Bayesian Joint Analysis of Longitudinal and Survival AIDS Data with a Spatial Fraction of Long-term Survivors

Rui Martins · **Giovani Silva** · **Valeska Andreozzi**

Abstract When we have a time-dependent covariate in the survival model that is measure with error (e.g. $CD4^+$T lymphocyte counts) we can modeling it longitudinally. A joint analysis of longitudinal and survival data allows an improvement in the results over a simple survival analysis. In the last decade we have witnessed a large wave of research in this field. Additionally, with rapid developments in medical and health sciences, researchers increasingly encounter data sets where a substantial portion of patients who never experience the event of interest of the survival analysis (*i.e.* they are long-term survivors or cured). Models accounting for long-term survivors in the population help in the prognosis of chronicle or potentially terminal diseases. Along with this medical researches have been interested in incorporate geographical information inside its analysis. These issues have led statisticians to develop models that account for spatial variation. This article proposes a Bayesian approach to model survival data with a spatial clustering in the presence of long-term survivors using a general class of cure models proposed in Cooner et al (2007), jointly with a longitudinal modeling of $CD4^+$T lymphocyte counts for a random sample of 4653 individuals collected in all the 27 states of Brazil during the years 2002-2006. As long as we

This work was partially funded by projects PTDC/MAT/118335/2010 and Pest-OE/MAT/UI0006/2011.

Rui Martins
Escola Superior de Saúde Egas Moniz, Centro de Investigação Interdisciplinar Egas Moniz, Quinta da Granja, Monte de Caparica, 2829-511 Caparica - Portugal
Tel.: +351 212946700
Fax: +351 212946868
E-mail: ruimartins@egasmoniz.edu.pt

Giovani Silva
Instituto Superior Técnico, Universidade de Lisboa, Centro de Estatística e Aplicações da Universidade de Lisboa, Bloco C6 - Piso 4, Campo Grande, 1749-016 Lisboa - Portugal

Valeska Andreozzi
Universidade de Lisboa, Centro de Estatística e Aplicações da Universidade de Lisboa, Bloco C6 - Piso 4, Campo Grande, 1749-016 Lisboa - Portugal

know this is the first paper dealing with AIDS data assuming that a fraction of the population are long-term survivors.

Keywords Joint Modeling · Longitudinal Data · Survival Data · Bayesian · Spatial · Cure fraction

1 Introduction

With the advent of the HAART therapy (Highly Active Antiretroviral Therapy) individuals living with HIV/AIDS are expected to live ever longer. As a result of this improvement recent researches show that these individuals have a life expectancy which is approximately the same as the non-infected individuals (Van Sighem et al, 2010)). We can assume in the limit that some of the infected patients will die of old age or because other reasons linked with the biological process of aging (e.g. cancer, heart attack, etc.). These patients are called in the literature of survival models with cure fraction *long-term survivors* or *cured* (it means that exists a non-zero tail probability in the survival function). Inside this framework we think that the latter designation is not a good choice because there is no cure for this disease. Instead *long-term survivors* is more meaningful in the sense that AIDS can be considered a chronicle disease where patients will live with it for the rest of their lifes.

Models that account for long-term survivors are important for understanding prognosis in chronicle diseases. Traditional survival models do not account for this fraction of patients, assuming instead that individuals who do not experience the event are censored. In these contexts one should distinguishes between the concepts of censoring and long-term survivors: censoring refers to a subject who does not fail within the time window of the experiment; long-term survivor refers to one who will never experience the event of interest of the survival analysis even if followed-up indefinitely.

As long as we know there are not papers published dealing with AIDS data assuming that a fraction of the population are long-term survivors. We think that a survival model which not assumes a proportion of subjects that will never experience the event of interest (death as a result of HIV/AIDS infection) is insufficient. Work with a survival model that can accommodate long-term survivors (because of the chronic characteristics of the sickness) and repeated measures of a longitudinal variable measured with error (CD4) seems a natural approach to this problem. This path may shed a new light on the biological behaviour of the HIV/AIDS disease.

Figure 1 shows a Kaplan-Meier estimate of the survival function from the AIDS data (described in section 2) with 95% confidence intervals. The estimate appears to decrease very slowly, indicating a possible presence of long-term survivors who may not coming to die because of HIV/AIDS.

There has been a great deal of work done on jointly modelling longitudinal and survival data with no survival fraction (DeGruttola and Tu (1994), Tsiatis et al (1995), Faucett and Thomas (1996), Wulfsohn and Tsiatis (1997), Henderson et al (2000), Wang and Taylor (2001), Xu and Zeger (2001), Tseng et al (2005), Elashoff et al (2008), Ibrahim et al (2004), Guo and Carlin (2004), Chi and Ibrahim (2006), Zhang

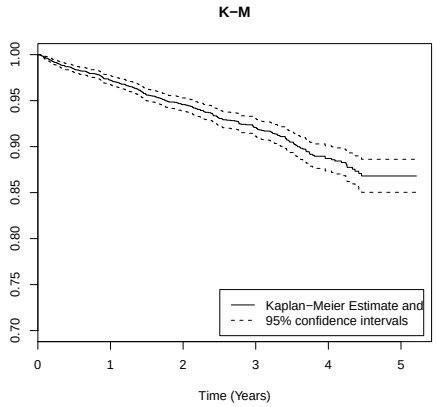

Fig. 1 Kaplan-Meier estimate of the survival curve with 95% confidence intervals.

et al (2010), Rizopoulos et al (2010); Rizopoulos and Ghosh (2011)). Currently existing methods that can accommodate a proportion of the population being cured, known as cure-rate models, include Berkson and Gage (1952), Farewell (1982), Yakovlev et al (1993), Chen et al (1999, 2004), Brown and Ibrahim (2003), Banerjee et al (2004), Cooner et al (2006), Cooner et al (2007), Gu et al (2011), among others. In these models, the cured proportion are sometimes referred to as the "survival fraction".

The remainder of this paper evolves as follows. In section 2 we describe the data set. In section 3 we outline the spatial joint model with long-term survivors that we work with. Section 4 derives the likelihood function. Predictions of future event times or longitudinal measures are discussed in section 5. In Section 6.1 a detailed analysis of the HIV/AIDS data set is conducted to apply the proposed method.

2 Data Structure

The Brazilian National AIDS Program has been acknowledged as a success in controlling the epidemic. It has generated three major electronic databases Fonseca et al (2010): SINAN-AIDS (Information System for Notifiable Diseases of AIDS Cases), SISCEL (Laboratory Test Control System) and SICLOM (System for Logistic Control of Drugs). We consider a random sample of the three databases referred above which have been combined in a single database with both HIV and AIDS cases applying the linkage strategy adopted by the Surveillance Unit of the Brazilian National AIDS Program, including all individuals in each database Fonseca et al (2010).

The longitudinal and survival data were collected in a network of 88 laboratories located in every 27 states of Brazil during the years 2002-2006. $CD4^{+}$T lymphocyte counts (a measure of immunologic and virologic status) and survival time were the responses collected in a random sample of $N = 4653$ individuals from the original data base. While the explanatory variables included were: age ("15–49" codded 0 and ≥ 50 codded $= 1$), gender (Female$= 0$, Male$= 1$), PrevOI (previous opportunis-

tic infection at study entry= 1, no previous infection= 0), state (patient's Brazilian state of residence), date of HIV/AIDS diagnosis and date of death (available if death happened before 31 December 2006 and censored otherwise). The survival time after diagnosis is calculated as the time period, in years, between date of diagnosis and date of death. As referred in Souza-Jr et al (2007) the variable that accounts for age was chosen basis on the Ministry of Health recommendations, as the over-50 age group showed a higher proportion of delayed initiation of therapy when compared to the population group aged 15-49 years.

Because CD4 counts showed a high degree of skewness toward high CD4 counts, we worked with the square root of these measures. Table 1 summarizes the total number of HIV/AIDS individuals per Brazilian state and the percentage of deaths. There were 320 deaths; 88% of the patients were between 15 and 49 years old; 2774 (60%) were males of whom 220 died. 61% percent of the individuals had no previous infection; 6.7% lived in the Central-West, 11.5% in the Northeast, 4.8% in the North, 60% in the Southeast region and 16.7% in the South. The median initial $CD4^+$ T lymphocyte count was 245 cells/mm^3 (men - 226 cells/mm^3; women - 263 cells/mm^3) and patients made on average 4.62 CD4 exams resulting in a total of 21508 observations.

Table 1 Number of HIV/AIDS individuals per Brazilian state and respective percentage of deaths.

State	Total	Dead (%)	State	Total	Dead (%)
Distrito Federal	62	5	Acre	125	14
Goiás	59	2	Amapá	12	0
Mato Grosso	272	8	Amazonas	12	8
Mato Grosso Sul	244	7	Pará	4	50
Alagoas	322	8	Rondônia	162	7
Bahia	79	3	Roraima	7	14
Maranhão	65	8	Tocantins	23	9
Pernambuco	540	6	Espírito Santo	16	6
Rio G. Norte	262	6	Minas Gerais	28	14
Sergipe	23	9	Rio de Janeiro	109	5
Ceará	22	14	São Paulo	1835	7
Piauí	34	3	Paraná	34	6
Paraíba	52	4	Rio G. Sul	98	10
			Santa Catarina	152	3

3 Joint model with a spatial fraction of long-term survivors

Joint models are a class of models to describe the joint behavior of two components. A longitudinal process, frequently a set of repeated measures of a biomarker (biological marker that can be used to asses the prognostic of the disease) and an associated survival process. The former is observed at a series of times and the latter gives rise to censored and event times. We usually start building separate models for each component and then link them together. One way to do this is building some characteristics of the longitudinal model into the survival model. A joint model allows simultane-

ously the estimation of the parameters that describe the biomarker process and those that describe the risk of failure as a function of the biomarker process.

For a clear focus and ease of exposition we will develop our model in the context of HIV/AIDS data where we typically deal with a survival process (time-to-event, i.e. death) where the biomarker ($CD4^+$T lymphocyte counts) is assumed to be a time-dependent covariate measured with error that should be longitudinally modeled.

Let us suppose that we have a sample of N individuals coming from K regions. The number of patients in the k-th region is n_k, where $k = 1,\ldots,K$. Given that the i-th subject of the k-th region $i = 1,\ldots,n_k$ is observed at time t, it means that he is at risk at time t. Let T_{ik} denote the event time for the ik-th individual which may be right censored. The event indicator is defined as δ_{ik} ($\delta_{ik} = 1$ indicates a failure and $\delta_{ik} = 0$ indicates a right censored observation).

Let $\mathcal{Y}_{ik}(t) = \{y^*_{ik}(u), 0 \leq u < t\}$ denote the history of the *true* and *unobserved* longitudinal process up to time point t. Although this formulation involves the longitudinal response at any time t, the response is collected on each subject only intermittently at some set of times $\{t_{ikj} \leq T_{ik}, j = 1,\ldots,n_{ik}, k = 1,\ldots,K\}$. Note that we do not assume a common set of measurements times for all subjects.

The observed data for the ik-th subject is $\mathcal{D}_{ik} = \{\mathbf{y}_{ik}, T_{ik}, \delta_{ik}\}$. The $\mathcal{D}_{ik}$'s are taken to be independent across ik, reflecting the belief that the disease process evolves independently for each subject. We will assume that time-to-event, T_{ik}, and the vector of the observed repeated measurements, $\mathbf{y}_{ik} = (y_{ik1},\ldots,y_{ikn_{ik}})$, are conditionally independent given some covariates of interest and a set of unobserved subject-level random effects linking the longitudinal and time-to-event process (shared parameter models – Wu and Carroll (1988), Henderson et al (2000), Rizopoulos et al (2008)).

3.1 Longitudinal Model

Let $y_{ik}(t)$ be an assessment of the number of CD4 cells for the ik-th subject at time t and $y^*_{ik}(t)$ be the corresponding trajectory function representing its true value. The longitudinal model for $y_{ik}(t)$ is given by

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

where $e_{ik}(t) \sim \mathcal{N}(0,\sigma^2)$ represents a measurement error that accounts for the unexplained variation in the data. To model the trajectory function we will assume a mixed effects model. This allow the average progression to be described as a function of population parameters and subject-specific deviations from this average evolution that are accounted for by using a vector of subject-level random effects, $\boldsymbol{b}_{ik}$. Letting $\mathbf{x}^\top_{1ik}(t)$ be the fixed effects design vector, $\boldsymbol{\beta}_1$ the corresponding fixed effects parameters for the ik-th individual at time t and $\mathbf{z}^\top_{1ik}(t)$ be the random effects design vector. The trajectory function is modelled as

$$y^*_{ik}(t) = \mathbf{x}^\top_{1ik}(t)\boldsymbol{\beta}_1 + \mathbf{z}^\top_{1ik}(t)\boldsymbol{b}_{ik}. \tag{2}$$

We assume an exchangeable normal prior distribution for the random effects, i.e. $\boldsymbol{b}_{ik}|\boldsymbol{\Sigma} \overset{iid}{\sim} \mathcal{N}_p(\mathbf{0},\boldsymbol{\Sigma})$ which are independent from $e_{ik}(t)$. Due to the significant separation in time between observations, correlation among residuals over time is assumed

to be negligible, so the error term, $e_{ik}(t)$, is a sequence of mutually independent measurement errors. The structure of $\boldsymbol{\Sigma}$ describes the association between repeated observations of the longitudinal data $\mathbf{y}_{ik}$.

3.2 Survival model with a spatial fraction of long-term survivors

Recent investigations (Van Sighem et al (2010), Samji et al (2013)), clearly show the effectiveness of the HAART therapy and reveal many patients that are expected to died because of other reasons rather than opportunistic infections that are a characteristic of HIV/AIDS patients. Such patients may be viewed as long-term survivors. Incorporating such fractions of individuals in survival models leads to *survival models with long-term survivors* (often referred to as cure models). These models have an extensive history in biostatistics, with the most popular being perhaps that of Berkson and Gage (1952) which has been extensively studied in the statistical literature by several authors, including Farewell (1982, 1986), Gray and Tsiatis (1989), Ewell and Ibrahim (1997) and Stangl and Greenhouse (1998). In this model, the *survival function for the entire population* is given by

$$P(T > t) = S_p(t) = \xi + (1 - \xi)S(t), \tag{3}$$

where $\lim_{t\to+\infty} S_p(t) = \xi$ is the called *cured fraction* (or non-susceptible fraction) and $S(t)$ with $\lim_{t\to+\infty} S(t) = 0$ is the proper survivor function for the non-cured group. Because $\lim_{t\to+\infty} S_p(t) \neq 0$, S_p is not a proper survival function. Following earlier work by Yakovlev and Tsodikov (1996), Chen et al (1999) studied this model in a more general setting by assuming a Bounded Cumulative Hazard (BCH), $\lim_{t\to+\infty} H(t) = \xi$.

The BCH model has been developed inside the cancer context and is derived assuming that some latent biological process is generating the observed data, i.e. the observed failure time T (if the individuals fails) is generated by M *latent event times* (times of activation), T^*_m, $m = 1,\ldots,M$. If $M = 0$ the individual is not exposed to any of the latent factors and is considered a *long-term survivor* (not at risk of failure) and $T = \infty$. Given a fixed $M \geq 1$, the $\{T^*_m\}_{m=1}^M$ are assumed to be independently distributed with a common *latent survival function*, $P(T^* > t) = S(t) = 1 - F(t)$, that does not depend upon M. Cooner et al (2006, 2007) generalized this framework to a flexible class of cure models under several latent activation schemes. They assume that r out of M latent factors need to be activated for the subject to fail, so $T = T^*_{(r)}$, $r = 1,\ldots,M$ where $T^*_{(1)} < \cdots < T^*_{(M)}$ are the ordered $T^*_{(m)}$'s. Making $r = 1$ implies that the activation of any one of the latent factors leads to the observed failure, i.e. $T = \min_{1\leq m\leq M} T^*_{(m)}$. This is called the *first-activation* scheme. In contrast, setting $r = M$ implies $T = \max_{1\leq m\leq M} T^*_{(m)}$ and delivers a different scheme where an individual is able to resist up to $M - 1$ activations and fails with the last activation. They call this the *last-activation* scheme. More generally, an exposed subject $(M \geq 1)$ at any time point will not experience detectable failure if the number of latent occurrences at that time is less than r. In those two works they assume that the *conditional distribution of T given M*

can be written down in terms of the incomplete beta function, $IB(S(t);M-r+1,r)$, as

$$P(T>t|M) = \mathbb{1}(M=0) + IB(S(t);M-r+1,r)\,\mathbb{1}(M \geq r \geq 1), \tag{4}$$

where

$$\begin{aligned} IB(S(t);M-r+1,r) &= \sum_{j=0}^{r-1} \binom{M}{j} [F(t)]^j [S(t)]^{M-j} \\ &= M\binom{M-1}{r-1} \int_0^{S(t)} u^{M-r}(1-u)^{r-1} du. \end{aligned} \tag{5}$$

The unconditional survival function of T, i.e. the survival function of the population, $S_p(t)$, is related to the latent distribution, $S(t)$, as

$$\begin{aligned} S_p(t) &= E_M\left[P(T>t|M)\right] \\ &= P(M=0) + E_M\left[IB(S(t);M-r+1,r)\,\mathbb{1}(M \geq r \geq 1)\right]. \end{aligned} \tag{6}$$

As $\lim_{t\to+\infty} S(t) = 0$, we have $\lim_{t\to+\infty} S_p(t) = P(M=0)$, showing that $S_p(t)$ is improper whenever $P(M=0) > 0$. Indeed, $P(M=0)$ is the probability of a person being a long-term-survivor and depends only upon the distribution of M, irrespective of what r is. In fact, if $f(t)$ is the proper latent density function corresponding to $F(t)$ than

$$f_p(t) = f(t) E_M\left\{ M\binom{M-1}{r-1} [S(t)]^{M-r} [F(t)]^{r-1}\, \mathbb{1}(M \geq r \geq 1) \right\}. \tag{7}$$

The variable M can never be observed and should be modelled using a probabilistic assumption. We stress that the BCH model is a particular case of (6) where M is distributed as a Poisson random variable with mean θ, i.e, $M|\theta \sim \mathcal{P}(\theta)$ and $r=1$, which implies that activating any one of the M latent factors brings the observed failure, i.e. $T = \min_{1\leq m\leq M} T^*_{(m)}$. So (4) becomes

$$P(T>t|M) = \mathbb{1}(M=0) + [S(t)]^M\, \mathbb{1}(M \geq 1), \tag{8}$$

and (6) becomes

$$\begin{aligned} S_p(t) &= E_M\left[P(T>t|M)\right] \\ &= P(M=0) + E_M\left[S(t)^M\, \mathbb{1}(M \geq 1)\right] = \exp(-\theta F(t)). \end{aligned} \tag{9}$$

Let us suppose that the ik-th individual is potentially exposed to M_{ik} unobserved latent factors. The presence of any of which, i.e., if $M_{ik} \geq 1$ ultimately leads the patient to experiment the event. In cancer settings these latent factors may correspond to metastasis-competent tumour cells within the individual. In our database of HIV/AIDS certainly there could be several latent factors that increase the risk of death, but this is quite speculative and is not biologically motivated (as well in the cancer field). Instead we assume a single dominant latent factor. So, M_{ik} is a binary variable, with a Bernoulli distribution, $M_{ik}|\theta_k \sim Ber(\theta_k)$, where θ_k is the probability of an activation for the individuals in the k-th region and T^*_{ik} is the latent event time.

Note that we are assuming a common cure fraction for the individuals living in the same region (spatial cure fraction) and not an individual one. Given M_{ik} we have

$$\begin{aligned} S_p(t) &= E_M\left[P(T>t|M)\right] \\ &= P(M=0) + E_M\left[S(t)^M \mathbb{I}(M=1)\right] \\ &= 1-\theta+\theta S(t) \end{aligned} \tag{10}$$

which is the classic cure-rate model (Berkson and Gage (1952) and Farewell (1986)) with cure fraction $1-\theta$. The hazard function for the entire population is

$$h_p(t) = -\frac{d\left[\log\{S_p(t)\}\right]}{dt} = \frac{\theta f(t)}{1-\theta+\theta S(t)}. \tag{11}$$

4 Likelihood

Lets assume that T_m^* follows a two-parameter Weibull distribution $\mathcal{W}(\rho, e^{\eta(t)})$, whose density and survival functions are respectively

$$f(t|\rho, e^{\eta(t)}) = e^{\eta(t)}\rho t^{\rho-1}\exp(-t^\rho e^{\eta(t)}) \text{ and } S(t|\rho,\eta(t)) = \exp(-t^\rho e^{\eta(t)}). \tag{12}$$

We will introduce covariates through the scale parameter, $e^{\eta(t)}$, allowing it to vary across individuals. In order to use the information of the CD4 and account for the effect of these longitudinal measures in the time-to-event model we can also let the Weibull scale parameter to be a function of the random effects, $\boldsymbol{b}_{ik}$. This time-independent vector links the longitudinal and survival processes which means that it accounts for both the association between the two submodels and the correlation between the repeated measurements in the longitudinal process (conditional independence). So,

$$\eta_{ik}(t) = \mathbf{x}_{2ik}^\top \boldsymbol{\beta}_2 + \boldsymbol{\gamma}^\top \boldsymbol{b}_{ik}, \tag{13}$$

where $\boldsymbol{\gamma}$ is a vector that quantifies the effect of the CD4 values in the survival time; $\mathbf{x}_{2ik}$ is a vector of baseline covariates (can coincide with $\mathbf{x}_{1ik}$) and $\boldsymbol{\beta}_2$ is the respective vector of coefficients. Considering (12) we can rewrite (10) and (11) to get

$$S_p(t_{ik}) = \left[1-\theta_k+\theta_k\exp(-t_{ik}^\rho e^{\eta_{ik}})\right]. \tag{14}$$

and

$$h_p(t_{ik}) = \frac{\theta_k\rho t_{ik}^{\rho-1}e^{\eta_{ik}}}{S_p(t_{ik})\exp\{t_{ik}^\rho e^{\eta_{ik}}\}}. \tag{15}$$

where $1-\theta_k$ is the common cure fraction for the individuals living in the k-th region.

The issue of incorporating covariates in cure rate models is a challenge one. As pointed out in Banerjee et al (2004) and Cooner et al (2007) if M_{ik} is assumed to be Poisson distributed as in Chen et al (1999), covariates can be introduced through a suitable link function, g, in the cure fraction, so $\theta_k = g(\mathbf{x}_{ik}^\top\boldsymbol{\beta})$. In that case proper posteriors arise for the regression coefficients even under improper priors. But if M_{ik} is a Bernoulli variable, as is in our case, we should use vague but proper priors. Although this makes the parameters difficult to interpret and can often lead to poor

MCMC convergence. Regression on θ with flat priors on regression coefficients produces improper posteriors. On the other hand, regressing through the scale parameter in the latent Weibull distribution is always valid, yielding proper posteriors even with improper priors with an appropriate link.

Suppressing the covariates for easy of exposition let $\mathcal{D}_{ik} = \{\mathbf{y}_{ik}, T_{ik}, \delta_{ik}, M_{ik}\}$ be the complete data for the ik-th individual. The likelihood function for the joint model involves two components. The first component is the longitudinal process, denoted by L_{1ik}, and the second component involves the likelihood function of the time-to-event variable, T, denoted by L_{2ik}. The contribution of the ik-th subject to the complete data likelihood function (in a right-censored setting) can thus be written as

$$L_{ik}(\boldsymbol{\Omega}_{ik}|\mathcal{D}_{ik}) = L_{1ik}(\boldsymbol{b}_{ik}, \boldsymbol{\beta}_1, \sigma^2|\mathbf{y}_{ik}) L_{2ik}(\boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, M_{ik}|T_{ik}, \delta_{ik}) \tag{16}$$

where $\boldsymbol{\Omega}_{ik}$ stands for the collection of all the model parameters related to the ik-th individual,

$$\begin{aligned} &L_{1ik}(\boldsymbol{b}_{ik}, \boldsymbol{\beta}_1, \sigma^2|\mathbf{y}_{ik}) = \\ &= \prod_{j=1}^{n_{ik}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ - \frac{\left[y_{ik}(t_{ikj}) - \mathbf{x}_{1ik}^\top(t_{ikj})\boldsymbol{\beta}_1 - \mathbf{z}_{1ik}^\top(t_{ikj})\boldsymbol{b}_{ik}\right]^2}{2\sigma^2} \right\} \end{aligned} \tag{17}$$

and

$$\begin{aligned} &L_{2ik}(\boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, M_{ik}|T_{ik}, \delta_{ik}) = \\ &= P(T > t_{ik}|\boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, M_{ik})^{1-\delta_{ik}} \times \left(- \frac{d}{dt_{ik}} P(T > t_{ik}|\boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, M_{ik}) \right)^{\delta_{ik}} \end{aligned} \tag{18}$$

where $P(T > t_{ik}|\boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, Q_k, M_{ik})$ is as in equation (4) with $r = 1$ and

$$- \frac{d}{dt_{ik}} P(T > t_{ik}|\boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, M_{ik}) = M\left[S(t)\right]^{M-1} f(t)\, \mathbb{I}(M = 1) = f(t). \tag{19}$$

The introduction of the latent M_{ik} will facilitates the convergence of the MCMC algorithm. Although, is more meaningful to look to the conditional survival distribution $P(T > t_{ik}|\boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, M_{ik})$ after marginalizing out the M_{ik}. This marginalization, performed using an MCMC algorithm because of its analytical intractability, significantly reduces the estimation space to become the marginalized L_{2ik} (see Appendix 7):

$$\begin{aligned} \sum_{M_{ik}} L_{2ik}(\boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, M_{ik}|T_{ik}, \delta_{ik}) &= L_{2ik}(\boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}|T_{ik}, \delta_{ik}) \\ &= \left[h_p(t_{ik})\right]^{\delta_{ik}} S_p(t_{ik}). \end{aligned} \tag{20}$$

In order to have the joint likelihood one can simple take the product of all the N individual contributions to the likelihood:

$$\begin{aligned} L(\boldsymbol{\Omega}|\mathcal{D}) &= \prod_{k=1}^{K} \prod_{i=1}^{n_k} L_{ik}(\boldsymbol{\Omega}|\mathcal{D}) \\ &= \prod_{k=1}^{K} \prod_{i=1}^{n_k} L_{1ik}(\boldsymbol{b}_{ik}, \boldsymbol{\beta}_1, \sigma^2|\mathbf{y}_{ik}) \times \prod_{k=1}^{K} \prod_{i=1}^{n_k} L_{2ik}(\boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}|T_{ik}, \delta_{ik}), \end{aligned} \tag{21}$$

where $\mathcal{D}$ represents all the observed data.

5 Prediction of future values

An eminent issue of the joint models is the obtaintion of predictions for the longitudinal and survival outcomes (Sweeting and Thompson (2011, 2012), Rizopoulos (2011), Yu et al (2008) and Proust-Lima and Taylor (2009)). The ability to incorporate the trajectory of the 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 the survival probability in the coming years, from the time of the last visit or after the individual is censored. If the CD4 pattern is suggestive of an increase risk of death, the physician may decide to change the therapy to slow the progression of the disease.

Lets assume that we had obtained a longitudinal series of measurements from a new individual, along with its survival information up to time s. The data for this individual can be summarized in $\mathcal{D}_{new} = \{\mathbf{y}_{new}, T_{new} = s, \delta_{new} = 0\}$. Inferences about a new (future) longitudinal value for this individual at time $t > s$, say $\tilde{y}(t)$, can be obtained from its posterior predictive distribution,

$$p(\tilde{y}(t)|\mathcal{D}, \mathcal{D}_{new}) = \int\int p(\tilde{y}(t)|\mathcal{D}_{new}, \boldsymbol{b}_{new}, \boldsymbol{\Omega}) p(\boldsymbol{b}_{new}|\mathcal{D}_{new}, \boldsymbol{\Omega}) p(\boldsymbol{\Omega}|\mathcal{D}) d\boldsymbol{\Omega} d\boldsymbol{b}_{new}, \tag{22}$$

where $\boldsymbol{b}_{new}$ stands for the random effects vector of the new individual and $p(\boldsymbol{b}_{new}|.)$ is the respective posterior distribution. Similarly, the posterior predictive probability of surviving time $t > s$ given survival up to s is

$$\begin{aligned} & p(T_{new} > t \mid \mathcal{D}, T_{new} > s, \mathbf{y}_{new}) = \\ &= \int\int p(T_{new} > t|\mathcal{D}_{new}, \boldsymbol{b}_{new}) p(\boldsymbol{b}_{new}|T_{new} > s, \mathbf{y}_{new}, \boldsymbol{\Omega}) p(\boldsymbol{\Omega}|\mathcal{D})\, d\boldsymbol{\Omega}\, d\boldsymbol{b}_{new} \\ &= \int\int \frac{S_{new}(t \mid \mathbf{y}_{new}, \boldsymbol{b}_{new})}{S_{new}(s \mid \mathbf{y}_{new}, \boldsymbol{b}_{new})} p(\boldsymbol{b}_{new}|\mathcal{D}_{new}, \boldsymbol{\Omega}) p(\boldsymbol{\Omega} \mid \mathcal{D})\, d\boldsymbol{\Omega}\, d\boldsymbol{b}_{new}, \end{aligned} \tag{23}$$

where $S_{new}(.)$ is the marginal survival function (14) for the new individual.

Inside an MCMC framework obtain the posterior distribution of the random effects is easy. Besides that, in Winbugs, there exists the possibility to use the `cut` function, which allows the estimation of the the random effects for the new individual without contributing to the likelihood. So, the estimates of the remaining parameters are not influenced by the new data.

6 Analysis of the HIV/AIDS data

6.1 Illustration

Inside a Bayesian approach we applied the proposed methodology in sections 3 and 4 to the HIV/AIDS data described in section 2. To deal with the asymmetry in the longitudinal measures we transformed the observed CD4 counts to $\sqrt{\text{CD4}}$. We choose to model the transformed longitudinal biomarker trajectory for a specific patient through

a mixed effects model (section 3.1). Let $y_{ik}(t_{ikj}) \equiv y_{ikj}$ denote the square root of the j-th CD4 count measurement on the i-th patient living in one of the $27 = K$ Brazilian states, $k = 1, \ldots, 27$. We will assume

$$y_{ikj}|\boldsymbol{b}_{ik}, \boldsymbol{\beta}_1, \sigma^2 \sim \mathcal{N}(y^*_{ik}(t_{ikj}), \sigma^2),$$

$$y^*_{ik}(t_{ikj}) = \beta_{11} + \beta_{12}t_{ikj} + \beta_{13}\text{gender}_{ik} + \beta_{14}\text{age}_{ik} + \beta_{15}\text{PrevOI}_{ik} + b_{ik1} + b_{ik2}t_{ikj}, \quad (24)$$

where $\boldsymbol{b}_{ik} = (b_{ik1}, b_{ik2})$. For the latent survival times, T^*, two distributions were considered. A Weibull, $T^*_{ik} \sim \mathcal{W}(\rho, e^{\eta_{ik}(t)})$, and an Exponential, $T^*_{ik} \sim \mathcal{E}(e^{\eta_{ik}(t)})$, with

$$\eta_{ik}(t) = \beta_{21} + \beta_{22}\text{gender}_{ik} + \beta_{23}\text{age}_{ik} + \beta_{24}\text{PrevOI}_{ik} + \gamma_1 b_{ik1} + \gamma_2 b_{ik2}, \quad (25)$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$. Given this the posterior joint distribution, $\pi(\boldsymbol{\Omega}|\mathcal{D})$, will be proportional to

$$L(\boldsymbol{\Omega}|\mathcal{D}) \left[\prod_{k=1}^{K} \prod_{i=1}^{n_k} \pi(\boldsymbol{b}_{ik}|\boldsymbol{\Sigma}) \right] \pi(\boldsymbol{\Sigma})\pi(\boldsymbol{\beta}_1)\pi(\sigma^2)\pi(\boldsymbol{\beta}_2)\pi(\boldsymbol{\gamma})\pi(\rho)\pi(\boldsymbol{\theta}) \quad (26)$$

where $\boldsymbol{b} = (\boldsymbol{b}_{11}, \ldots, \boldsymbol{b}_{n_K K})$ stands for the complete random effects vector and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ is the vector of the different cure fractions which can be degenerated in a single point if we assume a common cure fraction for the population instead of a different one for each region. The parameter estimates for the proposed joint model in this work were obtained through the use of an MCMC simulation within the WinBUGS software. We used noninformative priors for all the parameters. With $\mathcal{G}$, $\mathcal{N}$, $\mathcal{W}ish$ and $\mathcal{U}$, denoting, respectively, Gamma, Normal, Wishart and Uniform distributions, the prior specifications were: $1/\sigma^2 \sim \mathcal{G}(0.01, 0.01)$, $\boldsymbol{\beta}_1 \sim \mathcal{N}_5(\mathbf{0}_5, 1000I_5)$, $\boldsymbol{\beta}_2 \sim \mathcal{N}_4(\mathbf{0}_4, 1000I_4)$, $\boldsymbol{b}_{ik}|\boldsymbol{\Sigma} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}^{-1} \sim \mathcal{W}ish(100I_2, \xi)$, $\gamma_1 \sim \mathcal{N}(0, 100)$, $\gamma_2 \sim \mathcal{N}(0, 100)$, $\theta_k \sim \mathcal{U}(0, 1)$, $\rho \sim \mathcal{G}(0.01, 0.01)$ if the latent survival times are assumed to have a Weibull distribution. We puted the degrees of freedom of the inverse Wishart $\xi = n/20 = 4653/20 \approx 233$ following a suggestion in Carlin and Louis (2001) page 279 to avoid confounding between fixed and random effects. The initial values of the parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ for sampling are obtained by modelling the longitudinal and survival data separately. Priors for σ^2, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\Sigma}$, $\boldsymbol{\gamma}$, ρ and $\boldsymbol{\theta}$ are motivated by their conjugacy and were assumed to be independent *a priori*.

A set of models was fitted in our analysis. The fixed effects from covariates `gender`, `age` and `PrevOI` are always included in both submodels. The best-fitting model will be selected based on the Deviance Information Criteria value (DIC, Spiegelhalter et al (2002)) and the value of the logarithm of the pseudomarginal likelihood (LPML, Ibrahim et al (2001)) which is well defined even in scenarios with an improper prior distribution.

Model I presents no cure fraction. Models II and III admit a fraction of individuals in the population that will not experiment the event of interest, i.e. $1 - \theta \neq 0$. Model II assumes that there is a unique probability of cure, common to all the 27 states. On the other hand model III assumes that the cure fraction varies between states.

Table 2 Models - Candidate Bayesian models assuming the sharing of the random effects. Several options were considered: absence of long-term survivors (a null cure fraction, $1-\theta=0$); a common cure fraction to all individuals in the dataset ($\theta_k=\theta$); a common cure fraction to the individuals in a region but different from other regions ($\theta_{k_1}\neq\theta_{k_2}$); two distributions were considered for the latent survival times - Weibull, $T^*\sim\mathcal{W}(.,.)$, and Exponential, $T^*\sim\mathcal{E}(.)$. DIC.L (DIC.S) and LPML.L (LPML.S) stands for the DIC and LPML values of the longitudinal (survival) submodel, respectively.

Model	**Cure fraction** $1-\theta$	$T^*\sim$	p_D	DIC	DIC.L	DIC.S	LPML	LPML.L	LPML.S
I	$1-\theta=0$	$\mathcal{W}$	6324.12	112071	109553	2517.67	-21318.18	-24335.55	3039.92
		$\mathcal{E}$	6326.45	112075	109555	2520.26	-21333.85	-24348.19	3039.87
II	$\forall k,\theta_k=\theta$	$\mathcal{W}$	6328.19	112084	109561	2522.91	-21363.54	-24379.72	3038.46
		$\mathcal{E}$	6330.92	112089	109564	2524.70	-21283.52	-24303.45	3039.02
III	$\theta_{k_1}\neq\theta_{k_2}$	$\mathcal{W}$	6332.32	112095	109558	2537.26	-21340.37	-24355.00	3033.51
		$\mathcal{E}$	6335.79	112103	109558	2545.35	-21368.56	-24385.93	3033.63

6.2 Results

Based on LPML and DIC measures, models II and VIII are those that provide the better fit to the data. They have in common the fact of not considering a cure fraction in the survival model. Furthermore, the use of the Weibull distribution for the latent survival time, comparatively to when the exponential distribution is used, greatly improves the value of these two measures – DIC decreases and LPML increases. The assumption that we have made considering a fraction of long-term survivors was not revealed in the results. The selected model conjugates spatial random effects with a Weibull distribution for the latent survival times with no cure fraction. As expected the DIC and LPML values for the longitudinal part (DIC.L and LPML.L, respectively) are approximately the same for all the 8 models, since the longitudinal model used is coincident. In what follows we will consider only the model II, because it produces better values of those two measures of adequacy.

Relatively to the posterior quantities of the parameters (*vide* table 3) we note that, for the longitudinal model, the differences between the two selected models are virtually non-existent. The survival model stand out some facts, namely indicating that none of the three baseline covariates has a significant impact on the risk of death.

The variance of the spatial frailties, σ_Q^2, in the selected model is very small, confirming that there are no geographical differences in the survival of these patients. Because the parameter γ_1 is negative, higher baseline CD4 values means a reduction in the risk. With the parameter γ_2 succeeds an interesting phenomenon, its estimate is positive. In practice this means that a greater slope in the CD4 trajectory is associated with a shorter survival time. Although not intuitive, this fact can find justification in the figure 3. The patient having the trajectory with the highest slope is simultaneously the one who suffers the event. From a biological point of view this is acceptable, because patients who have more oscillating trajectories, with high and low levels of CD4, followed by sharp falls or rises, are the ones in a more advanced stage of the disease. The healthy patients will maintain a more uniform CD4 trajectory over time.

Figure 2 show the spatial random effects for the selected model representing unobserved regional factors. To make the spatial distribution of the random effects visible, were obtained the quintiles of the distribution of the posterior medians of the

Table 3 Posterior quantities - Posterior mean and 95% HPD credibility intervals for the parameters of the model assuming a Weibull distribution (left side). Specific relative risk mean (RR) of the states (right side).

Parameter	*Mean*	*95% CI*	*State*	*RR*	*RR 95% CI*	*State*	*RR*	*RR 95% CI*
(β_{11})	17.37	$(17.13, 17.61)$	Dist. Federal	1.00	$]0.90, 1.11[$	Mato Grosso	1.02	$]0.85, 1.20[$
(β_{12})	1.80	$(1.71, 1.88)$	Goiás	1.01	$]0.87, 1.13[$	Bahia	1.00	$]0.90, 1.13[$
(β_{13})	−0.60	$(-0.88, -0.34)$	Minas Gerais	1.00	$]0.89, 1.10[$	Piauí	1.00	$]0.89, 1.14[$
(β_{14})	−0.48	$(-0.88, -0.021)$	Tocantins	1.00	$]0.86, 1.16[$	Pernambuco	1.01	$]0.88, 1.17[$
(β_{15})	−1.97	$(-2.26, -1.65)$	Roraima	1.00	$]0.91, 1.09[$	Alagoas	1.00	$]0.89, 1.10[$
σ^2	7.12	$(6.95, 7.29)$	Pará	1.00	$]0.93, 1.06[$	Sergipe	1.00	$]0.94, 1.07[$
σ^b_{11}	26.62	$(25.28, 27.80)$	Amapá	0.99	$]0.90, 1.08[$	São Paulo	0.99	$]0.89, 1.06[$
σ^b_{22}	4.52	$(4.20, 4.91)$	Maranhão	1.00	$]0.94, 1.07[$	Paraíba	1.01	$]0.92, 1.11[$
$\text{cor}(b_1, b_2)$	−0.42	$(-0.49, -0.34)$	Amazonas	1.00	$]0.90, 1.10[$	Rio G. Norte	1.00	$]0.93, 1.08[$
(β_{20})	−1.04	$(-1.39, -0.70)$	Paraná	1.00	$]0.94, 1.07[$	Ceará	1.00	$]0.90, 1.13[$
(β_{21})	0.13	$(-0.15, 0.38)$	St. Catarina	1.00	$]0.92, 1.07[$			
(β_{22})	−0.08	$(-0.39, 0.21)$	Mato Gr. Sul	1.00	$]0.92, 1.06[$			
(β_{23})	0.18	$(-0.07, 0.44)$	Acre	1.00	$]0.93, 1.09[$			
γ_1	−0.03	$(-0.06, -0.001)$	Rondônia	1.00	$]0.89, 1.10[$			
γ_2	0.24	$(0.15, 0.32)$	Rio de Janeiro	1.00	$]0.91, 1.08[$			
σ^2_Q	0.005	$(0.0001, 0.021)$	Espírito Santo	0.99	$]0.90, 1.08[$			
ρ	1.46	$(1.30, 1.61)$	Rio G. Sul	1.00	$]0.88, 1.14[$			

specific relative risks, $\exp\{Q_k\}$. Although the map exhibit some spatial variation, we emphasize that all the 95% HPD credibility intervals for the relative risk (table 3) indicate no states with a lower risk relatively to others. Either way it appears that the states in the South-east tend to have a relative risk lower than the others.

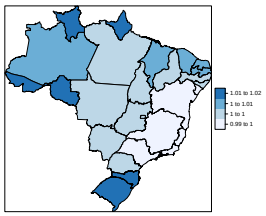

Fig. 2 Spatial frailties - Brazil's map where we present the spatial heterogeneities the model **II** with a **Weibull** distribution for the latent survival times.

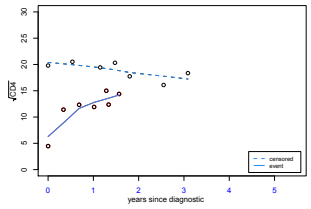

Fig. 3 Longitudinal profile of 2 patients - The lines in the graph were obtained through the LOWESS method in R.

6.3 Future values

In this section we will address the issue of predictions as discussed in section 5, in order to calculate the conditional probability of surviving for some time later relatively to the last measurement time of the CD4. It should be noted that we had some difficulties in performing predictions, since the individuals in the dataset who died were not necessarily those showing the worst CD4 trajectory, read to have a negative slope. In fact, as we see in Figure 3, the individual who suffers the event has a positive slope, while the censored one presents a negative.

We have made 24 predictions for real individuals (present in the database) who died and had 6 or more repeated measures (*vide* table 4). The aim was to obtain con-

ditional survival probabilities knowing that the individual was alive at the last CD4 measurement (censored) to see if the model was capable to predict the correct survival times. We were careful to remove these individuals from the sample at the time of obtaining their posterior estimates. Figures 4 - 5 show a dual information. On the one hand the CD4 median trajectory and its 95% credibility interval and on the other the conditional survival probability with their 95% credibility intervals. The predictions were made so that each individual presented 20 CD4 measurements, including the observed ones.

Based on the mean of the posterior median survival time of each individual, it is clear that, in general, the individual is supposed to live longer than what occurred in reality. This can be justified by the small percentage of deaths in our database, resulting in a shrinkage of these individuals towards the overall mean of the survival time.

For the individuals with the lowest CD4 counts, it is found that after 1 or 2 years, the range of credibility intervals for the CD4 count is high, making the predictions inaccurate.

What seems to have the most influence on survival time prediction (figures 4 and 5) is the value of the intercept, i.e. the first CD4 value. If not, watch out on the individuals 242 and 329 (two males under 50 years and a history of opportunistic infections). The former, even though he has a less favorable CD4 trajectory (decreasing) he has a posterior median survival time higher than the latter subject which has a increasing longitudinal biomarker trajectory. However, the first value of $\sqrt{\text{CD4}}$ of the individual 242 is higher than that of patient 329 (20.45 to 17.41). It is further noted that most of the 24 subjects, 15 in this case) have a increasing longitudinal trajectory of the biomarker, therefore it is difficult to the model to predict the median survival time. This goes back to what we said earlier about the fact that many patients who died had increasing CD4 trajectories. Moreover, there is strong evidence that the model indicates that a high slope may be a predictor of death. This is the case of patients 1349, 1415 and 3376. These three individuals also have in common the fact that the specific random effect for the slope, b_{2ik}, is greater than 1 (2.1, 1.45 and 1.15). Could this be a sign that the immune system is already responding poorly!

Acknowledgements The authors wish to thank Maria Goretti Fonseca and Cláudia Coeli by the database.

Table 4 24 real patients - Characteristics presented by 24 patients in the database used to make predictions. We can see the values of $\sqrt{\text{CD4}}$, sex, age, the presence/absence of previous opportunistic infections, the time (in years) of the last measure of CD4 and the real survival time (years).

Patient	gender	age	PrevOI	State	last measurement time	true survival time	predicted survival time - model II
835	0	0	0	Mato Grosso	2.40	2.54	3.40
2472	0	0	0	São Paulo	3.87	4.05	5.37
3013	0	0	0	São Paulo	2.74	2.96	4.24
944	0	0	1	Pernambuco	1.48	2.60	2.48
832	1	0	0	Mato Grosso	2.65	3.30	3.65
1076	1	0	0	São Paulo	2.54	3.41	5.04
1129	1	0	0	São Paulo	1.74	1.88	3.24
1349	1	0	0	Alagoas	1.47	1.61	2.48
2931	1	0	0	Mato G. Sul	2.66	3.43	3.66
161	1	0	1	Rio G. Norte	2.45	3.47	3.44
242	1	0	1	São Paulo	1.98	4.09	3.98
329	1	0	1	São Paulo	2.15	3.53	3.15
767	1	0	1	Rio G. Norte	2.34	3.34	3.34
1353	1	0	1	Alagoas	3.49	3.57	4.49
1404	1	0	1	Pernambuco	1.59	1.69	2.59
1793	1	0	1	Mato Grosso	1.59	2.20	2.59
3472	1	0	1	Alagoas	2.76	3.30	4.26
4560	1	0	1	São Paulo	2.13	2.77	3.13
536	1	1	0	Pernambuco	3.85	3.96	4.85
1415	1	1	0	Pernambuco	2.52	3.05	3.51
2956	1	1	0	Mato Grosso do sul	3.30	3.45	4.30
419	1	1	1	São Paulo	3.62	4.37	5.11
1453	1	1	1	São Paulo	2.59	3.21	4.09
3376	1	1	1	São Paulo	1.56	1.61	2.57

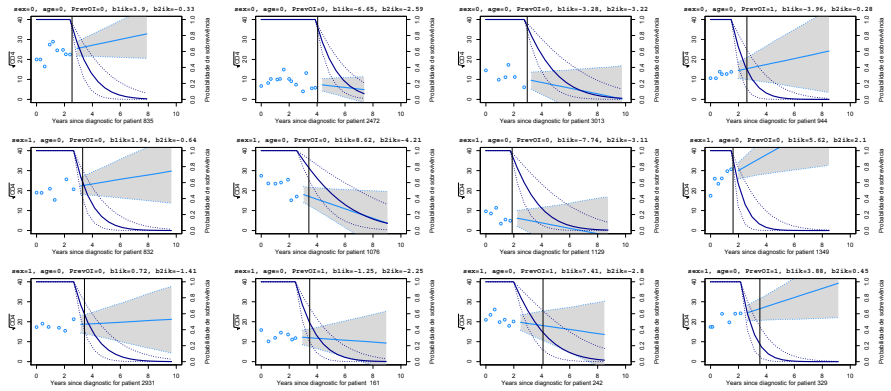

Fig. 4 Predictions for the first 12 patients - Median trajectory and predicted survival probability after the last CD4 measurement for the model II with a Weibull distribution for the latent survival times. b_{1ik} and b_{2ik} stand for the specific random effects values. The dotted lines delimit the 95% credibility intervals and the vertical lines indicates the true survival time for each individual.

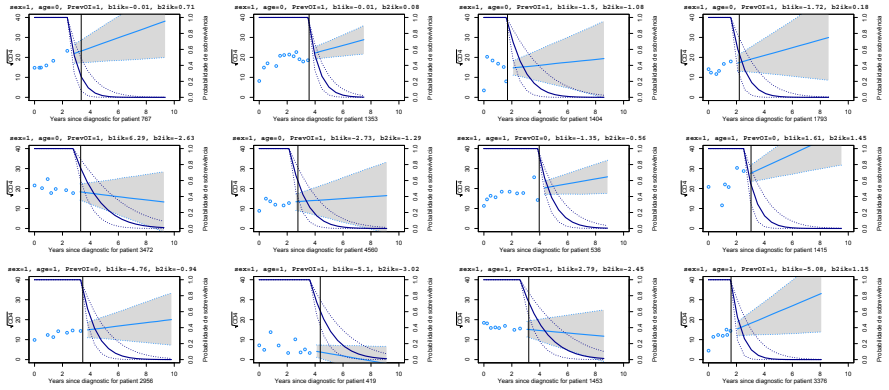

Fig. 5 Predictions for the last 12 patients - Median trajectory and predicted survival probability after the last CD4 measurement for the model II with a Weibull distribution for the latent survival times. b_{1ik} and b_{2ik} stand for the specific random effects values. The dotted lines delimit the 95% credibility intervals and the vertical lines indicates the true survival time for each individual.

7 Appendix

$$
\begin{aligned}
& L_{2ik}(\boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma} | T_{ik}, \delta_{ik}) \\
=& \sum_{M_{ik}} L_{2ik}(\boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, M_{ik} | T_{ik}, \delta_{ik}) \\
=& E_{M_{ik}} \left[\left(-\frac{d}{dt_{ik}} P(T > t_{ik} | \boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, M_{ik}) \right)^{\delta_{ik}} \times \right. \\
& \left. \times P(T > t_{ik} | \boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, M_{ik})^{1-\delta_{ik}} \right] \\
=& \left[-\frac{d}{dt_{ik}} S_p(t_{ik}) \right]^{\delta_{ik}} S_p(t_{ik})^{\delta_{ik}} \\
=& \left[h_p(t_{ik}) \right]^{\delta_{ik}} S_p(t_{ik}) \qquad (27)
\end{aligned}
$$

where we use the fact that δ_{ik} equals 1 or 0 and that the derivative can be interchanged with the expectation.

In particular, note that $P(T > t_{ik} | \boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, M_{ik})$ equals $S(t_{ik} | \boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma})$ if $M_{ik} = 1$ and equals 1 if $M_{ik} = 0$. That is, if the latent factor is absent, the subject is a long-term-survivor. Marginalizing over the Bernoulli distribution for M_{ik}, we obtain for the ikth patient the survival function $S_p(t_{ik} | \boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}) \equiv P(T > t_{ik} | \boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}) = 1 - \theta_k + \theta_k S(t_{ik})$, which is the classic cure-rate model with cure fraction $1 - \theta_k$, as in Berkson and Gage (1952) and Farewell (1986).

References

Banerjee S, Carlin B, Gelfand A (2004) Hierarchical modeling and analysis for spatial data. Chapman & Hall/CRC, Boca Raton, Florida.

Berkson J, Gage R (1952) Survival curve for cancer patients following treatment. Journal of the American Statistical Association 47(259):501–515, URL http://www.jstor.org/stable/2281318

Brown ER, Ibrahim JG (2003) Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. Biometrics 59(3):686–693

Carlin B, Louis T (2001) Bayes and Empirical Bayes Methods for Data Analysis, Second Edition. Chapman & Hall

Chen M, Ibrahim J, Sinha D (1999) A new bayesian model for survival data with a surviving fraction. Journal of the American Statistical Association 94(447):909–919, URL http://www.jstor.org/stable/2670006

Chen MH, Ibrahim JG, Sinha D (2004) A new joint model for longitudinal and survival data with a cure fraction. Journal Of Multivariate Analysis 91(1):18–34

Chi Y, Ibrahim J (2006) Joint models for multivariate longitudinal and multivariate survival data. Biometrics 62:432–445, DOI DOI:10.1111/j.1541-0420.2005.00448.x, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1541-0420.2005.00448.x/abstract

Cooner F, Banerjee S, McBean AM (2006) Modelling geographically referenced survival data with a cure fraction. Statistical Methods in Medical Research 15(4):307–324, DOI 10.1191/0962280206sm453oa, URL http://smm.sagepub.com/content/15/4/307.abstract, http://smm.sagepub.com/content/15/4/307.full.pdf+html

Cooner F, Banerjee S, Carlin B, Sinha D (2007) Flexible cure rate modeling under latent activation schemes. Journal of the American Statistical Association 102(478):560–572, DOI doi:10.1198/016214507000000112

DeGruttola V, Tu XM (1994) Modelling progression of cd4 lymphocyte count and its relationship to survival time. Biometrics 50:1003–1014

Elashoff R, Li G, Li N (2008) A joint model for longitudinal measurements and survival data in the presence of multiple failure types. Biometrics 64:762–771

Ewell M, Ibrahim J (1997) The large sample distribution of the weighted log rank statistic under general local alternatives. Lifetime Data Analysis 3:5–12, URL http://dx.doi.org/10.1023/A:1009690200504, 10.1023/A:1009690200504

Farewell V (1982) The use of mixture models for the analysis of survival data with long-term survivors. Biometrics 38(4):1041–1046

Farewell V (1986) Mixture models in survival analysis: Are they worth the risk? Canadian Journal of Statistics 14:257–262

Faucett CL, Thomas DC (1996) Simultaneously modelling censored survival data and repeatedly measured covariates: a gibbs sampling approach. Statistics in Medicine 15(15):1663–1685, DOI 3.0.CO;2-1, URL http://dx.doi.org/3.0.CO;2-1

Fonseca M, CM C, Lucena F, Veloso V, Carvalho M (2010) Accuracy of a probabilistic record linkage strategy applied to identify deaths among cases reported to the brazilian aids surveillance database. Cad Saúde Pública [online] 26:1431–1438, DOI 10.1590/S0102-311X2010000700022., URL http://www.scielosp.org/scielo.php?pid=S0102-311X2010000700022&script=sci_arttext&tlng=en

Gray R, Tsiatis A (1989) A linear rank test for use when the main interest is in differences in cure rates. Biometrics 45(3):899–904, URL http://www.jstor.org/stable/2531691

Gu Y, Sinha D, Banerjee S (2011) Analysis of cure rate survival data under proportional odds model. Lifetime Data Analysis 17:123–134, URL http://dx.doi.org/10.1007/s10985-010-9171-z, 10.1007/s10985-010-9171-z

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

Henderson R, Diggle P, Dobson A (2000) Joint modelling of longitudinal measurements and event time data. Biostatistics 1(4):465–480, DOI 10.1093/biostatistics/1.4.465, URL http://dx.doi.org/10.1093/biostatistics/1.4.465

Ibrahim J, Chen MH, Sinha D (2004) Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine studies. Statistica Sinica 14(3):863–883

Ibrahim JG, Chen MH, Sinha D (2001) Bayesian Survival Analysis. Springer-Verlag

Proust-Lima C, Taylor J (2009) Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: a joint modeling approach. Biostatistics 10(3):535–549, DOI 10.1093/biostatistics/kxp009, URL http://biostatistics.oxfordjournals.org/content/10/3/535.abstract, http://biostatistics.oxfordjournals.org/content/10/3/535.full.pdf+html

Rizopoulos D (2011) Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. Biometrics 67:819–829, DOI 10.1111/j.1541-0420.2010.01546.x, URL http://dx.doi.org/10.1111/j.1541-0420.2010.01546.x

Rizopoulos D, Ghosh P (2011) A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. Statistics in Medicine 30(12):1366–1380, DOI 10.1002/sim.4205, URL http://dx.doi.org/10.1002/sim.4205

Rizopoulos D, Verbeke G, Molenberghs G (2008) Shared parameter models under random effects misspecification. Biometrika 95(1):63–74, DOI 10.1093/biomet/asm087, URL http://biomet.oxfordjournals.org/content/95/1/63.abstract

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

Samji H, Cescon A, Hogg R, Modur S, Althoff KN, Buchacz K, Burchell AN, Cohen M, Gebo KA, Gill MJ, Justice A, Kirk G, Klein MB, Korthuis PT, Martin J, Napravnik S, Rourke SB, Sterling TR, Silverberg MJ, Deeks S, Jacobson LP, Bosch RJ, Kitahata MM, Goedert JJ, Moore R, Gange SJ (2013) Closing the gap: Increases in life expectancy among treated hiv-positive individuals in the united states and canada. PLoS ONE 8(12):e81,355, DOI 10.1371/journal.pone.0081355, URL http://dx.doi.org/10.1371%2Fjournal.pone.0081355

Souza-Jr P, Szwarcwald C, Castilho E (2007) Delay in introducing antiretroviral therapy in patients infected by hiv in brazil, 2003-2006. Clinical Science 62(5):579–584, DOI 10.1590/S1807-59322007000500008, URL

http://www.scielo.br/scielo.php?pid=S1807-59322007000500008&script=sci_abstract&tlng=pt

Spiegelhalter D, Best N, Carlin B, Van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society (Series B) 64(4):583–639

Stangl D, Greenhouse J (1998) Assessing placebo response using bayesian hierarchical survival models. Lifetime Data Analysis 4:5–28, URL http://dx.doi.org/10.1023/A:1009644308160, 10.1023/A:1009644308160

Sweeting M, Thompson S (2011) Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. Biometrical Journal 53(5):750–763, DOI 10.1002/bimj.201100052, URL http://dx.doi.org/10.1002/bimj.201100052

Sweeting M, Thompson S (2012) Making predictions from complex longitudinal data, with application to planning monitoring intervals in a national screening programme. Journal of the Royal Statistical Society: Series A (Statistics in Society) 175(2):569–586, DOI 10.1111/j.1467-985X.2011.01005.x, URL http://dx.doi.org/10.1111/j.1467-985X.2011.01005.x

Tseng YK, Hsieh FS, Wang JL (2005) Joint modelling of accelerated failure time and longitudinal data. Biometrika 92(3):587–603

Tsiatis AA, DeGruttola V, Wulfsohn MS (1995) Modelling the relationship of survival to longitudinal data measured with error: applications to survival cd4 counts in patients with aids. J Amer Statist Assoc 90:27–37

Van Sighem A, Gras L, Reiss P, Brinkman K, de Wolf F (2010) Life expectancy of recently diagnosed asymptomatic hiv-infected patients approaches that of uninfected individuals. AIDS 24(10):1527–1535, DOI 10.1097/QAD.0b013e32833a3946, URL http://journals.lww.com/aidsonline/Abstract/2010/06190/Life_expectancy_of_recently_diagnosed_asymptomatic.15.aspx

Wang Y, Taylor JMG (2001) Jointly modelling longitudinal and event time data, with applications to aids studies. J Amer Statist Assoc 96:895–905

Wu M, Carroll R (1988) Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. Biometrics 44:175–188

Wulfsohn M, Tsiatis A (1997) A joint model for survival and longitudinal data measured with error. Biometrics 53:330–339

Xu J, Zeger SL (2001) Joint analysis of longitudinal data comprising repeated measures and times to events. Journal Of The Royal Statistical Society Series C-Applied Statistics 50:375–387

Yakovlev A, Tsodikov A (1996) Stochastic models of tumor latency and their biostatistical applications. World Scientific, New Jersey

Yakovlev A, Asselain B, Bardou VJ, Fourquet A, Hoang T, Rochefordiere A, Tsodikov A (1993) A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. In: Biometrie et Analyse de Donnees Spatio-Temporelles, vol 12, Société Française de Biométrie, ENSA Rennes, France, pp 66–82

Yu MG, Taylor J, Sandler H (2008) Individual prediction in prostate cancer studies using a joint longitudinal survival cure model. Journal of the American Statistical

Association 103:178–187
Zhang S, Mueller P, Do KA (2010) A bayesian semiparametric survival model with longitudinal markers. Biometrics 66(2):435–443, DOI 10.1111/j.1541-0120.2009.01276.x