median of the CD4 counts was 245 cells/mm^3 (226 cells/mm^3 for males and 263 cells/mm^3 for females). All patients made on average 4.62 CD4 exams resulting in a total of 21,508 observations (Figure 2 – SuppMat).

3 Joint modeling framework

Suppose a set of N subjects coming from K regions with n_k patients each, $\sum_{k=1}^{K} n_k = N$, followed over a certain time period for which were collected both longitudinal and survival response variables, as well as a set of explanatory variables. Our goal is to understand the relation of all these variables modeling the *true* value of the longitudinal outcome at time point t, $y^*_{ik}(t)$, and the survival component, T^*_{ik}, to a certain endpoint for the ith patient living in the kth region, $i=1,\dots,n_k$, $k=1,\dots,K$. Time-to-event, T^*_{ik}, may be subject to the usual right censoring mechanism and then only the minimum, T_{ik}, of the time-to-event and censoring time, C_{ik}, is observed, $T_{ik} \equiv \min(T^*_{ik}, C_{ik})$. We define the event indicator as δ_{ik}, where $\delta_{ik} = 1$ indicates a failure ($T^*_{ik} \leq C_{ik}$) and $\delta_{ik} = 0$ indicates a right censored observation ($T^*_{ik} > C_{ik}$).

Longitudinal outcomes are collected on each subject intermittently at some set of times $\{t_{ikj} \leq T_{ik} : i = 1,\dots,n_k; k = 1,\dots,K; j = 1,\dots,n_{ik}\}$ producing the observed vector $\mathbf{y}_{ik} = (y_{ik1},\dots,y_{ikn_{ik}})^\top$, where $y_{ikj} \equiv y_{ik}(t_{ikj})$ and n_{ik} is the repeated measurements number of the longitudinal outcome for the ikth individual. Note that the observed value of the longitudinal response at time t_{ikj}, $y_{ik}(t_{ikj})$, is the true value with error, i.e.

$$y_{ik}(t_{ikj}) \;=\; y^*_{ik}(t_{ikj}) + e_{ik}(t_{ikj}), \tag{1}$$

where $e_{ik}(t_{ikj}) \equiv e_{ikj}$ is an intra-subject error, $j=1,\dots,n_{ik}$, $i=1,\dots,n_k$ and $k=1,\dots,K$. Now, as in Henderson *et al.* [9], and Guo and Carlin [6], we introduce the joint model starting with the longitudinal and survival components separately.

3.1 Longitudinal component

We postulate a mixed effects model to describe the longitudinal latent process in (1), $y^*_{ik}(t_{ikj})$, by specifying it as a function of “fixed” and random effects,

$$y^*_{ik}(t_{ikj}) \;=\; \mu_{ik}(t_{ikj}) + W_{ik}(t_{ikj}), \tag{2}$$

where $\mu_{ik}(t_{ikj})$ is the “fixed” component that can be described by a curve (polynomial) growth model, providing more flexibility in the longitudinal trajectories, and $W_{ik}(t_{ikj})$ is the random component for which can be considered a zero mean latent Gaussian process. Specifically we will define: $\mu_{ik}(t_{ikj}) = \mathbf{x}^\top_{1ik}(t_{ikj})\boldsymbol{\beta}_1$, where $\boldsymbol{\beta}_1$ is the population parameters vector (fixed effects) related to a covariate vector $\mathbf{x}_{1ik}(t_{ikj})$, and $W_{ik}(t_{ikj}) = \mathbf{z}^\top_{1ik}(t_{ikj})\boldsymbol{b}_{ik}$, where $\mathbf{z}_{1ik}(t_{ikj})$ denotes a design vector corresponding to a random effects vector, $\boldsymbol{b}_{ik}$, $i=1,\dots,n_k$, $k=1,\dots,K$, $j=1,\dots,n_{ik}$.

3.2 Spatial survival component

A traditional framework to link a longitudinal process to a disease outcome is the relative risk model (Kalbfleisch and Prentice [22]),

$$\begin{aligned} h_{ik}(t \mid \mathcal{Y}^*_{ik}(t), \mathbf{x}_{2ik}) &\equiv \lim_{dt \to 0} \mathbb{P}\{t \leq T^*_{ik} < t + dt \mid T^*_{ik} \geq t, \mathcal{Y}^*_{ik}(t), \mathbf{x}_{2ik}\} dt^{-1} \\ &= h_0(t) \exp\{\boldsymbol{\beta}_2^\top \mathbf{x}_{2ik} + \gamma\, y^*_{ik}(t)\}, \end{aligned} \tag{3}$$

where $\mathcal{Y}^*_{ik}(t) = \{y^*_{ik}(u), 0 \leq u < t\}$ denotes the history of the *true* and *unobserved* longitudinal process up to time point t, $\boldsymbol{\beta}_2$ is the vector of regression parameters associated to the vector of covariates $\mathbf{x}_{2ik}$, $h_0(t)$ is the baseline risk function and γ quantifies the effect of the underlying longitudinal outcome to the risk for an event. For instance, it measures the effect of CD4 counts to the risk of death in the HIV/AIDS data.

A common criticism to the model (3) has been its dependence regarding the history of the longitudinal biomarker up to time t, $\mathcal{Y}^*(t)$. First, it can include improper extrapolation beyond the range of the longitudinal measurements because the last registration of the biomarker may be quite temporally distanced from the moment of failure. Second, it is not obvious that the imputed value for the longitudinal variable is the more relevant biological summary. For example, changes in the slope of the trajectory may be more predictive of patient's survival time than the current value of the marker.

Alternatively, one can induce the association between the survival and longitudinal processes by using only the Gaussian process $W_{ik}(t)$ in (2). In addition, assuming a Weibull baseline hazard function, *i.e.*, $h_0(t) = at^{a-1}$, model (3) can be replaced by

$$h_{ik}(t) \;\; = \;\; at^{a-1} \exp\{\mathbf{x}_{2ik}^\top(t)\boldsymbol{\beta}_2 + \boldsymbol{\gamma}^\top g(W_{ik}(t))\}, \tag{4}$$

where $g(\cdot)$ is a link function specifying which components of the longitudinal process are related to $h_{ik}(.)$, $\boldsymbol{\beta}_2$ represents the vector of regression coefficients associated with the vector of the possibly time-dependent explanatory variables, $\mathbf{x}_{2ik}(t)$ (may coincide with $\mathbf{x}_{1ik}(t)$); $\boldsymbol{\gamma}$ denotes a vector of parameters which measure the association between the survival and longitudinal components, and $a > 0$ is the shape parameter of the Weibull distribution, denoted by $\mathcal{W}(a, \lambda)$, being λ the $\exp(\cdot)$ function in (4).

Sometimes individuals are clustered in a hierarchical structure such that subjects within the same cluster share a common frailty, for example, the incidence of some diseases is lower or higher in regions with better health services or more environmental problems, respectively. We consider here a special case of the frailty survival model, introducing region-specific random effects exhibiting spatial dependence (Banerjee *et al.* [17]). In order to accommodate this spatial extra-variation we extend our survival model by adding appropriate random effects into its hazard function (4). Let Q_k be the spatial effect of latent risk factors related to the kth region, $k=1,\dots,K$. Thus, the spatial survival model is defined by

$$h_{ik}(t) \;=\; at^{a-1}\exp\{\mathbf{x}_{2ik}^{\top}(t)\boldsymbol{\beta}_2+\boldsymbol{\gamma}^{\top}g(W_{ik}(t))+Q_k\}, \tag{5}$$

where Q_k captures the residual or unexplained log-relative risk of an event (*e.g.* death) in the kth region.

The HIV/AIDS data described in Section 2 have well-defined spatial boundaries associated with the residence geographic regions (Brazilian states) which is a typical example of areal or lattice data. In addition, we believe that there exists a "neighborhood effect", where neighboring locations have a similar risk-of-death, and also a "grouping effect", where subjects living in the same region are assumed to have identical risk. We will develop this matter more deeply in Subsection 3.4, namely discussing the appropriate prior distributions for Q_k.

The introduction of the spatial random effects in (5) serves three main purposes: (i) capturing the unexplained risk-of-death in each of the 27 Brazilian states; (ii) mapping the spatial risk-of-death for an epidemiological interpretation purpose (Figure 2) and (iii) investigate the need to include spatially varying covariates (*vide* Subsection 5.2).

3.3 Likelihood

We propose a spatial joint model assuming that the longitudinal (2) and spatial survival (5) components share the same set of time-independent random effects, $\boldsymbol{b}_{ik}$. We will define

$$W_{ik}(t) \;=\; \mathbf{z}_{1ik}^{\top}(t)\,\boldsymbol{b}_{ik} \tag{6}$$

and

$$g(W_{ik}(t)) \;=\; \mathbf{Z}_{2ik}(t)\,\boldsymbol{b}_{ik}, \tag{7}$$

where $\mathbf{z}_{1ik}(t)$ and $\mathbf{Z}_{2ik}(t)$ are appropriate design vector and matrix, respectively. For instance, considering $\boldsymbol{b}_{ik}=(b_{1ik},b_{2ik})^{\top}$, $\boldsymbol{\gamma}=(\gamma_1,\gamma_2)^{\top}$ and $\mathbf{Z}_{2ik}(t)$ an identity matrix, we may have $\mathbf{z}_{1ik}^{\top}(t)\,\boldsymbol{b}_{ik}=b_{1ik}+b_{2ik}\,t$ and $\boldsymbol{\gamma}^{\top}\mathbf{Z}_{2ik}(t)\,\boldsymbol{b}_{ik}=\gamma_1 b_{1ik}+\gamma_2 b_{2ik}$. This specification allows different subjects to have different baseline repeated measures and different time trends for longitudinal responses during the follow-up. Note that $\boldsymbol{\gamma}=0$ means a separated analysis of the longitudinal and survival data.

For the spatial joint model hereafter cited by joint model (6)-(7), we are assuming that the repeated measures and time-to-event are independent given the random effects and the covariates of interest. We also assume a normal distribution, $\mathcal{N}(0, \sigma^2)$, for the measurement errors, e_{ikj}. Due to the significant separation in time between observations, correlation among residuals over time is assumed to be negligible, so the error e_{ikj} belongs to a sequence of independent and identically distributed random variables assumed as independent of the random effects, $\boldsymbol{b}_{ik}$. Rizopoulos *et al.* [23] remarked that as the number of repeated measurements per subject increases, a misspecification of the random effects distribution has a minimal effect in the parameter estimates and their standard errors. Under a matrix approach, we assume that $\boldsymbol{b}_{ik}|\boldsymbol{\Sigma} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$. The structure of the $p \times p$ covariance matrix, $\boldsymbol{\Sigma}$, describes the association between repeated measures of the observed longitudinal data. Because $\boldsymbol{b}_{ik}$ links both the longitudinal and survival processes it accounts for both the association between the two model components and the correlation between the repeated measurements in the longitudinal process.

Let $\boldsymbol{\theta}$ be a generic vector of all parameters of the spatial joint model and $L(\boldsymbol{\theta}|\mathcal{D})$ the related likelihood function, where $\mathcal{D} = \{\mathbf{y}_{ik}, T_{ik}, \delta_{ik}; i = 1, \ldots, n_k, k = 1, \ldots, K\}$ represents the observed data, composed of the survival (T_{ik}, δ_{ik}) and longitudinal $\mathbf{y}_{ik} = (y_{ik1}, \ldots, y_{ikn_{ik}})^\top$ components, whose elements are observations from the normal distribution, $\mathcal{N}(y^*_{ikj}, \sigma^2)$. Covariates in $\mathcal{D}$ have been suppressed to facilitate the exposition. $L_{ik}(\boldsymbol{\theta}|\mathcal{D})$ denotes the contribution of the ikth individual to the likelihood, $L(\boldsymbol{\theta}|\mathcal{D})$, defined as

$$L_{ik}(\boldsymbol{\theta}|\mathcal{D}) = L_{1ik}(\boldsymbol{b}_{ik}, \boldsymbol{\beta}_1, \sigma^2|\mathcal{D}) \times L_{2ik}(\boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, Q_k|\mathcal{D}), \tag{8}$$

where $L_{1ik}(\cdot|\mathcal{D})$ and $L_{2ik}(\cdot|\mathcal{D})$ denote the corresponding contributions for the longitudinal and survival components, respectively. The related longitudinal contribution, $L_{1ik}(\boldsymbol{b}_{ik}, \boldsymbol{\beta}_1, \sigma^2|\mathcal{D})$, is

$$\prod_{j=1}^{n_{ik}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\left[y_{ikj} - \mathbf{x}_{1ik}^\top(t_{ikj})\boldsymbol{\beta}_1 - \mathbf{z}_{1ik}^\top(t_{ikj})\boldsymbol{b}_{ik}\right]^2}{2\sigma^2}\right\}, \tag{9}$$

and the corresponding survival contribution, $L_{2ik}(\boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, Q_k|\mathcal{D})$, is

$$h_{ik}(T_{ik}|\boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, Q_k)^{\delta_{ik}} \times \exp\left\{-\int_0^{T_{ik}} h_{ik}(s|\boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, Q_k)ds\right\}, \tag{10}$$

where $h_{ik}(\cdot)$ is the hazard function (5). Consequently, the likelihood of the spatial joint model (6)-(7) is the product of all N individual contributions to the likelihood:

$$L(\boldsymbol{\theta}|\mathcal{D}) = \prod_{k=1}^{K}\prod_{i=1}^{n_k} L_{ik}(\boldsymbol{\theta}|\mathcal{D}) = \prod_{k=1}^{K}\prod_{i=1}^{n_k} L_{1ik}(\boldsymbol{b}_{ik}, \boldsymbol{\beta}_1, \sigma^2|\mathcal{D}) \times L_{2ik}(\boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, Q_k|\mathcal{D}). \tag{11}$$

For example, if $\boldsymbol{b}_{ik} = (b_{1ik}, b_{2ik})^\top$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)^\top$, $\mathbf{x}_{2ik}(t) = \mathbf{x}_{2ik}\ \forall\, t$, $\mathbf{z}_{1ik}^\top(t)\, \boldsymbol{b}_{ik} = b_{1ik} + b_{2ik}\, t$, $\boldsymbol{\gamma}^\top \mathbf{Z}_{2ik}(t)\, \boldsymbol{b}_{ik} = \gamma_1 b_{1ik} + \gamma_2 b_{2ik}$, and $a = 1$ in Equation (5), corresponding to an exponential distribution, the likelihood (11) is expressed as

$$\begin{aligned}\prod_{k=1}^{K}\prod_{i=1}^{n_k}\prod_{j=1}^{n_{ik}}\tfrac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\tfrac{1}{2\sigma^2}\left[y_{ikj}-\mathbf{x}_{1ik}^\top(t_{ikj})\boldsymbol{\beta}_1-b_{1ik}-b_{2ik}t_{ikj}\right]^2\right\}\times\\ \times\prod_{k=1}^{K}\prod_{i=1}^{n_k}\left\{\exp[\mathbf{x}_{2ik}^\top\boldsymbol{\beta}_2+\gamma_1 b_{1ik}+\gamma_2 b_{2ik}+Q_k]\right\}^{\delta_{ik}}\times\\ \times\exp\left\{-\sum_{k=1}^{K}\sum_{i=1}^{n_k}T_{ik}\exp[\mathbf{x}_{2ik}^\top\boldsymbol{\beta}_2+\gamma_1 b_{1ik}+\gamma_2 b_{2ik}+Q_k]\right\}.\end{aligned}\tag{12}$$

3.4 Bayesian approach

For a Bayesian approach of the spatial joint model (6)-(7), we added prior distributions for all model parameters. In particular, for the longitudinal component we took, respectively, multivariate normal, $\mathcal{N}_{p_1}(\mathbf{0},\mathbf{V}_1^*)$, and inverse gamma, $\mathcal{IG}(c_1,d_1)$, priors for the vector of fixed effects, $\boldsymbol{\beta}_1$, and the measurement error variance, σ^2. For the survival component, we designated normal priors for both $\boldsymbol{\beta}_2$ and $\boldsymbol{\gamma}$, respectively denoted by $\mathcal{N}_{p_2}(\mathbf{0},\mathbf{V}_2^*)$ and $\mathcal{N}_{p_3}(\mathbf{0},\mathbf{V}_3^*)$. A gamma prior distribution, $\mathcal{G}(c_2,d_2)$, for the Weibull shape parameter, a, and an inverse Wishart prior, $\mathcal{IW}ish(\boldsymbol{V}_4^*,\kappa)$, for the covariance matrix, $\boldsymbol{\Sigma}$, of the random effects, $\boldsymbol{b}_{ik}$, with $\boldsymbol{V}_4^*$ representing a $p\times p$ positive definite matrix prespecified and with κ degrees of freedom. In consonance with Guo and Carlin [6], we chose very low precision (high variance) for these priors, including an inverse Wishart prior that is vague but does provide at least some shrinkage of the random effects toward 0, ensuring good identifiability of the main effects.

Concerning the spatial frailty, Q_k, we incorporated that dependence by specifying an intrinsic conditionally autoregressive (ICAR) prior proposed in [24], i.e. the prior on $\boldsymbol{Q}=(Q_1,\ldots,Q_K)$ is specified as a set of K univariate full conditional distributions, $Q_k|\sigma_Q^2\sim ICAR(\sigma_Q^2)$, allowing us to deal with the risk's spatial autocorrelation, capturing the "local" extra-variability in the log-relative risk so that nearby regions will have more similar risks [17] (structured effect). For the spatial frailties variance, σ_Q^2, we assigned an inverse gamma prior, $\mathcal{IG}(c_3,d_3)$, similarly assumed with high variance.

Although the convenience of the ICAR prior, one may certainly employ independent priors for each spatial random effect. For instance replacing Q_k by V_k in (5) we may consider an exchangeable normal prior, $V_k|\sigma_V^2\sim\mathcal{N}(0,\sigma_V^2)$, if independence across areal units is a plausible assumption (unstructured effect). This specification is appropriate if the covariates included in (5) account for all of the spatial structure, leaving V_k to account for the "global" region heterogeneity. We can also allow for both structured and unstructured random effects, however, it requires two random effects to be estimated for each region, whereas only their sum is identifiable from the data. Another problem is the decrease in algorithm performance because identifiability problems [17]. Discussion of these issues is given in [25].

The joint posterior distribution of the Bayesian hierarchical spatial joint model (6)-(7), denoted by $p(\boldsymbol{\theta}|\mathcal{D})$, is proportional to

$$\begin{aligned} L(\boldsymbol{\theta}|\mathcal{D}) \times \pi(\boldsymbol{\beta}_1|\mathbf{V}_1^*) \times \pi(\boldsymbol{\beta}_2|\mathbf{V}_2^*) \times \textstyle\prod_{k=1}^{K} \prod_{i=1}^{n_k} \pi(\mathbf{b}_{ik}|\boldsymbol{\Sigma}) \times \prod_{k=1}^{K} \pi(Q_k|\sigma_Q^2) \times \\ \times \pi(\boldsymbol{\gamma}|\mathbf{V}_3^*) \times \pi(\boldsymbol{\Sigma}|\boldsymbol{V}_4^*, \kappa) \times \pi(\sigma^2|c_1, d_1) \times \pi(a|c_2, d_2) \times \pi(\sigma_Q^2|c_3, d_3), \end{aligned} \tag{13}$$

where $L(\boldsymbol{\theta}|\mathcal{D})$ is defined in (11) and $\pi(\cdot|\cdot)$ generically denotes a prior distribution specified in the previous paragraphs. Typically, the marginal posterior distributions cannot be carried out in closed form and, therefore, to avoid the analytic intractable integral problem involved in the marginalized functions, we propose to apply MCMC methods in OpenBUGS ([26]).

4 Model assessment

Due to recent computational advances, sophisticated techniques for Bayesian model assessment are becoming increasingly popular (see some summary in [12, 17, 27]). The next two Subsections are devoted to two of those techniques in a joint models context. First, the prediction of future values and then a residual analysis.

4.1 Prediction of future values

The ability to incorporate the trajectory of the longitudinal biomarker over time in a survival model gives to joint models the possibility to act as a dynamic prognostic tool, which can drive to a more accurate clinical decision. For example, the full history of CD4 counts observed in a patient with HIV/AIDS can be used to predict his survival probability in the coming years, from the time of the last visit or after censoring. If the CD4 trajectory indicates an increasing risk of death, the physician may decide to change the therapy in order to slow the progression of the disease.

Proust-Lima and Taylor [7] proposed a dynamic prognostic tool for joint models providing a measure of variability obtained from the parameters asymptotic distribution and validating this prognostic tool based on predictive accuracy measures. Rizopoulos [28] focused particularly on the assessment of the predictive ability of the longitudinal outcome for the survival outcome, assessing how well the former is capable of discriminating between subjects who will experience or not the event within a certain period. Sweeting and Thompson [29] have compared shared random effects models with two approximation-based approaches, concluding that these latter should be avoided since they can severely underestimate any association between the longitudinal and event processes.

Suppose we have a series of repeated measurements from a new individual, along with its survival information up to time t. Let $\tilde{\mathcal{D}} = \{\tilde{\mathbf{y}}, \tilde{T}=t, \tilde{\delta}=0\}$ be the summarized data for this individual. Inferences on a future longitudinal value for this individual at time $s > t$, denoted by $\tilde{y}(s)$, can be obtained from its posterior predictive distribution conditional on the existing data, $\mathcal{D}$, and the new data, $\tilde{\mathcal{D}}$ [7, 28, 29],

$$p(\tilde{y}(s) \mid \mathcal{D}, \tilde{\mathcal{D}}) = \iint p(\tilde{y}(s)|\tilde{\mathcal{D}}, \tilde{\boldsymbol{b}}, \boldsymbol{\theta})\, p(\tilde{\boldsymbol{b}}|\tilde{\mathcal{D}}, \boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}\, d\tilde{\boldsymbol{b}}, \tag{14}$$

where $\tilde{\boldsymbol{b}}$ represents the random effects vector of the new individual, $p(\tilde{\boldsymbol{b}}|\tilde{\mathcal{D}}, \boldsymbol{\theta})$ is here the posterior distribution of the new random effects and $p(\boldsymbol{\theta}|\mathcal{D})$ is the joint posterior distribution defined in (13). Similarly, the posterior predictive probability for the time-to-event, $\tilde{T}^*$, of the new individual at time s given survival up to time t, is expressed as

$$p(\tilde{T}^* > s \mid \mathcal{D}, \tilde{T}^* > t, \tilde{\mathbf{y}}) = \iint \frac{\tilde{S}(s \mid \tilde{\mathbf{y}}, \tilde{\boldsymbol{b}})}{\tilde{S}(t \mid \tilde{\mathbf{y}}, \tilde{\boldsymbol{b}})}\, p(\tilde{\boldsymbol{b}}|\tilde{\mathcal{D}}, \boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta}\, d\tilde{\boldsymbol{b}}, \tag{15}$$

where $\tilde{S}(\cdot \mid \tilde{\mathbf{y}}, \tilde{\boldsymbol{b}})$ is the survival function for the new individual [7, 28, 29].

4.2 Residual analysis

Because model selection measures provide no information about the absolute adequacy of the models, other diagnostic tools (*e.g.* residuals analysis) are needed to assess the model adequacy. In checking model assumptions via the inspection of residuals, Dobson and Henderson [30] pointed out some properties of the residuals conditioned by the dropout information and Zhu *et al.* [31] developed a series of influence measures to quantify the degree of perturbation introduced into the model during a sensitivity analysis. Recently, for longitudinal and survival joint models Zhang *et al.* [32] developed a novel decomposition of the well-known model selection criteria AIC and BIC in order to assess the fit of each component of the joint model, whereas Park and Qiu [33] discussed several model selection criteria applying them to the joint model for comparing two crossing hazard rate functions proposing hypothesis testing and graphical methods for model diagnostics. Rizopoulos *et al.* [34] discussed the difficulty in using standard model diagnostics in joint models because of the nonrandom dropout in the longitudinal outcome caused by the occurrence of events proposing a multiple-imputation-based approach as diagnostic and model-assessment tool.

For the residual analysis of the survival component of the spatial joint model (6)-(7), we can employ Cox-Snell residuals [35] by using the well-known relationship $r_{ik}^{CS}(t|\boldsymbol{\theta}) \equiv \int_0^t h_{ik}(s|\boldsymbol{\theta})ds = -\log S_{ik}(t|\boldsymbol{\theta})$. Conforming to Rizopoulos and Ghosh [36], we get $r_{ik}^{CS}(t)$ calculating the expected value for $r_{ik}^{CS}(t|\boldsymbol{\theta})$ averaged over the parameters posterior distribution, $p(\boldsymbol{\theta}|\mathcal{D})$, i.e.

$$r_{ik}^{CS}(t) = \mathbb{E}_{\boldsymbol{\theta}|\mathcal{D}}\left[r_{ik}^{CS}(t|\boldsymbol{\theta})\right] = \int r_{ik}^{CS}(t|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}. \tag{16}$$

In practice, we compute $r_{ik}^{CS}(t)$ at the observed event times T_{ik}, being censored if the event of interest did not occur for the related individuals. In order to check the fit of the survival model taking into account the censoring times, we can graphically compare the associated Kaplan–Meier estimate for $r_{ik}^{CS}(T_{ik})$ with the survival function of the unit exponential distribution [36].

Regarding the residual analysis for the longitudinal component of the joint model, we can make use of the widely used standardized marginal and standardized subject-specific residuals for mixed models [37], respectively defined by

$$\mathbf{r}_{ik}^{ym} = \hat{\mathbf{V}}_{ik}^{-\frac{1}{2}}\left[\mathbf{y}_{ik} - \mathbf{X}_{1ik}\hat{\boldsymbol{\beta}}_1\right] \quad \text{and} \quad \mathbf{r}_{ik}^{ys} = \hat{\sigma}^{-1}\left[\mathbf{y}_{ik} - \mathbf{X}_{1ik}\hat{\boldsymbol{\beta}}_1 - \mathbf{Z}_{1ik}\hat{\boldsymbol{b}}_{ik}\right], \tag{17}$$

where $\hat{\boldsymbol{\beta}}_1$, $\hat{\sigma}$, $\hat{\boldsymbol{b}}_{ik}$ and $\hat{\mathbf{V}}_{ik}$ are posterior estimates (*e.g.* mean or median), respectively, for the vector of regression coefficients $\boldsymbol{\beta}_1$, the residual standard deviation σ, the vector of the random effects $\boldsymbol{b}_{ik}$ and the covariance matrix of the repeated measurements $\mathbf{y}_{ik}$, *i.e.*, $\mathbf{V}_{ik} = \mathbf{Z}_{1ik}\boldsymbol{\Sigma}\mathbf{Z}_{1ik}^{\top} + \sigma^2\mathbf{I}$, with $\mathbf{I}$ denoting the identity matrix of appropriate dimensions, and $\mathbf{X}_{1ik}$ and $\mathbf{Z}_{1ik}$ are design matrices whose rows are, respectively, $\mathbf{x}_{1ik}^{\top}(t_{ikj})$ and $\mathbf{z}_{1ik}^{\top}(t_{ikj})$.

Rizopoulos *et al.* [34] pointed out some issues in using the residuals (17) in joint models, especially because the occurrence of events causes a nonrandom dropout in the longitudinal outcome. Accordingly, they proposed to augment the observed longitudinal data with a multiple-imputation-based scheme, under the assumed joint model. The main advantage of using both the augmented and observed data is to calculate residuals, such as (17), that inherit now the properties of the complete data model and, therefore, they can be directly used in diagnostic plots without requiring to take dropout into account.

Let $\mathbf{y}_{ik}^m = \{y_{ik}(t_{ikj}) \equiv y_{ikj} : t_{ikj} \geq T_{ik}, j = 1, \ldots, n'_{ik}\}$ be the missing part of the longitudinal response vector, where n'_{ik} is the total of augmented measurements concerning the ikth individual. The multiple-imputation-based method consists of sampling from the posterior predictive distribution of $\mathbf{y}_{ik}^m$

$$p(\mathbf{y}_{ik}^m \mid \mathcal{D}) = \int p(\mathbf{y}_{ik}^m|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}. \tag{18}$$

Notice that (i) we are assuming that $\mathbf{y}_{ik}^m$ and $\mathcal{D}$ are independent, conditionally on $\boldsymbol{\theta}$; (ii) the predicted $\mathbf{y}_{ik}^m$ and observed $\mathbf{y}_{ik}$ values form the complete longitudinal data.

If the visiting times, t_{ikj}, of the repeated measurements are determined by the patients themselves, we should model that random visiting process before obtaining the multiple-imputation-based residuals (Rizopoulos *et al.* [34]). Let u_{ikq} denote the time elapsed between $(q{-}1)$th and qth visits for the ikth subject with n_{ik} measures/visits, $q{=}2,\ldots,n_{ik}$. Assuming that all subjects have at least one measurement and the visiting process is non-informative, we can let the distribution of the elapsed time, u_{ikq}, to depend only on the last observed longitudinal measurement, *i.e.*, $p(u_{ikq}|y_{ik(q{-}1)};\boldsymbol{\theta}_v)$, where $\boldsymbol{\theta}_v$ is the vector of the visiting process parameters and $t_{ikq} = t_{ik(q{-}1)} + u_{ikq}$ (see [34] for different formulations of the visiting process and for a comprehensive simulation scheme of the elapsed times).

We propose to model the elapsed times vector, $\mathbf{u}_{ik} = (u_{ik2},\ldots,u_{ikn_{ik}})^\top$, by using a Weibull model, $\mathcal{W}(a_v,\lambda_v)$, with individual and spatial frailties expressed in terms of its hazard function

$$h_v(u_{ikq}|\mathbf{x}_{vik},\boldsymbol{\theta}_v) = a_v u_{ikq}^{a_v-1}\exp(\mathbf{x}_{vik}^\top\boldsymbol{\beta}_v + Q_k)\omega_{ik}, \tag{19}$$

where $\boldsymbol{\beta}_v$ is the vector of regression coefficients associated with the design vector $\mathbf{x}_{vik}$ containing possibly a functional form of the last observed longitudinal response, $y_{ik(q-1)}$; ω_{ik} is an individual frailty taking a gamma distribution, $\mathcal{G}(\eta,\eta)$, and Q_k is a spatial frailty as in (5). We note that in Rizopoulos *et al.* [34] there is not a spatial frailty in the visiting process definition. In order to obtain the estimates of the various elapsed times, we propose to resort to the posterior distribution of $\boldsymbol{\theta}_v$ given all the visit data, $\mathcal{D}_v$,

$$\begin{aligned} p(\boldsymbol{\theta}_v \mid \mathcal{D}_v) &\propto \prod_{k=1}^{K}\prod_{i=1}^{n_{ik}} L_{vik}(\boldsymbol{\theta}_v|\mathbf{u}_{ik},\mathbf{x}_{vik}) \times \pi(\boldsymbol{\theta}_v) \\ &= \prod_{k=1}^{K}\prod_{i=1}^{n_{ik}}\prod_{q=2}^{n_{ik}} \{[h_v(u_{ikq}|\mathbf{x}_{vik},\boldsymbol{\theta}_v)]\,[S_v(u_{ikq}|\mathbf{x}_{vik},\boldsymbol{\theta}_v)]\} \times \pi(\boldsymbol{\theta}_v), \end{aligned} \tag{20}$$

where $L_{vik}(.|.)$ is the ikth individual contribution to the likelihood, $S_v(.)$ is the Weibull survival function and $\pi(\boldsymbol{\theta}_v)$ is the prior distribution on $\boldsymbol{\theta}_v$. Then we carry on using the posterior distribution, $p(\boldsymbol{\theta}_v|\mathcal{D}_v)$, to simulate the future elapsed times, u_{ikq}, $q = n_{ik}+q'$, $q'{=}1,\ldots,n'_{ik}$, from its posterior predictive distribution

$$p(u_{ikq} \mid \mathcal{D}_v) = \int p(u_{ikq}|\boldsymbol{\theta}_v)\, p(\boldsymbol{\theta}_v|\mathcal{D}_v)\, d\boldsymbol{\theta}_v, \tag{21}$$

in order to get the missing y^m_{ikq}'s at times $t_{ikq} = t_{ik(q-1)} + u_{ikq}$, $q = n_{ik} + n'_{ik}$, $q' = 1, \ldots, n'_{ik}$, via its predictive distribution (18). Along with the observed data, $\mathbf{y}_{ik}$, we calculate the residuals (17) for the complete longitudinal data of the spatial joint model (6)-(7).

5 Analysis of the HIV/AIDS data

5.1 Spatial joint model

The HIV/AIDS data described used in this work have well-defined spatial boundaries associated with the residence geographic regions (Brazilian states) which is a typical example of areal or lattice data. In addition, we believe that there exists a "neighborhood effect", where region's risk-of-death is similar to that of neighboring locations, and also a "grouping effect", where subjects living in the same region are assumed to have identical risk. The methodology developed in Section 3 is now applied to the HIV/AIDS data described in Section 2. Based on exploratory analysis partially introduced in Section 2, we assumed a square root transformation of the longitudinal measures (*i.e.* $\sqrt{\text{CD4}}$), as well as the particular case ($a = 1$) of the Weibull survival model (exponential survival model). Those practical considerations are in agreement with other AIDS joint analysis, such as Guo and Carlin [6].

Let y_{ikj} denote the square root of the jth CD4 count measurement on the ith patient living in the kth Brazilian state, $j = 1, \ldots, n_{ik}$, whereas (T_{ik}, δ_{ik}) represents both the AIDS survival time and the death indicator of the patient, $i = 1, \ldots, n_k$, $k = 1, \ldots, 27$. Several spatial joint models were fitted. For the longitudinal measures, an individual polynomial trajectory inside the random effects model was considered to account for patient-specific $\sqrt{\text{CD4}}$ counts over time (see *e.g.* Wu and Zhang [38] for some discussion on polynomial mixed-effects models for longitudinal data). Particularly, we assumed $y_{ikj}|\boldsymbol{b}_{ik}, \boldsymbol{\beta}_1, \sigma^2 \sim \mathcal{N}(y^*_{ikj}, \sigma^2)$, where $y^*_{ikj} = \mu_{ik}(t_{ikj}) + W_{ik}(t_{ikj})$, and

$$\begin{aligned} \mu_{ik}(t_{ikj}) &= \beta_{11} + \beta_{12} t_{ikj} + \beta_{13} t^2_{ikj} + \beta_{14} t^3_{ikj} + \beta_{15}\text{gender}_{ik} + \beta_{16}\text{age}_{ik} + \beta_{17}\text{PrevOI}_{ik} \\ W_{ik}(t_{ikj}) &= b_{1ik} + b_{2ik} t_{ikj} + b_{3ik} t^2_{ikj} + b_{4ik} t^3_{ikj}. \end{aligned} \tag{22}$$

In regard to the survival times, we assumed $T^*_{ik}|\boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, Q_k \sim \mathcal{W}(1, \lambda_{ik}(t)) \equiv \mathcal{E}(\lambda_{ik}(t))$, where

$$\log(\lambda_{ik}(t)) = \beta_{21} + \beta_{22}\text{gender}_{ik} + \beta_{23}\text{age}_{ik} + \beta_{24}\text{PrevOI}_{ik} + \textstyle\sum_{s=1}^4 \gamma_s b_{sik} + Q_k, \tag{23}$$

being the latent parameters, b_{sik}, $s = 1, \ldots, 4$, random effects related to the intercept, slope, curvature and rate of change of the curvature. Notice that the γ_s coefficients,

$s=1,\ldots,4$, quantify the extent to which each of the random effects influences the hazard of death. For example, $\gamma_3 = -0.5$ means that an individual with a positive curvature will have a negative association with the hazard function implying a lower risk of death.

For the parameters of the various fitted spatial joint models, we assumed vague but proper prior distributions because we had little prior information about them. In particular, we considered: $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{15}, \beta_{16}, \beta_{17})^\top \sim \mathcal{N}_7(\mathbf{0}, 1000\,\mathbf{I})$; $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23}, \beta_{24})^\top \sim \mathcal{N}_4(\mathbf{0}, 1000\,\mathbf{I})$; $\boldsymbol{b}_{ik} = (b_{1ik}, b_{2ik}, b_{3ik}, b_{4ik})^\top \sim \mathcal{N}_4(\mathbf{0}, \boldsymbol{\Sigma})$; $\boldsymbol{\Sigma} \sim \mathcal{IW}ish(1000\,\mathbf{I}, \kappa)$; $\gamma_s \sim \mathcal{N}(0, 100)$, $s=1,\ldots,4$. $\mathbf{0}$ and $\mathbf{I}$ denote the null vector and identity matrix of appropriate dimensions, respectively, and $\kappa = N/20 \approx 233$ as stated in Guo and Carlin [6] to ensure good identifiability of the main effects; for the spatial structured random effects we use the ICAR prior distribution, $Q_k|\sigma_Q^2 \sim ICAR(\sigma_Q^2)$.

In some scenarios assuming a Gamma prior for the precision (inverse of the variance) can be problematic, because of its sensitivity to prior choices of the parameters causing it to be inappropriately biased away from 0 [39]. For instance, if we want to allow for the possibility of no within-individual variability or a negligible spatial dependence between areas this prior should not be the way forward! Although in our dataset having an individual with zero variance, $\sigma^2 = 0$, is implausible, so in accordance we assume $\sigma^{-2} \sim \mathcal{G}(0.01, 0.01)$. Concerning the prior for the precision of the spatial structured frailties, σ_Q^{-2}, Kelsall and Wakefield [40] circumvented the problem suggesting an alternative prior for the precision parameter, $\sigma_Q^{-2} \sim \mathcal{G}(0.5, 0.0005)$, expressing the prior belief that the spatial random effects standard deviation is centered around 0.05 with a 1% prior probability of being smaller than 0.01 or larger than 2.5.

Initial values for the parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ were obtained by modeling the longitudinal and survival data individually. The choice of the prior distributions for σ^2, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\Sigma}$ and σ_Q^2 were motivated by their conjugacy, assuming that they are independent *a priori*. The covariates `gender`, `age` and `PrevOI` were always included into both longitudinal and survival components of the fitted spatial joint models. Estimates of the parameters were obtained through MCMC simulation within the OpenBugs [26] (*vide* SuppMat Section 5), based on sampling chains of 100,000 iterations following the 20,000 iterations of "burn-in" period. In order to eliminate autocorrelation among samples within the chains, we selected every 50th iteration of the chains. A study of the trace and density plots of the posterior distributions indicated no convergence problems concerning these samples.

5.2 Model selection

There are several summary measures for model comparison and selection. Namely, we choose the Deviance Information Criterion (DIC) (Spiegelhalter *et al.* [41]) and the so-called Watanabe-Akaike Information Criterion (WAIC) (Watanabe [42] and Gelman *et al.* [43]) which is a recent penalized likelihood-based measure. DIC and WAIC handle Bayesian models of any degree of complexity and smaller values indicate a better adjustment. The computation of these measures is straightforward using MCMC methods because it is particularly convenient to compute them from posterior samples.

Table 1 reports DIC and WAIC values for a variety of fitted joint models with different forms for the latent processes, $W(t)$, for the linking structure, $g(W(t))$, and for the spatial random effect, Q. We noted some inability of our data to reliably identify both the shape parameter, a, and the survival intercept, β_{21}, in the model (23), exhibiting strong negative correlation between the two-parameter samples and strong positive autocorrelations in their individual samples. That was already mentioned by Guo and Carlin [6]. Meanwhile, we fitted a few models from Table 1 including these two parameters, but increasing both the thin and the number of iterations per chain. After one day running in a Quad core desktop computer, the posterior mean of a was 1.04 being similar to consider the exponential survival model.

Based on the model selection measures in Table 1, considering only the first ten rows, (linear) models sharing both random effects (individual intercept and time trend) result in the best scenario. Models IX and X, which extend models VII and VIII by introducing the spatial frailty, exhibit better comparison measures values. That can suggest some latent spatial effect in the HIV/AIDS data. Models XI-XVIII assume higher degree polynomial functions in order to look for more flexible time trends, being extensions of the models VIII, IX and X. The decreasing values of model selection measures for the current models set indicate that $\sqrt{\text{CD4}}$ longitudinal profile is better captured by a non-linear trajectory, especially for Models XIV and XVIII. The latter has the lowest DIC and WAIC values among all joint models and therefore is the selected spatial joint model (6)-(7). Other joint models were fitted, namely considering only unstructured spatial random effects, where V_k's are assumed to have an exchangeable Gaussian prior; and simultaneously considering both structured and unstructured spatial heterogeneity (*vide* Subsection 3.2 and SuppMat at Section 2). Despite the values of the summary measures for these models being very close to the ones in Table 1 we note that they are always larger. Furthermore we can also note that considering both Q_k and V_k at the same time (SuppMat models XXV–XXX) has virtually no impact on DIC or WAIC (compared to models XIII–XVIII) which means that Q_k are accounting for virtually all the residual variation between the states.

For the selected joint model (Model XVIII), we present the posterior mean and the 95% credibility interval (CI) for their parameters of interest in Table 2. Additionally, in order to compare separate and joint HIV/AIDS data analysis, we include the corresponding estimates for Model IV that was the best separate model *i.e.* $g(W(t))=0$ and for Model XIV, which is the same as Model XVIII but without the spatial component, Q_k. Notice that, for Model XVIII, the (symmetric) covariance matrix, $\mathbf{\Sigma}$, is presented in terms of its components: σ^b_{11}, σ^b_{12}, σ^b_{13}, σ^b_{14}, σ^b_{22}, σ^b_{23}, σ^b_{24}, σ^b_{33}, σ^b_{34}, and σ^b_{44}. Based on the posterior estimates we conclude that: (i) `gender`, `age` and `PrevOI` have "significant" effect both in the CD4 count mean and the relative risk of death; (ii) male patients have lower CD4 counts and higher death risk during the follow-up than female ones; (iii) both patients aged above 50 and with previous opportunistic infectious disease at study entry have lower CD4 counts and higher death risk than patients in the opposite category of each group. Moreover, the posterior estimates of the parameters γ_1, γ_2, γ_3 and γ_4 provide strong evidence of a negative (latent) association between the two components. In fact, as these coefficients represent the strength of the influence that each longitudinal individual random effect, $\boldsymbol{b}_i$, has on the survival time we can say that if the individual trajectory is above the population mean trajectory, that individual will have a good survival prognostic. In simple terms, if a particular $\boldsymbol{b}_i$ with each of its elements being positive, i.e., $\boldsymbol{b}_i = (b_{i1} > 0, b_{i2} > 0, b_{i3} > 0, b_{i4} > 0)$, then his $\sqrt{\text{CD4}}$ longitudinal trajectory will be above the population mean trajectory implying that his survival time will be greater than the population survival mean because he has a lowering in the risk of death.

We also conclude that our joint model present improvements over the survival time when compared with a separate modeling. To illustrate this we considered Model IV, Model XVIII and two patients: (i) `Patient 85` - male, 31 years old, without previous opportunistic infection and censored time 1,645 days; (ii) `Patient 105` - male, 29 years old, with previous opportunistic infection and censored time 1,508 days. Figure 1 shows patient 105 with a relatively "good" CD4 trajectory (starts relatively low and then increases), while Patient 85 has a "not so good" one (starts low and do not increase much). Joint results substantially differ from the separate ones, increasing the posterior median survival time for Patient 105, and decreasing it for Patient 85. Moreover, the joint model actually reverses separate models findings, in the sense that the patient with the "good" CD4 trajectory is now predicted to survive much longer than the patient with the "bad" trajectory. For the median survival times of the patients 105 and 85, we obtained estimates of roughly 39 and 15, respectively. One referee suggested to provide information on the performance of this phenomenon by summarizing all the subjects studied, by means of the percentage of subjects with this right "reverse pattern" when considering his/her CD4 trajectory. We conduct a series of simulations, whose results are summarized in SuppMat Section 3.

Figure 2 shows two maps for the HIV/AIDS data by Brazilian states, representing the posterior spatial mean risk (left) and posterior spatial relative risk (right) based on model XVIII, respectively defined by $\bar{\lambda}_k = \sum_{i=1}^{n_k} \lambda_{ik}/n_k$ and $\exp(Q_k)$, with λ_{ik} and Q_k as in Equation (23), $k = 1, \ldots, 27$. For convenience, the posterior means of these state-specific quantities were ordered according to the quintiles of their distributions. The Brazilian states with higher HIV/AIDS spatial mean risk are located in North region (3 out of 7 states: **Acre**, **Amazonas**, **Pará**) and Northeast region (2 out of 9 states: **Paraíba**, **Sergipe**), being these states more distant from the most populous states in Brazil, especially the first set of states. It is interesting to note that when we are not considering the covariates effects and only the unobserved spatial variation (map on the right) the colors are reversed indicating an increasing spatial relative risk for the South region (3 states: **Paraná**, **Santa Catarina**, **Rio Grande do Sul**), Central-West region (2 out of 4 states: **Goiás**, **Mato Grosso do Sul**) and North region (1 out 7 states: **Acre**). Apart from the last, the first five states have moderate population density and economic growth and, even expecting to have better Public Health conditions, they still have some latent risk factors for HIV/AIDS issue. As reported by Teixeira *et al.* [44], AIDS epidemic in Brazil was only found to be expanding in the North and Northeast regions, while declining in the rest of the country, especially in the Southeast. Similar maps for models XXIV and XIV can be found on SuppMat Figure 3 and Figure 4, respectively.

Finally, note (Figure 2 - right panel) that the values of Q_k's are in the range $(\log(0.98), \log(1.04)) = (-0.02, 0.04)$ suggesting that missing regional covariates have nearly a null impact (around to 2% to 4%) on the hazard. Such a small value suggests that covariates like, for example, region economic status, quality of health care or population total per region might not be needed in explaining the spatial epidemiology.

In order to investigate the influence of the spatial specification we carried out a sensitivity analysis with respect to the hyperprior distribution for the spatial variance component, σ_Q^2, of the selected model in Subsection 5.2, assuming several different inverse gamma priors [45]. Change the distribution of the spatial variance, σ_Q^2, does not seem to affect the value of the summary measures and therefore the selected model should not change with that variation (see SuppMat – Section 1).

5.3 Residual analysis

To assess Model XVIII, we employed residual analysis as presented in Subsection 4.2. For the survival outcome the posterior estimates of the Cox-Snell residuals (16) were analyzed. To be easier to understand we plotted the Kaplan–Meier curve for the posterior mean of the Cox-Snell residuals (thick line) in Figure 3 (bottom panel), along with the unit exponential distribution (thin line), corresponding to a perfect fitting model. Although there is some deviation at the middle of the curves, the majority of the estimates are close to the "perfect" survival curve. Actually, this deviation represents only a small percentage of the total observations, about 7% of the sample size. Further examination reveals that most of these observations correspond to individuals with only one CD4 measure. Therefore, considering again that most individuals have two or more CD4 measurements, model adequacy may be deemed reasonable.

Concerning the longitudinal outcome, we combined the standardized marginal and subject-specific residuals (17) with the multiple-imputation-based residual approach (18). To generate the random visiting process, we consider the Weibull model (19), whose hazard function $h_v(u_{ikq}|\mathbf{x}_{vik}, \boldsymbol{\theta}_v)$ is given by

$$a_v u_{ikq}^{a_v - 1} \exp(\beta_{v0} + \beta_{v1}\text{age}_{ik} + \beta_{v2}\text{gender}_{ik} + \beta_{v3}\text{PrevOI}_{ik} + \beta_{v4}\, y_{ik(q-1)} + Q_k)\, \omega_{ik}. \qquad (24)$$

We assigned vague prior distributions to the regression coefficients, β_{vs}, $s = 0, \ldots, 4$, to the Weibull shape parameter, a_v, and to the individual frailty, ω_{ik}. Namely, each $\beta_{vs} \sim \mathcal{N}(0, 1000)$, $a_v \sim \mathcal{G}(0.01, 0.01)$ and $\omega_{ik} \sim \mathcal{G}(0.01, 0.01)$. As mentioned before $\sigma_Q^2 \sim \mathcal{IG}(0.5, 0.0005)$.

Aiming to get an easier reading of the residual analysis of the longitudinal component, we produce plots of the corresponding residuals using $L = 5$ imputations and check for systematic trends using weighted loess fits, with weight one for the observed residuals, and $1/L$ for the imputed ones (see Rizopoulos *et al.* [34]). The plot of the standardized subject-specific residuals in Figure 3 (left top panel) shows a slight but systematic growing for the observed residuals (dark gray line). That behavior is alleviated when we consider the imputed residuals (light gray line) and, therefore, the homoscedasticity of the errors e_{ikj} is verified. In Figure 3 (right top panel), the plot of the standardized marginal residuals point out that the fitted weighted loess curve, based on the observed data alone *versus* the fitted values of $\sqrt{\text{CD4}}$, shows a slight systematic decrease (dark gray line) but that behavior is not present when we look to the imputed residuals (light gray line), indicating that after taking dropout into account the fitted joint model seems to be a plausible model for this data set.

Finally, the little differences observed between the dark gray and the light gray lines may be due to the visits frequency of our follow-up. In a CD4 counts context Geskus R. [46] shows that if follow-up is frequent the nonrandom dropout may not be a source of bias.

5.4 Prediction of future values

We performed predictions among patients in HIV/AIDS data as stated in Subsection 4.1, for 11 individuals who died and had 6 or more CD4 repeated measures. We aimed to obtain the conditional probability of surviving for some time later relatively to the last CD4 measurement time considering that the individual was censored immediately after it, in order to verify the time-to-event predictive ability of the model. In this sense, we have removed these 11 individuals from the data before obtaining its posterior quantities. Predictions were made such that each individual presented 20 CD4 measurements. Figure 4 shows plots with the two types of predictions: (i) CD4 median trajectory obtained accordingly to (14) (dashed line) and its 95% CI (gray area); (ii) conditional survival probability (solid line) obtained accordingly to (15).

Generally subjects are predicted to live longer than what occurred in reality. This can be justified by the small percentage of death in the data, resulting in a shrinkage of these individuals towards the overall mean of the survival time. For individuals with the lowest CD4 counts, after 1 or 2 years the predictions are very inaccurate because the 95% CI for the CD4 counts is very large (*e.g.* individuals 242 and 329). We note that what seems to have the most influence on survival time prediction is the overall time trend. When there is an upward trajectory, the survival curve remains almost constant and equal to 1 (*e.g.* individuals 329, 767, 1349 and 1415). The longitudinal component of the selected model seems to capture the variations in the longitudinal trajectory because the most part of the observations lies within the 95% CI (gray area). It should be noted that the patients who died were not necessarily those ones with the worst CD4 trajectory, *i.e.*, a decreasing overall slope in that trajectory. Possibly this caused some difficulties in performing predictions.

We also run a few more simulations, namely to compare the longitudinal predictive performance of our model XVIII against the separated model IV, the simple joint model IX and two other models without and with unstructured spatial random effects, models XIV and XXIV, respectively (*vide* SuppMat Section 4). Model XVIII always outperforms its competitors in terms of coverage.

6 Concluding remarks

The introduction of functional time and spatial frailty effects in longitudinal and survival joint models adds new tools for analyzing them. The associated maps provide visual representations of the regions in study, allowing to identify areas of high spatial relative risk that should receive more attention and resource from the public health policy. From this point of view it is of a great value to know that apparently there are no spatial differences in the risk of death. It means that patients across all regions have e.g. access to different health cares and their survival depends on the region where they live.

For our HIV/AIDS data, i) $\sqrt{CD4}$ longitudinal profile is better captured by a non-linear trajectory; ii) joint analysis substantially differs from the separate ones, increasing (decreasing) the posterior median of the survival times for patients with a relatively "good" ("not so good") CD4 trajectory; iii) gender, age and PrevOI have "significant" effect both in the CD4 count mean and the relative risk of death, for instance, male patients have lower CD4 counts and higher death risk during the follow-up than female ones; iv) for the prediction of future values, we note that the overall time trend seems to have the most influence on survival time prediction and the patients who died were not necessarily those ones with the worst CD4 trajectory.

In addition, the Brazilian states with higher HIV/AIDS spatial mean risk are located in North region (Acre, Amazonas, Pará) and Northeast region (Paraíba, Sergipe), being these states more distant from the most populous states in Brazil. Although we have found small spatial unobserved heterogeneity at state level in Brazil, taking into account the spatial dependence structure improves the corresponding predictions of both the CD4 trajectory and the HIV/AIDS survival curve. (*vide* Subsection 5.4 and Sections 3 and 4 in SuppMat). In order to detect more spatial extra-variation in the Brazilian HIV/AIDS data, we should have used another area definition instead of states which, unfortunately, was not available in the database that has been provided.

Some of the posterior estimates from the non-spatial separate model IV, non-spatial joint model XIV and the selected one (XVIII) are similar (Table 2). This could hide the advantages in use the spatial model, but there are indeed important issues associated with the proposed model and its results. For instance, i) some apparent overall stabilization of the AIDS epidemic in Brazil tends to mask regional disparities and the susceptibility of given specific locations and should, thus, be evaluated carefully through analyses with lower levels of aggregation such as municipalities and micro-regions instead of states [44]; ii) motivated by the absence of past AIDS studies or expert conjectures in Brazil, we have used non-informative prior for spatial variance components. That assumption is well-accept with a prior sensitivity analysis but not consensual, e.g. Gelman [39] discussed prior distributions for variance parameters in hierarchical models illustrating some problems with the inverse-gamma family of non-informative prior distributions.

With more biomedical studies taking measures of various outcomes over time in an effort to evaluate a patient's health or risk to some event, a joint modeling approach is indeed useful to link these longitudinal and survival outcomes. Despite the reasonable ease of implementing Bayesian joint models, they have some potential limitations, for example, slow convergence of MCMC methods due to the large number of parameters that need to be estimated. Alternative methods are the integrated nested Laplace approximation methods (INLA), proposed by Rue *et al.* [47], which is a recent approach to statistical inference based on latent Gaussian Markov random field models.

There is an undeniable appeal in applying joint models, but there is still a long way to address issues such as identification of the appropriate association structure [48, 49]. Joint models are on the front line of the statistical methods applied to a personalized medicine mainly because of its ability in deriving individualized predictions both in the longitudinal and survival responses, such the ones in Subsection 5.4.

AIDS is regarded today as a chronic disease. Indeed, because of the small percentage of death in our database we could investigate a possible cure fraction in HIV/AIDS Brazilian population, especially in Southwest region. It would be interesting to include other longitudinal measures along with CD4 counts (*e.g.* viral load), thus generating a multivariate longitudinal component of the spatial joint model.

Acknowledgements

The authors wish to thank Maria Goretti Fonseca and Cláudia Coeli by the database. This work was partially funded by Fundação para a Ciência e a Tecnologia (FCT) projects PTDC/MAT/118335/2010 and PEst-OE/MAT/UI0006/2014. We also thank two anonymous referees and an associate director for their helpful comments and suggestions that have substantially improved the presentation of this article.

References

[1] Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica* 2004; **14**(3):809–834.

[2] Tsiatis AA, DeGruttola V, Wulfsohn MS. Modelling the relationship of survival to longitudinal data measured with error: applications to survival CD4 counts in patients with AIDS. *Journal of the American Statistical Association* 1995; **90**(429):27–37.

[3] Faucett CL, Thomas DC. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine* 1996; **15**(15):1663–1685, doi:10.1002/(SICI)1097-0258(19960815)15:15<1663::AID-SIM294>3.0.CO;2-1.

[4] Wang Y, Taylor JMG. Jointly modelling longitudinal and event time data, with applications to AIDS studies. *Journal of the American Statistical Association* 2001; **96**(455):895–905.

[5] Ibrahim JG, Chen MH, Lipsitz SR. Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika* 2001; **88**(2):551–564, doi:10.1093/biomet/88.2.551.

[6] Guo X, Carlin BP. Separate and joint modelling of longitudinal and event time data using standard computer packages. *The American Statistician* 2004; **58**(1):16–24.

[7] Proust-Lima C, Taylor JMG. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. *Biostatistics* 2009; **10**(3):535–549, doi:10.1093/biostatistics/kxp009.

[8] Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* 1982; **69**(2):331–342.

[9] Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostatistics* 2000; **1**(4):465–480, doi:10.1093/biostatistics/1.4.465.

[10] Chi YY, Ibrahim JG. Joint models for multivariate longitudinal and multivariate survival data. *Biometrics* 2006; **62**(2):432–445, doi:10.1111/j.1541-0420.2005.00448.x.

[11] Ratcliffe SJ, Guo WS, Have TRT. Joint modeling of longitudinal and survival data via a common frailty. *Biometrics* 2004; **60**(4):892–899, doi:10.1111/j.0006-341X.2004.00244.x.

[12] Ibrahim JG, Chen MH, Sinha D. *Bayesian Survival Analysis*. Springer-Verlag: New York, 2001.

[13] Rizopoulos D. *Joint Models for Longitudinal and Time-to-Event Data with Applications in R*. Chapman and Hall/CRC: Boca Raton, Florida, 2012.

[14] Rizopoulos D. JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software* 2010; **35**(9):1–33, doi:10.18637/jss.v035.i09.

[15] Rizopoulos D. *JMbayes: Joint modeling of longitudinal and time-to-event data under a Bayesian Approach* 2014. R package version 0.6-1.

[16] Gould AL, Boye ME, Crowther KJ, Ibrahim JG, Quartey G, Sandrine Micallef S, Bois FY. Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Statist. Med.* 2015; **34**(14):2202–2203, doi:10.1002/sim.6141.

[17] Banerjee S, Carlin BP, Gelfand AE. *Hierarchical Modeling and Analysis for Spatial Data*. 2nd edn., Chapman & Hall/CRC: Boca Raton, Florida, 2014.

[18] Fonseca MG, Coelli CM, Lucena F, Veloso V, Carvalho M. Accuracy of a probabilistic record linkage strategy applied to identify deaths among cases reported to the Brazilian AIDS surveillance database. *Cadernos de Saúde Pública* 2010; **26**(7):1431–1438, doi:10.1590/S0102-311X2010000700022.

[19] Deapen D, Cockburn M, Pinder R, Lu S, Wohl A. Population-based linkage of AIDS and cancer registries. *American Journal of Preventive Medicine* Aug 2007; **33**(2):134–136, doi:10.1016/j.amepre.2007.03.015.

[20] Souza-Jr PRB, Szwarcwald CL, Castilho EA. Delay in introducing antiretroviral therapy in patients infected by HIV in Brazil, 2003-2006. *Clinical Science* 2007; **62**(5):579–584, doi:10.1590/S1807-59322007000500008.

[21] Taylor JMG, Law N. Does the covariance structure matter in longitudinal modelling for the prediction of future CD4 counts? *Statistics in Medicine* 1998; **17**(20):2381–2394.

[22] Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. 2nd edn., John Wiley: New York, 2002.

[23] Rizopoulos D, Verbeke G, Molenberghs G. Shared parameter models under random effects misspecification. *Biometrika* 2008; **95**(1):63–74, doi:10.1093/biomet/asm087.

[24] Besag J, York J, Mollié A. Bayesian image restoration with two application in spatial statistics. *Annals of the Institute of Statistical Mathematics* 1991; **43**(1):1–59.

[25] Eberly LE, Carlin BP. Identifiability and convergence issues for markov chain monte carlo fitting of spatial models. *Stat. Med.* Sep 2000; **19**(17-18):2279–94, doi:10.1002/1097-0258(20000915/30)19:17/18<2279::AID-SIM569>3.0.CO;2-R.

[26] Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine* 2009; **28**(25):3049–3067, doi:10.1002/sim.3680.

[27] Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis*. 3rd edn., CRC Press: London, 2013.

[28] Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* 2011; **67**(3):819–829, doi:10.1111/j.1541-0420.2010.01546.x.

[29] Sweeting MJ, Thompson SG. Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal* 2011; **53**(5):750–763, doi:10.1002/bimj.201100052.

[30] Dobson A, Henderson R. Diagnostics for joint longitudinal and dropout time modeling. *Biometrics* 2003; **59**(4):741–751, doi:10.1111/j.0006-341X.2003.00087.x.

[31] Zhu H, Ibrahim JG, Chi YY, Tang N. Bayesian influence measures for joint models for longitudinal and survival data. *Biometrics* 2012; **68**(3):954–964, doi:10.1111/j.1541-0420.2012.01745.x.

[32] Zhang D, Chen MH, Ibrahim JG, Boye ME, Wang P, Shen W. Assessing model fit in joint models of longitudinal and survival data with applications to cancer clinical trials. *Statistics in Medicine* Nov 30 2014; **33**(27):4715–4733, doi:10.1002/sim.6269.

[33] Park KY, Qiu P. Model selection and diagnostics for joint modeling of survival and longitudinal data with crossing hazard rate functions. *Statistics in Medicine* 2014; **33**(26):4532–4546, doi:10.1002/sim.6259.

[34] Rizopoulos D, Verbeke G, Molenberghs G. Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics* 2010; **66**(1):20–29, doi:10.1111/j.1541-0420.2009.01273.x.

[35] Cox DR, Snell EJ. A general definition of residuals. *Journal of the Royal Statistical Society (Series B)* 1968; **30**(2):248–254.

[36] Rizopoulos D, Ghosh P. A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine* 2011; **30**(12):1366–1380, doi:10.1002/sim.4205.

[37] Nobre JS, Singer JM. Residual analysis for linear mixed models. *Biometrical Journal* 2007; **49**(6):863–875, doi:10.1002/bimj.200610341.

[38] Wu H, Zhang JT. Local polynomial mixed-effects models for longitudinal data. *Journal of the American Statistical Association* 2002; **97**(459):883–897, doi:10.1198/016214502388618672.

[39] Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 2006; **1**(3):1–19, doi:10.1214/06-BA117A.

[40] Kelsall J, Wakefield J. Discussion of "Bayesian models for spatially correlated disease and exposure data" by Best, N.G. and Waller, L.A. and Thomas, A. and Conlon, E.M. and Arnold, R. *Sixth Valencia International Meeting on Bayesian Statistics*, Bernardo J, Berger J, Dawid A, Smith A (eds.), Oxford University Press: London, 1999.

[41] Spiegelhalter DJ, Best NG, Carlin BP, Van der Linde A. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society (Series B)* 2002; **64**(4):583–639, doi:10.1111/1467-9868.00353.

[42] Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* 2010; **11**(455):3571–3591.

[43] Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. *Statistics and Computing* 2014; **24**(6):997–1016, doi:10.1007/s11222-013-9416-2.

[44] Teixeira TRA, Gracie R, Malta MS, Bastos FI. Social geography of AIDS in Brazil: identifying patterns of regional inequalities. *Cadernos de Saúde Pública* 2014; **30**(2):259–271, doi:10.1590/0102-311X00051313.

[45] Silva GL, Dean CB, Niyonsenga T, Vanasse A. Hierarchical Bayesian spatiotemporal analysis of revascularization odds using smoothing splines. *Statistics in Medicine* 2008; **27**(13):2381–2401, doi:10.1002/sim.3094.

[46] Geskus R. Which individuals make dropout informative? *Statistical Methods in Medical Research* 2014; **23**(2):91–106, doi:10.1177/0962280212445840.

[47] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society (Series B)* 2009; **71**(2):319–392, doi:10.1111/j.1467-9868.2008.00700.x.

[48] Gould AL, Boye M, Crowther M, Ibrahim J, Quartey G, Micallef S, Bois FY. Responses to discussants of "Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group". *Statistics in Medicine* Jun 2015; **34**(14):2202–2203, doi:10.1002/sim.6502.

[49] Rizopoulos D. Comments on "Joint modeling of survival and longitudinal non-survival data: current methods and issues. report of the DIA Bayesian joint modeling working group". *Statistics in Medicine* Jun 2015; **34**(14):2196–2197, doi:10.1002/sim.6260.

Table 1: Candidate Bayesian joint models for the HIV/AIDS data analysis.

Model	$W(t)$	$g(W(t))$	Q	DIC	WAIC
Part I - None or one degree polynomial function:					
no random effects					
I	0	0	0	139004	215495
random intercept					
II	b_1	0	0	119686	196486
III	b_1	$\gamma_1 b_1$	0	119359	195970
random intercept and random slope					
IV	$b_1 + b_2 t$	0	0	112251	189537
V	$b_1 + b_2 t$	$\gamma_1 b_1$	0	112026	189069
VI	$b_1 + b_2 t$	$\gamma_2 b_2$	0	112228	189433
VII	$b_1 + b_2 t$	$\gamma(b_1 + b_2)$	0	111948	188984
VIII	$b_1 + b_2 t$	$\gamma_1 b_1 + \gamma_2 b_2$	0	111907	188863
spatial random effects					
IX	$b_1 + b_2 t$	$\gamma_1 b_1 + \gamma_2 b_2$	Q	111891	188883
X	$b_1 + b_2 t$	$\gamma(b_1 + b_2)$	Q	111952	188960
Part II - Two or more degree polynomial function:					
no spatial random effects					
XI	$b_1 + b_2 t + b_3 t^2$	$\gamma_1 b_1 + \gamma_2 b_2 + \gamma_3 b_3$	0	108992	185051
XII	$b_1 + b_2 t + b_3 t^2 + b_4 t^3$	$\gamma_1 b_1 + \gamma_2 b_2$	0	108483	184680
XIII	$b_1 + b_2 t + b_3 t^2 + b_4 t^3$	$\gamma_1 b_1 + \gamma_2 b_2 + \gamma_3 b_3$	0	108372	184492
XIV	$b_1 + b_2 t + b_3 t^2 + b_4 t^3$	$\gamma_1 b_1 + \gamma_2 b_2 + \gamma_3 b_3 + \gamma_4 b_4$	0	108100	183649
spatial random effects					
XV	$b_1 + b_2 t + b_3 t^2$	$\gamma_1 b_1 + \gamma_2 b_2 + \gamma_3 b_3$	Q	108983	185005
XVI	$b_1 + b_2 t + b_3 t^2 + b_4 t^3$	$\gamma_1 b_1 + \gamma_2 b_2$	Q	108475	184703
XVII	$b_1 + b_2 t + b_3 t^2 + b_4 t^3$	$\gamma_1 b_1 + \gamma_2 b_2 + \gamma_3 b_3$	Q	108388	184555
XVIII	$b_1 + b_2 t + b_3 t^2 + b_4 t^3$	$\gamma_1 b_1 + \gamma_2 b_2 + \gamma_3 b_3 + \gamma_4 b_4$	Q	108083	183568

Table 2: Posterior parameters estimates for separate (IV) and joint (XIV and XVIII) models.

	Model IV		**Model XIV**		**Model XVIII**	
Parameter	**Mean**	**95% CI**	**Mean**	**95% CI**	**Mean**	**95% CI**
Longitudinal:						
Intercept (β_{11})	17.39	$(17.14, 17.66)$	16.93	$(16.66, 17.2)$	16.94	$(16.66, 17.21)$
Time (β_{12})	1.81	$(1.71, 1.90)$	4.26	$(4.01, 4.52)$	4.27	$(4.01, 4.53)$
Time^2(β_{13})	–	–	-1.68	$(-1.85, -1.51)$	-1.68	$(-1.87, -1.51)$
Time^3(β_{14})	–	–	0.24	$(0.20, 0.28)$	0.24	$(0.20, 0.28)$
Gender (β_{15})	-0.63	$(-0.93, -0.32)$	-0.64	$(-0.93, -0.35)$	-0.65	$(-0.93, -0.36)$
Age (β_{16})	-0.51	$(-0.96, -0.05)$	-0.59	$(-1.0, -0.14)$	-0.59	$(-1.05, -0.14)$
PrevOI (β_{17})	-2.01	$(-2.33, -1.71)$	-2.01	$(-2.32, -1.70)$	-2.01	$(-2.32 - 1.72)$
σ^2	7.04	$(6.87, 7.20)$	5.48	$(5.33, 5.64)$	5.47	$(5.32, 5.63)$
σ_{11}^b	26.92	$(25.64, 28.21)$	30.45	$(29.02, 31.93)$	30.44	$(29.01, 31.93)$
σ_{12}^b	-4.72	$(-5.28, -4.14)$	-13.56	$(-15.34, -11.9)$	-13.55	$(-15.29, -11.82)$
σ_{13}^b	–	–	3.89	$(2.88, 4.98)$	3.85	$(2.78, 4.94)$
σ_{14}^b	–	–	-0.25	$(-0.47, -0.03)$	-0.23	$(-0.46, 0.01)$
σ_{22}^b	5.20	$(4.82, 5.59)$	25.10	$(21.9, 28.73)$	25.42	$(22.13, 28.32)$
σ_{23}^b	–	–	-8.20	$(-10.19, -6.48)$	-8.38	$(-9.94, -6.61)$
σ_{24}^b	–	–	0.71	$(0.41, 1.07)$	0.74	$(0.43, 1.02)$
σ_{33}^b	–	–	3.19	$(2.38, 4.23)$	3.29	$(2.42, 4.10)$
σ_{34}^b	–	–	-0.34	$(-0.50, -0.21)$	-0.35	$(-0.48, -0.22)$
σ_{44}^b	–	–	0.15	$(0.12, 0.16)$	0.14	$(0.12, 0.16)$
Survival:						
Intercept(β_{21})	-4.30	$(-4.54, -4.07)$	-5.91	$(-6.37, -5.47)$	-5.90	$(-6.40, -5.44)$
Gender (β_{22})	0.33	$(0.10, 0.59)$	0.48	$(0.17, 0.79)$	0.47	$(0.18, 0.78)$
Age (β_{23})	0.62	$(0.33, 0.88)$	0.88	$(0.53, 1.25)$	0.87	$(0.51, 1.23)$
PrevOI (β_{24})	0.87	$(0.63, 1.10)$	1.09	$(0.82, 1.38)$	1.09	$(0.81, 1.39)$
γ_1	–	–	-0.21	$(-0.25, -0.17)$	-0.21	$(-0.24, -0.17)$
γ_2	–	–	-0.46	$(-0.61, -0.32)$	-0.45	$(-0.58, -0.30)$
γ_3	–	–	-1.57	$(-2.05, -1.13)$	-1.55	$(-1.95, -1.08)$
γ_4	–	–	-5.95	$(-6.84, -5.08)$	-5.94	$(-6.88, -5.02)$
σ_Q^2	–	–	–	–	0.014	$(0.001, 0.120)$

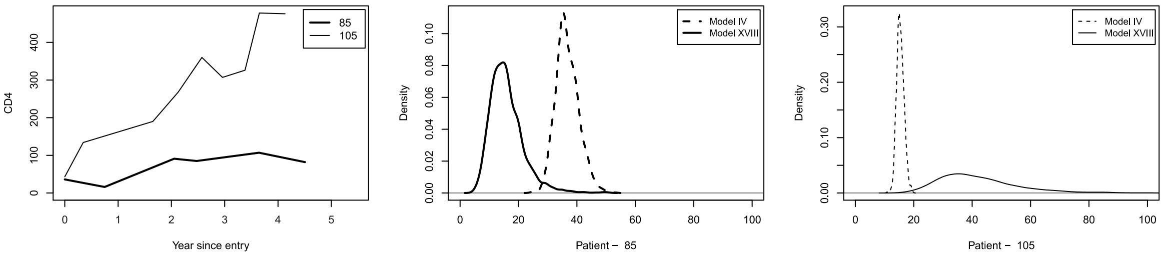

Figure 1: CD4 trajectory for patients 85 and 105 (left) and the posterior distributions of the median survival time for the patients 85 (middle) and 105 (right) using model XVIII (solid line) and model IV (dashed line).

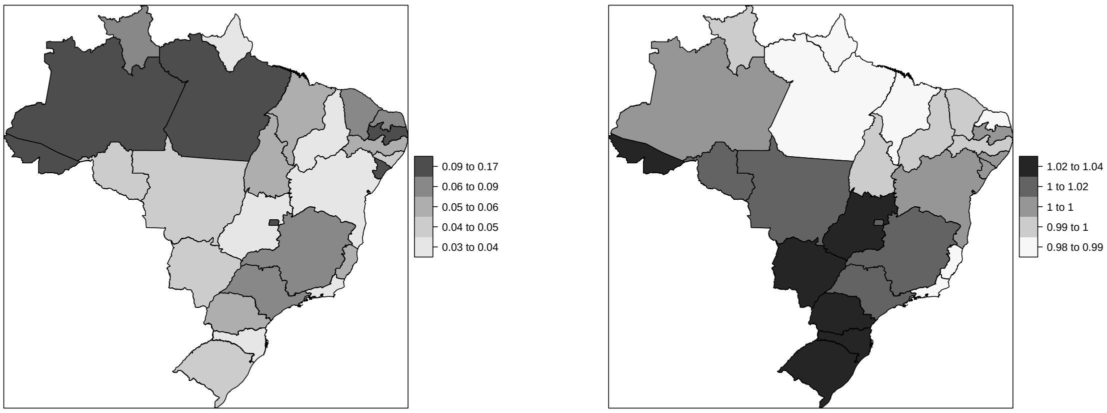

Figure 2: Maps of the spatial mean (left) and relative (right) risks based on model XVIII.

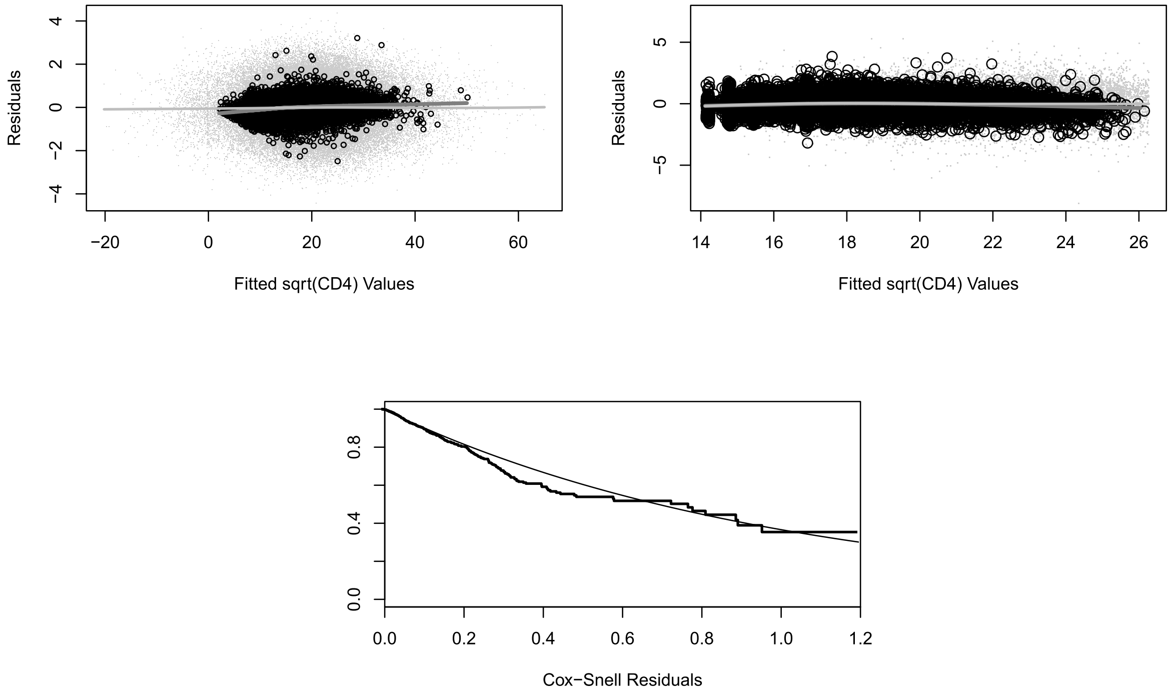

Figure 3: Standardized subject-specific (top left) and standardized marginal (top right) residuals (black circles), augmented with all the multiply imputed residuals produced by the $L = 5$ imputations (gray points). The superimposed dark gray and light gray lines represent a loess fit based only on the observed residuals and a weighted loess fit based on all residuals, respectively. The empirical survival curves (bottom panel) based on the Kaplan-Meier posterior estimates of the Cox-Snell residuals (thick line) and the unit exponential distribution (thin line).

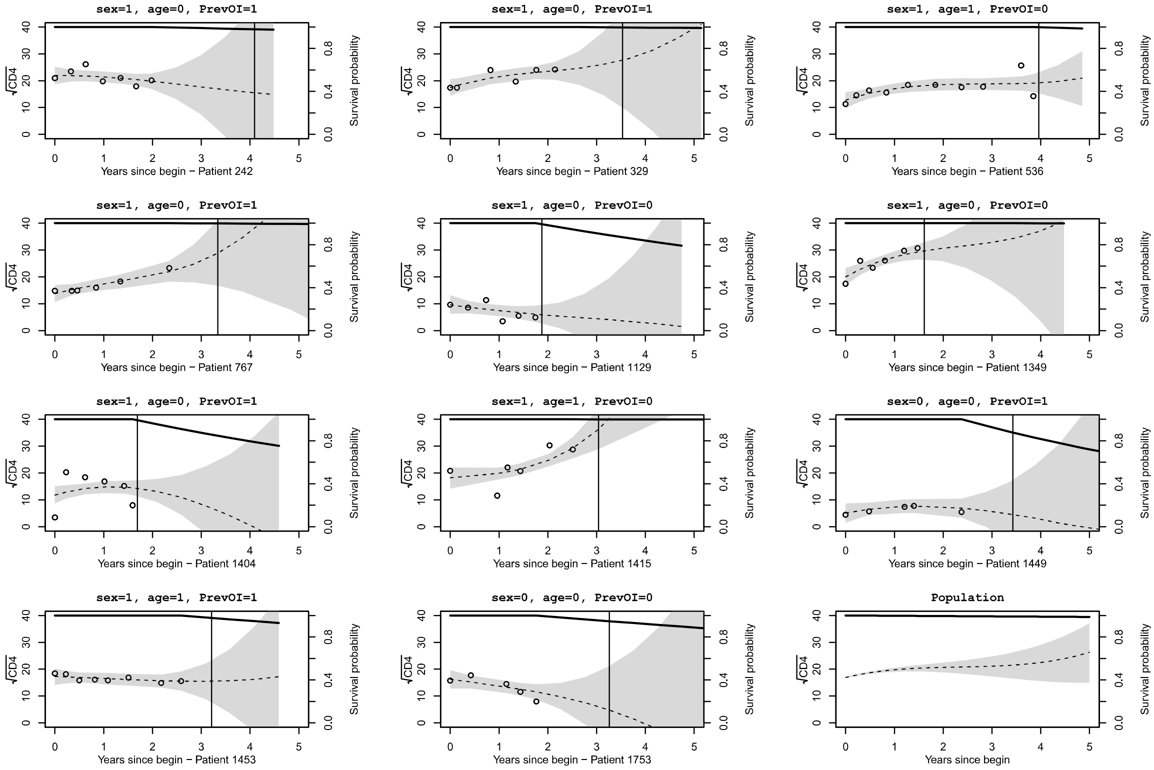

Figure 4: Predictions of the $\sqrt{\text{CD4}}$ trajectory and the survival curve for 11 patients (first eleven panels) and all patients (bottom right panel) based on model XVIII. Median $\sqrt{\text{CD4}}$ trajectory (dashed line) and predicted survival curve (solid line) after last CD4 measurement. Gray area delimits the 95% CI, whereas vertical line is the observed survival time for each patient.