Biometrics 67, 819–829
September 2011
DOI: 10.1111/j.1541-0420.2010.01546.x

Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data

Dimitris Rizopoulos*

Department of Biostatistics, Erasmus Medical Center, PO Box 2040, 3000 CA Rotterdam, the Netherlands
* *email:* d.rizopoulos@erasmusmc.nl

Summary. In longitudinal studies it is often of interest to investigate how a marker that is repeatedly measured in time is associated with a time to an event of interest. This type of research question has given rise to a rapidly developing field of biostatistics research that deals with the joint modeling of longitudinal and time-to-event data. In this article, we consider this modeling framework and focus particularly on the assessment of the predictive ability of the longitudinal marker for the time-to-event outcome. In particular, we start by presenting how survival probabilities can be estimated for future subjects based on their available longitudinal measurements and a fitted joint model. Following we derive accuracy measures under the joint modeling framework and assess how well the marker is capable of discriminating between subjects who experience the event within a medically meaningful time frame from subjects who do not. We illustrate our proposals on a real data set on human immunodeficiency virus infected patients for which we are interested in predicting the time-to-death using their longitudinal CD4 cell count measurements.

Key words: Area under the curve; Discrimination; ROC methodology; Shared parameter model; Survival analysis; Time-dependent covariates.

1. Introduction

In longitudinal studies often interest lies in the relation between a longitudinally measured marker and a time-to-event outcome. Standard examples of this setting include studies on human immunodeficiency virus (HIV) patients for whom we are interested in associating longitudinal CD4 cell count measurements with the time to death, and prostate cancer studies in which we are interested in associating longitudinal prostate specific antigen (PSA) level measurements with the time to recurrence. Such studies have given rise to a fast-developing area of biostatistics research that deals with the joint modeling of longitudinal and time-to-event data. Nice overviews of this field are given by Tsiatis and Davidian (2004) and Yu et al. (2004).

Several extensions of the standard joint model formulated by Faucett and Thomas (1996) and Wulfsohn and Tsiatis (1997) have been proposed in the literature. These include, among others, the consideration of multiple longitudinal outcomes (Brown, Ibrahim, and DeGruttola, 2005; Rizopoulos and Ghosh, 2011), handling multiple failure times (Elashoff, Li, and Li, 2008), modeling flexibly the subject-specific longitudinal profiles (Ding and Wang, 2008; Rizopoulos, Verbeke, and Lesaffre, 2009), and replacing the relative risk models by accelerated failure time models (Tseng, Hsieh, and Wang, 2005). Another use of these models that has lately gained some increasing interest is to obtain subject-specific predictions for either the longitudinal or survival outcomes (Taylor, Yu, and Sandler, 2005; Garre et al., 2008; Yu, Taylor, and Sandler, 2008; Proust-Lima and Taylor, 2009). It is more than evident that such prediction tools would be valuable in everyday medical practice. That is, medical experience coupled with an estimate of the survival probability of a patient that takes into account all available information for this patient would enable physicians to make better informed decisions regarding their actions and thus improve clinical output. However, in order to move toward this direction and the use of joint models in clinical practice certainly more work is required. In this article, we build on the previous work in this new field within the joint modeling framework, and we propose several new developments. In particular, we take a rather different approach than Proust-Lima and Taylor (2009), and Garre et al. (2008), who assumed latent classes to explain the interrelationships between the longitudinal and survival outcomes. Even though a latent class formulation of the joint model simplifies computations (i.e., dealing with sums rather than integrals), the existence of latent heterogeneity in the population that drives both processes is not always justifiable. Therefore, we follow here the more traditional formulation that assumes continuous random effects to underlie both processes (Wulfsohn and Tsiatis, 1997; Henderson, Diggle, and Dobson, 2000). Under the continuous-random-effects joint modeling framework, we primarily focus on the survival outcome, and we propose a Monte Carlo approach to estimate survival probabilities and their standard errors based on the output of a fitted joint model. Using this machinery, we consider dynamic subject-specific predictions, and illustrate how survival probabilities are updated as additional measurements of the longitudinal outcome are recorded. Following,

we focus on prospective accuracy measures for the longitudinal marker, and assess its capability in distinguishing between subjects who are about to experience the event and subjects who have a much lower risk. In particular, under a general definition of prediction rules, we present suitable definitions of sensitivity and specificity measures, and we determine the longitudinal marker's accuracy using receiver operating characteristic (ROC) methodology. In this last part we follow similar arguments to Zheng and Heagerty (2007); Antolini, Boracchi, and Biganzoli (2005); and Heagerty and Zheng (2005), but suitably adapted to the joint modeling framework.

Our proposals are exemplified in a case study that considers 467 patients with advanced HIV infection during antiretroviral treatment who had failed or were intolerant of zidovudine therapy. The main aim of this study was to compare the efficacy and safety of two alternative antiretroviral drugs, namely didanosine (ddI) and zalcitabine (ddC) in the time to death. Patients were randomly assigned to receive either ddI or ddC, and CD4 cell counts were recorded at study entry, where randomization took place, as well as at 2, 6, 12, and 18 months thereafter. By the end of the study 188 patients had died, resulting in about 60% censoring. More details regarding the design of this study can be found in Abrams et al. (1994). For our illustrations our aim is to investigate how well CD4 cell count is capable of discriminating between subjects who died within a medically relevant time interval after their last assessment and subjects who lived longer than that. That is, for a future patient from the same population, we would like to inform the treating physicians about her survival probability that is calculated based on her baseline covariates and available CD4 measurements, such that they could further decide upon their actions.

The rest of the article is organized as follows. Section 2 briefly reviews the joint modeling framework and Section 3 illustrates how survival probabilities can be estimated in a joint modeling framework. Section 4 presents the development of accuracy measures for joint models and in Section 5 we illustrate their use in the AIDS data set. Finally, Section 6 refers to the results of a simulation study and Section 7 concludes the article.

2. Joint Modeling Framework

2.1 *Submodels Specification*

Let T_i denote the observed failure time for the ith subject ($i = 1, \ldots, n$), which is taken as the minimum of the true event time T_i^* and the censoring time C_i, that is, $T_i = \min(T_i^*, C_i)$. Furthermore, we define the event indicator as $\delta_i = I(T_i^* \leqslant C_i)$, where $I(\cdot)$ is the indicator function that takes the value 1 if the condition $T_i^* \leqslant C_i$ is satisfied, and 0 otherwise. For the longitudinal responses, let $y_i(t)$ denote the value of the longitudinal outcome at time point t for the ith subject. The actual observed longitudinal data for subject i consist of the measurements $y_{ij} = \{y_i(t_{ij}),\ j = 1, \ldots, n_i\}$ taken at time points t_{ij}.

We will denote the true and unobserved value of the longitudinal outcome at time t as $m_i(t)$. To quantify the effect of $m_i(t)$ on the risk for an event, a standard option is to use a relative risk model of the form

$$\begin{aligned} h_i(t \mid \mathcal{M}_i(t), w_i) &= \lim_{dt \to 0} \Pr\left\{t \leqslant T_i^* < t + dt \,\middle|\, T_i^* \geqslant t, \mathcal{M}_i(t), w_i\right\} / dt \\ &= h_0(t) \exp\left\{\gamma^\top w_i + \alpha m_i(t)\right\}, \end{aligned} \tag{1}$$

where $\mathcal{M}_i(t) = \{m_i(u); 0 \leqslant u < t\}$ denotes the history of the true unobserved longitudinal process up to time point t, $h_0(\cdot)$ denotes the baseline risk function, and w_i is a vector of baseline covariates with a corresponding vector of regression coefficients γ. The baseline risk function can be left unspecified or can be approximated using step functions or splines.

In the above definition of the survival model we used the true unobserved value of the underlying longitudinal covariate $m_i(t)$. In order to quantify the effect of this covariate to the risk for an event, we need to estimate $m_i(t)$ and successfully reconstruct the complete longitudinal history $\mathcal{M}_i(t)$. To achieve this we will use the available measurements $y_{ij} = \{y_i(t_{ij}),\ j = 1, \ldots, n_i\}$ of each subject and a set of modeling assumptions. In particular, for the remainder of this article we focus on normal data and we postulate a linear mixed effects model to describe the subject-specific longitudinal evolutions

$$\begin{aligned} y_i(t) &= m_i(t) + \varepsilon_i(t) \\ &= x_i^\top(t)\beta + z_i^\top(t) b_i + \varepsilon_i(t), \quad \varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2), \end{aligned} \tag{2}$$

where β denotes the vector of the unknown fixed-effects parameters, b_i the vector of random effects, $x_i(t)$ and $z_i(t)$ denote row vectors of the design matrices for the fixed and random effects, respectively, and $\varepsilon_i(t)$ is the measurement error term with variance σ^2. Finally, the random effects b_i are assumed normally distributed with mean zero and covariance matrix D, and independent of $\varepsilon_i(t)$.

2.2 *Maximum Likelihood Estimation*

We base the estimation of the joint models' parameters presented in Section 2.1 on a maximum likelihood approach that maximizes the log-likelihood function corresponding to the joint distribution of the time-to-event and longitudinal outcomes $\{T_i, \delta_i, y_i\}$. To define this distribution, we will assume that the vector of time-independent random effects b_i underlies both the longitudinal and survival processes. This means that these random effects account for both the association between the longitudinal and event process, and the correlation between the repeated measurements in the longitudinal outcome. Formally, we have that,

$$p(T_i, \delta_i, y_i \mid b_i; \theta) = p(T_i, \delta_i \mid b_i; \theta)\, p(y_i \mid b_i; \theta) \tag{3}$$

$$p(y_i \mid b_i; \theta) = \prod_j p\{y_i(t_{ij}) \mid b_i; \theta\}, \tag{4}$$

where θ denotes the parameter vector, y_i the vector of longitudinal responses of the ith subject, and $p(\cdot)$ a probability density function. Under the modeling assumptions presented in the previous section, and the above conditional independence assumptions, the joint log-likelihood contribution for

the ith subject can be formulated as

$$\log p(T_i, \delta_i, y_i; \theta) = \log \int p(T_i, \delta_i \mid b_i; \theta) \times \prod_j p\{y_i(t_{ij}) \mid b_i; \theta\} p(b_i; \theta)\, db_i, \quad (5)$$

where the likelihood of the survival part is written as

$$p(T_i, \delta_i \mid b_i; \theta) = \{h_i(T_i \mid \mathcal{M}_i(T_i); \theta)\}^{\delta_i} S_i(T_i \mid \mathcal{M}_i(T_i); \theta),$$

with $h_i(\cdot)$ given by (1), and

$$S_i(t \mid \mathcal{M}_i(t), w_i) = \Pr\big(T_i^* > t \mid \mathcal{M}_i(t), w_i\big) = \exp\left\{-\int_0^t h_i(s \mid \mathcal{M}_i(s); \theta)\, ds\right\}. \quad (6)$$

Maximization of the log-likelihood function $\ell(\theta) = \sum_i \log p(T_i, \delta_i, y_i; \theta)$ is a computationally challenging task due to the fact that the integral with respect to the random effects in (5), and the integral in the definition of the survival function in (6) do not have an analytical solution. For the approximation of these integrals standard numerical integration techniques, such as Monte Carlo and Gaussian quadrature, have been employed in the joint modeling literature (Henderson et al., 2000; Song, Davidian, Tsiatis, 2002). Furthermore, Rizopoulos et al. (2009) and Ye, Lin, and Taylor (2008) have recently discussed the use of Laplace approximations for joint models, which can be especially useful in high-dimensional random-effects settings (e.g., when splines or high-order polynomials are used in the random-effects design matrix).

3. Predicted Survival Probabilities

In this section, we focus on expected survival. In particular, based on a joint model fitted on a sample of size n, we are interested in predicting survival probabilities for a new subject i that has provided a set of longitudinal measurements $\mathcal{Y}_i(t) = \{y_i(s); 0 \leqslant s \leqslant t\}$ (dependence on baseline covariates is assumed but suppressed for ease of exposition). An important feature that we need to carefully take into account is the fact that $y_i(t)$ represents an endogenous time-dependent covariate (Kalbfleisch and Prentice, 2002, Section 6.3). The implication this feature presents in the calculation of subject-specific survival probabilities is that $y_i(t)$ is directly related to the failure mechanism, that is providing longitudinal measurements up to t in fact implies survival up to this time point. Hence, it is more relevant to focus on the conditional probability of surviving time $u > t$ given survival up to t, that is,

$$\pi_i(u \mid t) = \Pr\big(T_i^* \geqslant u \mid T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_n; \theta\big), \quad (7)$$

where $\mathcal{D}_n = \{T_i, \delta_i, y_i; i = 1, \ldots, n\}$ denotes the sample on which the joint model was fitted and on which we wish to base our predictions. Using assumption (3), we observe that (7) can be written as

$$\begin{aligned}
&\Pr(T_i^* \geqslant u \mid T_i^* > t, \mathcal{Y}_i(t); \theta) \\
&= \int \Pr\big(T_i^* \geqslant u \mid T_i^* > t, \mathcal{Y}_i(t), b_i; \theta\big) \\
&\quad \times p\big(b_i \mid T_i^* > t, \mathcal{Y}_i(t); \theta\big)\, db_i \\
&= \int \Pr\big(T_i^* \geqslant u \mid T_i^* > t, b_i; \theta\big)\, p\big(b_i \mid T_i^* > t, \mathcal{Y}_i(t); \theta\big)\, db_i \\
&= \int \frac{S_i\{u \mid \mathcal{M}_i(u, b_i, \theta); \theta\}}{S_i\{t \mid \mathcal{M}_i(t, b_i, \theta); \theta\}}\, p\big(b_i \mid T_i^* > t, \mathcal{Y}_i(t); \theta\big)\, db_i, \quad (8)
\end{aligned}$$

where $S_i(\cdot)$ is given by (6), and furthermore we have explicitly noted that the longitudinal history $\mathcal{M}_i(\cdot)$, as approximated by the linear mixed effects model, is a function of both the random effects and the parameters. Based on (8), we can derive a first-order estimate of $\pi_i(u \mid t)$ using the empirical Bayes estimate for b_i, that is

$$\widetilde{\pi}_i(u \mid t) = S_i\{u \mid \mathcal{M}_i(u, \widehat{b}_i, \widehat{\theta}); \widehat{\theta}\} \times S_i\{t \mid \mathcal{M}_i(t, \widehat{b}_i, \widehat{\theta}); \widehat{\theta}\} + O\big(n_i^{-1}\big), \quad (9)$$

where $\widehat{\theta}$ denotes the maximum likelihood estimates, $\widehat{b}_i = \operatorname{argmax}_b \{\log p(T_i^* > t, \mathcal{Y}_i(t), b; \widehat{\theta})\}$, and n_i is the number of longitudinal measurements for subject i. We expect that this estimator will work well in practice; however, deriving its standard error and/or confidence intervals for $\pi_i(u \mid t)$ is a rather difficult task due to the fact that we need to account for the variability of both the maximum likelihood and empirical Bayes estimates. To overcome this problem and produce valid standard errors for the estimate of $\pi_i(u \mid t)$, we propose to follow an asymptotic Bayesian formulation of the joint model and derive the posterior expectation of (7). More specifically, we are interested in

$$\Pr\big(T_i^* \geqslant u \mid T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_n\big) = \int \Pr\big(T_i^* \geqslant u \mid T_i^* > t, \mathcal{Y}_i(t); \theta\big)\, p(\theta \mid \mathcal{D}_n)\, d\theta. \quad (10)$$

The first part of the integrand, and as shown above, is given by (8). For the second part, which is the posterior distribution of the parameters given the observed data, we use arguments of standard asymptotic Bayesian theory (Cox and Hinkley, 1974, Section 10.6), and assume that the sample size n is sufficiently large such that $\{\theta \mid \mathcal{D}_n\}$ can be well approximated by $\mathcal{N}(\widehat{\theta}, \widehat{\mathcal{H}})$, with $\widehat{\mathcal{H}} = \widehat{\operatorname{var}}(\widehat{\theta})$. Combining (10) with (8) and $\{\theta \mid \mathcal{D}_n\} \sim \mathcal{N}(\widehat{\theta}, \widehat{\mathcal{H}})$, a Monte Carlo estimate of $\pi_i(u \mid t)$ can be obtained using the following simulation scheme:

Step 1. Draw $\theta^{(l)} \sim \mathcal{N}(\widehat{\theta}, \widehat{\mathcal{H}})$.
Step 2. Draw $b_i^{(l)} \sim \{b_i \mid T_i^* > t, \mathcal{Y}_i(t), \theta^{(l)}\}$.
Step 3. Compute $\pi_i^{(l)}(u \mid t) = S_i\{u \mid \mathcal{M}_i(u, b_i^{(l)}, \theta^{(l)}); \theta^{(l)}\} \times [S_i\{t \mid \mathcal{M}_i(t, b_i^{(l)}, \theta^{(l)}); \theta^{(l)}\}]^{-1}$.
Step 4. Repeat Steps 1–3 for each subject i, $l = 1, \ldots, L$ times, where L denotes the number of Monte Carlo samples.

Steps 1 and 3 are straightforward to perform; on the contrary, the posterior distribution of the random effects given the observed data in Step 2 is of nonstandard form, and thus

a more sophisticated approach is required to sample from it. Here, we make use of a Metropolis–Hastings algorithm with independent proposals from a multivariate t distribution centered at the empirical Bayes estimates $\widehat{b}_i$, with scale matrix $\widehat{\text{var}}(\widehat{b}_i) = \{-\partial^2 \log p(T_i^* > t, \mathcal{Y}_i(t), b; \widehat{\theta})/\partial b^\top \partial b\big|_{b=\widehat{b}_i}\}^{-1}$, and four degrees of freedom. Our justification for a multivariate t proposal is twofold. First, Rizopoulos, Verbeke, and Molenberghs (2008) have recently shown that, as n_i increases, the leading term of the log posterior distribution of the random effects is the logarithm of the density of the linear mixed model $\log p\{\mathcal{Y}_i(t) \mid b_i; \theta^{(l)}\} = \sum_j \log p\{y_i(t_{ij}) \mid b_i; \theta^{(l)}\}$, which is quadratic in b_i and will resemble the shape of a multivariate normal distribution, and second, for small n_i, the heavier tails of the t distribution will ensure sufficient coverage. The realizations $\{\pi_i^{(l)}(u \mid t),\ l = 1, \ldots, L\}$ can be used to derive estimates of $\pi_i(u \mid t)$, such as

$$\widehat{\pi}_i(u \mid t) = \text{median}\big\{\pi_i^{(l)}(u \mid t), l = 1, \ldots, L\big\}$$

or

$$\widehat{\pi}_i(u \mid t) = L^{-1} \sum_{l=1}^{L} \pi_i^{(l)}(u \mid t), \tag{11}$$

and compute, standard errors using the sample variance over the Monte Carlo samples, and confidence intervals using the Monte Carlo sample percentiles. Compared to (9), estimators (11) not only provide a straightforward manner to calculate standard errors but they are also expected to yield more accurate results because they properly approximate the integrals in the definition of $\pi_i(u \mid t)$.

4. Prospective Accuracy for Endogenous Time-Dependent Covariates

4.1 *Definitions*

The assessment of the predictive performance of time-to-event models has received a lot of attention in the statistical literature. In general, the developed methodology has concentrated on either calibration, that is, how well the model predicts the observed data (Schemper and Henderson, 2000) or discrimination, that is, how well can the model discriminate between patients who will have the event from patients who will not (Harrell et al., 1982; Pencina et al., 2008). Here, we focus on discrimination and in particular, we rely on an ROC approach to assess the predictive ability of the longitudinal biomarker.

As also noted in the previous section, an inherent characteristic of the studies that require a joint modeling approach to answer the scientific questions of interest is their dynamic nature. Namely, as longitudinal information is collected for a subject, we can continuously update the predictions of her survival probabilities. Taking this feature into account, it is often of medical relevance to distinguish between patients who are about to experience the event within a time frame after the last measurement, and patients who are going to surpass this time frame. Therefore, in this setting a useful property of the longitudinal marker would be to be capable of discriminating between these patients. In particular, let us assume that we have collected longitudinal measurements $\mathcal{Y}_i(t) = \{y_i(s); 0 \leqslant s \leqslant t\}$ up to time point t for subject i. We are interested in events occurring in the medically relevant time frame $(t,\ t + \Delta t]$ within which the physician can take an action to improve the survival chance of the patient. Using an appropriate function of the marker history $\mathcal{Y}_i(t)$, we can define a prediction rule to discriminate between patients of high and low risk for an event. For instance, in the AIDS data set introduced in Section 1, we could consider values of CD4 cell count smaller than a specific threshold as predictive for death. Since we are in a longitudinal context, we have the flexibility of determining which values of the longitudinal history of the patient will contribute to the specification of the prediction rule. To allow for full generality we consider a vector of threshold values c based on which we term $\mathcal{S}_i(t, k, c) = \{y_i(s) \leqslant c_s; k \leqslant s \leqslant t\}$ as success (i.e., occurrence of the event of interest), and $\mathcal{F}_i(t, k, c) = \mathbb{R}^{r(k,t)} \setminus \{y_i(s) \leqslant c_s; k \leqslant s \leqslant t\}$ as a failure, where $\mathbb{R}^n$ denotes the n-dimensional Euclidean space, and $r(k, t)$ the number of longitudinal measurements taken in the interval $[k, t]$. The value of $k \geqslant 0$ specifies which past marker values contribute to the rule, and c_s denotes the threshold value at time point s. Typically, all thresholds c_s will depend on a single scalar parameter that will posit a meaningful relation between them (see example below). In the definitions of $\mathcal{S}_i(t, k, c)$ and $\mathcal{F}_i(t, k, c)$ we have made the convention that smaller values for the marker are associated with higher risk for death—in the opposite case these should be adjusted accordingly. Under these definitions, we formulate sensitivity as

$$\Pr\big\{\mathcal{S}_i(t, k, c) \mid T_i^* > t, T_i^* \in (t, t + \Delta t]; \theta\big\}, \tag{12}$$

and specificity as

$$\Pr\big\{\mathcal{F}_i(t, k, c) \mid T_i^* > t, T_i^* > t + \Delta t; \theta\big\}. \tag{13}$$

Three important notes regarding these formulations are the following. First, under the general definitions of Heagerty and Zheng (2005), (12) and (13) correspond to a cumulative sensitivity and dynamic specificity, respectively. Second, contrary to the analogous definitions of Zheng and Heagerty (2007) who used the last available measurement of the longitudinal outcome $y_i(t)$ to define a prediction rule and formulate sensitivity and specificity, (12) and (13) provide us with the capability of using a more elaborate function of the longitudinal history to reach a decision regarding the predicted failure status of subject i at time $t + \Delta t$. For example, in the AIDS data set we could define a composite prediction rule that is not based only on the last available measurement but rather on the last two or last three measurements of a patient. Furthermore, it could be of relevance to consider different threshold values for each of these measurements, for instance, we could define as success the event that the pre-last CD4 cell count is c and the last one $0.5c$, indicating that a 50% decrease is strongly indicative for death. As we illustrate in the following section, such prediction rules are evaluated in a straightforward manner under the joint modeling framework. Finally, baseline covariate adjustments are seamlessly incorporated in the above definitions by introducing them in design matrix W of the survival submodel (1).

The overall discrimination capability of the longitudinal marker can be assessed using the corresponding ROC curve

$$\text{ROC}_t^{\Delta t}(p) = \text{TP}_t^{\Delta t}\Big\{\big[\text{FP}_t^{\Delta t}\big]^{-1}(p)\Big\},$$

where p is in [0, 1], $\mathrm{TP}_t^{\Delta t}(c)$ denotes the true positive rate, $\mathrm{FP}_t^{\Delta t}(c)$ the false positive rate, and $[\mathrm{FP}_t^{\Delta t}]^{-1}(p) = \inf_c\{c : \mathrm{FP}_t^{\Delta t}(c) \leqslant p\}$. The corresponding area under the ROC curve (AUC) is obtained via $\mathrm{AUC}_t^{\Delta t} = \int_0^1 \mathrm{ROC}_t^{\Delta t}(p)\, dp$. Time-dependent ROCs and AUCs can be used to assess the performance of the longitudinal marker at different time points over the follow-up period. Alternatively, we may also summarize the discrimination power of the marker over the whole follow-up period, using a weighted average of AUCs. More specifically, we propose the use of

$$\mathrm{C}_{dyn}^{\Delta t} = \int \mathrm{AUC}_t^{\Delta t} \Pr(T_i^* > t)\, dt \Big/ \int \Pr(T_i^* > t)\, dt,$$

where $\Pr(T_i^* > t)$ is the marginal survival probability and is used to take into account that not all time points contribute equally to the comparison due to censoring. This is similar in spirit to the approach of Antolini et al. (2005). Other weight functions can be utilized as well, and the optimal choice depending on the focus of $\mathrm{C}_{dyn}^{\Delta t}$ is an open question. Note also that $\mathrm{C}_{dyn}^{\Delta t}$ depends on the length Δt of the time interval of interest, and therefore different models may exhibit different discrimination power for different Δt.

4.2 *Estimation*

As also noted in Section 3, longitudinal markers are typical examples of endogenous time-dependent covariates, a feature that should be accounted for in the estimation of the accuracy measures introduced in Section 4.1. In that respect, the joint modeling framework is greatly advantageous because it provides a complete specification for the bivariate distribution of the true event times and the longitudinal measurements $\{T_i^*,\ y_i\}$. Thus, we can simultaneously take both censoring and measurement error into account. In the following, we assume that we have fitted the joint model in our data set and we have obtained the maximum likelihood estimates $\widehat{\theta}$ and their asymptotic covariance matrix $\widehat{\mathcal{H}}$. Moreover, for ease of exposition, we will only focus on the estimation of sensitivity, since estimation of specificity proceeds analogously.

We will follow a similar approach as in Section 3 and derive an appropriate simulation scheme to produce a Monte Carlo estimate of sensitivity and its standard error. More specifically, we observe that (12) is written as (condition on parameters and covariates is assumed but omitted from the notation):

$$\begin{aligned}&\Pr\{\mathcal{S}_i(t,k,c) \mid T_i^* > t, T_i^* \in (t, t+\Delta t]\}\\ &\quad = \frac{\Pr\big\{\mathcal{S}_i(t,k,c), T_i^* \in (t, t+\Delta t] \mid T_i^* > t\big\}}{1 - \Pr\big(T_i^* > t + \Delta t \mid T_i^* > t\big)}.\end{aligned}$$

Under assumptions (3) and (4), and the joint model's definition, we can obtain further simplifications for the numerator and denominator. In particular, the numerator takes the form

$$\begin{aligned}&\Pr\big\{\mathcal{S}_i(t,k,c), T_i^* \in (t, t+\Delta t] \mid T_i^* > t\big\}\\ &\quad = \int \Pr\big\{\mathcal{S}_i(t,k,c), T_i^* \in (t, t+\Delta t] \mid T_i^* > t, b_i\big\}\\ &\qquad \times p\big(b_i \mid T_i^* > t\big)\, db_i\\ &\quad = \int \left[\prod_{s=k}^{t} \Phi\left\{\frac{c_s - m_i(s)}{\sigma}\right\}\right]\\ &\qquad \times \left[1 - \frac{S_i\{t + \Delta t \mid \mathcal{M}_i(t+\Delta t, b_i)\}}{S_i\{t \mid \mathcal{M}_i(t, b_i)\}}\right]\\ &\qquad \times p\big(b_i \mid T_i^* > t\big)\, db_i,\end{aligned} \tag{14}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Similarly for the denominator, we obtain

$$\begin{aligned}&\Pr\big(T_i^* > t + \Delta t \mid T_i^* > t\big)\\ &\quad = \int \frac{S_i\{t + \Delta t \mid \mathcal{M}_i(t+\Delta t, b_i)\}}{S_i\{t \mid \mathcal{M}_i(t, b_i)\}}\, p\big(b_i \mid T_i^* > t\big)\, db_i.\end{aligned} \tag{15}$$

Thus, we observe that sensitivity is rewritten as the ratio of the expected values of $\mathcal{E}_1(b_i, \theta) = [\prod_{s=k}^{t} \Phi\{\frac{c_s - m_i(s, b_i, \theta)}{\sigma}\}] \times [1 - \frac{S_i\{t+\Delta t \mid \mathcal{M}_i(t+\Delta t, b_i, \theta); \theta\}}{S_i\{t \mid \mathcal{M}_i(t, b_i, \theta); \theta\}}]$ and $\mathcal{E}_2(b_i, \theta) = S_i\{t + \Delta t \mid \mathcal{M}_i(t + \Delta t, b_i, \theta); \theta\} \times S_i\{t \mid \mathcal{M}_i(t, b_i, \theta); \theta\}^{-1}$ with respect to the marginal posterior distribution $p(b_i \mid T_i^* > t)$. Noting that this distribution is proportional to

$$p\big(b_i \mid T_i^* > t\big) \propto \int p(\mathcal{Y}_i(t) \mid b_i)\, S_i\{t \mid \mathcal{M}_i(t, b_i)\}\, p(b_i)\, d\mathcal{Y}_i(t), \tag{16}$$

and combining equations (14)–(16), we derive the following simulation scheme

Step 1. Draw $\theta^{(l)} \sim \mathcal{N}(\widehat{\theta}, \widehat{\mathcal{H}})$.
Step 2. Draw $\mathcal{Y}_i^{(l)}(t) \sim \mathcal{N}\{X_i\beta^{(l)} + Z_i b_i^{(l-1)}, [\sigma^{(l)}]^2\}$.
Step 3. Draw $b_i^{(l)} \sim \{b_i \mid T_i^* > t, \mathcal{Y}_i^{(l)}(t), \theta^{(l)}\}$.
Step 4. Compute $\mathcal{E}_1(b_i^{(l)}, \theta^{(l)})$ and $\mathcal{E}_2(b_i^{(l)}, \theta^{(l)})$.
Step 5. Repeat Steps 1–4, $l = 1, \ldots, L$ times, where L denotes the desired number of Monte Carlo samples.

Similarly to Section 3, Step 1 is used to account for the variability of the maximum likelihood estimates, and Step 3 is implemented using a Metropolis–Hastings algorithm with independent proposals from a multivariate t distribution centered at the empirical Bayes estimates $\widehat{b}_i^{(l)}$, and with scale matrix the covariance matrix of these estimates $\widehat{\mathrm{var}}(\widehat{b}_i^{(l)})$. Based on the realizations $\{\mathcal{E}_1(b_i^{(l)}, \theta^{(l)}), l = 1, \ldots, L\}$ and $\{\mathcal{E}_2(b_i^{(l)}, \theta^{(l)}), l = 1, \ldots, L\}$ we obtain the Monte Carlo estimate of sensitivity with

$$\widehat{\Pr}\big\{\mathcal{S}_i(t,k,c) \mid T_i^* > t, T_i^* \in (t, t+\Delta t]\big\} = \frac{\sum_l \mathcal{E}_1\big(b_i^{(l)}, \theta^{(l)}\big)}{L - \sum_l \mathcal{E}_2\big(b_i^{(l)}, \theta^{(l)}\big)}.$$

The corresponding standard error can be estimated using the Monte Carlo standard errors of the numerator and denominator of (12), and the Delta method. In particular, we have

$$s.e.\big(\widehat{\Pr}\big\{\mathcal{S}_i(t,k,c) \mid T_i^* > t, T_i^* \in (t, t+\Delta t]\big\}\big) = \{gVg^\top\}^{1/2},$$

where

$$g = L\Bigg[1 \Big/ \Big\{L - \sum_l \mathcal{E}_2\big(b_i^{(l)}, \theta^{(l)}\big)\Big\}, \sum_l \mathcal{E}_1\big(b_i^{(l)}, \theta^{(l)}\big) \Big/ \Big\{L - \sum_l \mathcal{E}_2\big(b_i^{(l)}, \theta^{(l)}\big)\Big\}^2\Bigg],$$

and

$$\begin{aligned}&\text{vech}(V)\\ &= L^{-1}\big[\text{var}\big\{\mathcal{E}_1\big(b_i^{(l)}, \theta^{(l)}\big)\big\},\\ &\qquad \text{cov}\big\{\mathcal{E}_1\big(b_i^{(l)}, \theta^{(l)}\big), \mathcal{E}_2\big(b_i^{(l)}, \theta^{(l)}\big)\big\}, \text{var}\big\{\mathcal{E}_2\big(b_i^{(l)}, \theta^{(l)}\big)\big\}\big].\end{aligned}$$

As stated above, the estimation of specificity proceeds in an analogous manner. Having estimated both sensitivity and specificity, it is straightforward to construct the corresponding ROC and calculate the AUC. Finally, for the estimation of $C_{dyn}^{\Delta t}$ we also require an estimate of the marginal survival function $\Pr(T_i^* > t)$, which can be easily obtained using the Kaplan–Meier product limit estimator. The integral in the numerator of $C_{\text{dyn}}^{\Delta t}$ does not have a closed-form solution, and thus a numerical method must be employed for its evaluation. Standard choices are Simpson's and Gaussian quadrature rules; here we use the latter and in particular a 15-point Gauss–Kronrod rule.

5. AIDS Data Set

We return to the AIDS data set introduced in Section 1. As mentioned there, we are interested in using the CD4 cell count measurements of the patients in order to estimate expected survival probabilities and also assess how well we are able to discriminate between patients with high risk and low risk of dying. The CD4 cell counts are known to exhibit right skewed shapes of distribution, and therefore, for the remainder of this analysis we will work with the $\text{CD4}^{1/4}$ cell count values. Descriptive plots and further information for the observed data are given in Web Section 1.1.

We start our analysis by fitting a joint model to the available data. In particular, in the survival submodel we include as baseline covariates the treatment, the gender, a dummy variable for patients who had azidothymidine (AZT) failure and a dummy variable for patients who had a previous opportunistic infection, and as a time-dependent covariate the true underlying CD4 cell count. The baseline risk function is assumed piecewise constant with seven knots placed at equally spaced percentiles of the observed event times. For the longitudinal submodel, in the fixed-effects part we include the main effects of time, treatment, gender, the dummy for AZT failure, and the dummy for previous opportunistic infection, and the interaction effects of time with all the other covariates. In the random-effects design matrix, we include an intercept and a time term. The functional form of this joint model, as well as the estimated regression coefficients with their standard errors are presented in Web Section 1.2 and Web Table 1, respectively. The analysis has been performed in `R` using package `JM` (Rizopoulos, 2010).

We continue by answering a typical question of medical relevance, that is, given the baseline information for a patient and his available CD4 cell count measurements, what is the shape of his survival function along with its standard error. Under the fitted joint model, survival probabilities can be estimated using the Monte Carlo procedure described in Section 3. As an example we consider Patient 7 who is a male belonging to the ddC group, who had showed AZT intolerance, had a previous opportunistic infection, he provided four CD4 cell count measurements, and he was lost to follow-up at 14.33 months. Figure 1 depicts the median of 200 Monte Carlo realizations of $\pi_i(u \mid t)$ along with 95% pointwise confidence intervals based on the percentiles of $\pi_i(u \mid t)$ over the Monte

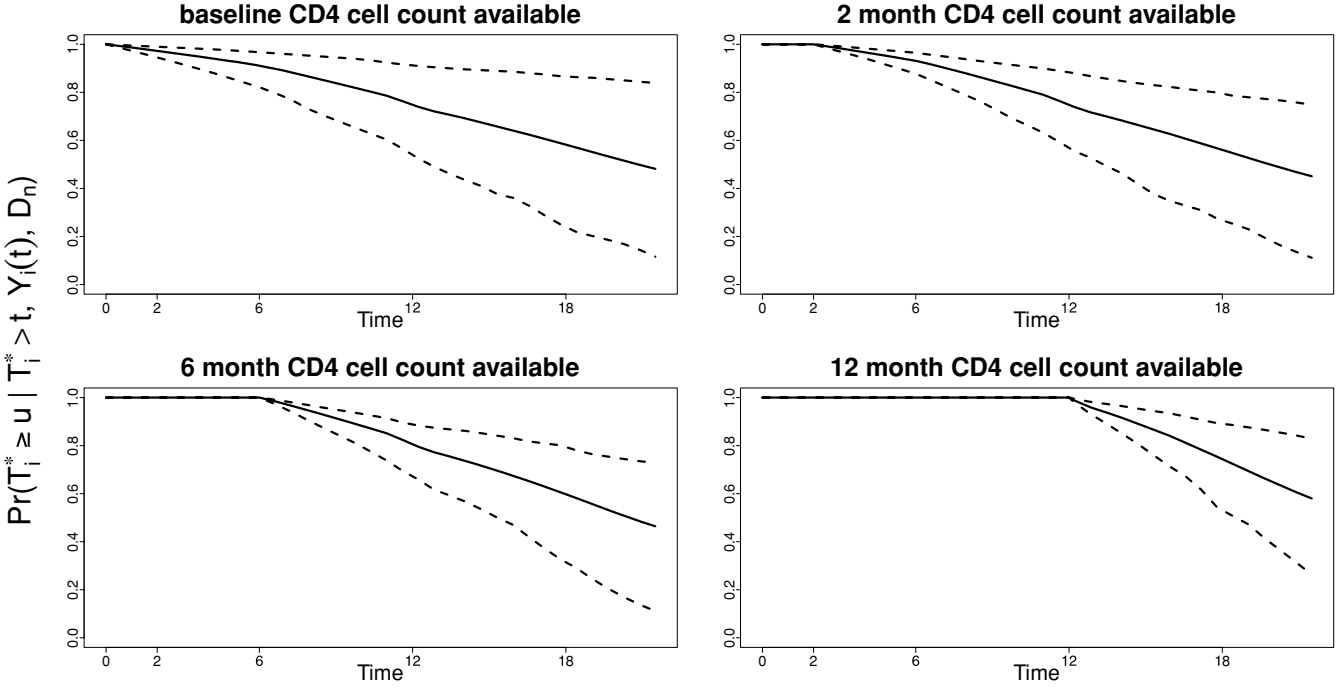


Figure 1. Predicted probabilities of survival for Patient 7, based on 200 Monte Carlo samples. The solid line depicts the median $\pi_i(u \mid t)$ over the Monte Carlo samples. The dashed lines depict 95% pointwise confidence intervals.

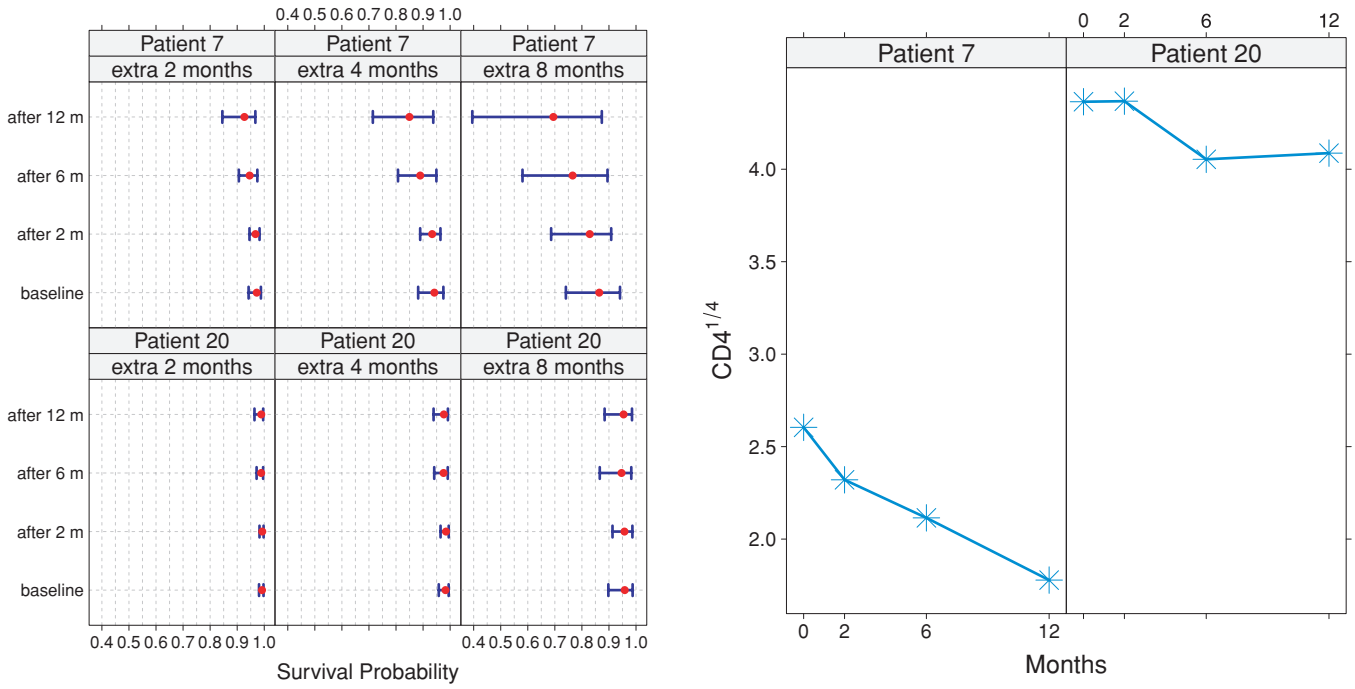


Figure 2. Left panel: predicted probabilities of survival for Patients 7 and 20 for a time frame of 2, 4, and 8 months after each recorded CD4 cell count measurement. The dot denotes the median $\pi_i(u \mid t)$ over 200 Monte Carlo samples, and the error bars correspond to 95% pointwise confidence intervals. Right panel: observed $CD4^{1/4}$ cell counts for Patients 7 and 20. This figure appears in color in the electronic version of this article.

Carlo samples for $t = 0$, 2, 6, and 12 months, and $u \in (t, t_{\max}]$ with $t_{\max}$ denoting the end of the study.

Another useful comparison is illustrated in Figure 2, which depicts $\pi_i(t + \Delta t \mid t)$ for $t = 0$, 2, 6, and 12 months, and $\Delta t = 2$, 4, and 8 months for Patient 7 and Patient 20 who had exactly the same baseline covariates, with the only difference that Patient 20 showed AZT failure and he was censored at 14.4 months.

We clearly observe how the changes in the CD4 cell count for the two patients are reflected in the point estimates of $\pi_i(u \mid t)$. In particular, focusing on $\Delta t = 8$ months (where differences are more noticeable), we observe that Patient 20 showed a rather stable CD4 cell count profile, which translates to high survival probabilities. In contrast, Patient 7 showed a deterioration in his CD4 cell counts and therefore was much less likely to survive an extra 8 months. An additional comparison with respect to the calculation of conditional survival probabilities, which illustrates the virtues of the joint modeling approach compared to a simple Cox analysis, and that only relies on the baseline CD4 cell count measurement, is illustrated in Web Section 1.3 and Web Figures 4 and 5.

In the following we focus on accuracy and how well the CD4 cell count can discriminate between patients. For our illustrations, we will consider a generic patient with the same baseline characteristics as of Patient 7. We first focus on a simple prediction rule that uses the last available CD4 cell count measurement to assess the predicted failure status for the generic patient. Figure 3 depicts the corresponding ROC curves, for $t = 0$, 2, 6, and 12 months, and $\Delta t = 2$, 4, and 8 months.

We clearly observe that for the target group of patients the CD4 cell count as a marker for death does not exhibit great discrimination power. Moreover, there do not seem to be considerable differences between the different time windows Δt. To check if a more elaborate prediction rule could improve the discriminative capability of the CD4 cell count, we consider the last two available measurements marker, and more specifically a decrease of 20% from the pre-last to the last one. That is $\mathcal{S}_i(t, k = t - 1, c) = \{y_i(t-1) \leqslant c,\ y_i(t) \leqslant 0.8c\}$, for various threshold values c. The ROC curves for $t = 0$, 2, 6, and 12 months under the simple and composite predictions rules and $\Delta t = 8$ months are illustrated in Figure 4.

Again we observe very small differences in the estimates of ROC curves, with the composite prediction rule performing slightly better at $t = 2$ months. As a further attempt to extract additional information from the CD4 cell count longitudinal profiles, we extend the relative risk submodel (1) and include in the linear predictor the term $m_i'(t) = \partial m_i(t)/\partial t$,

$$h_i(t \mid \mathcal{M}_i(t), w_i) = h_0(t) \exp\{\gamma^\top w_i + \alpha m_i(t) + \alpha_d m_i'(t)\},$$

that is we assume that the risk for an event at time t depends on both the true value of the trajectory at time t and the slope of the true trajectory at the same time point. The estimated regression coefficients with their standard errors for this joint model are presented in Web Table 2. Figure 5 presents a comparison of the corresponding ROC curves for $t = 0$, 2, 6, and 12 months under the two parameterizations and the simple prediction rule, for $\Delta t = 8$ months.

The association between the risk for an event and the slope of the true trajectory (parameter α_d in Web Table 2) was found very weak, which is also reflected in the ROC curves, since we observe an almost identical behavior between the two parameterizations. The corresponding AUCs and the

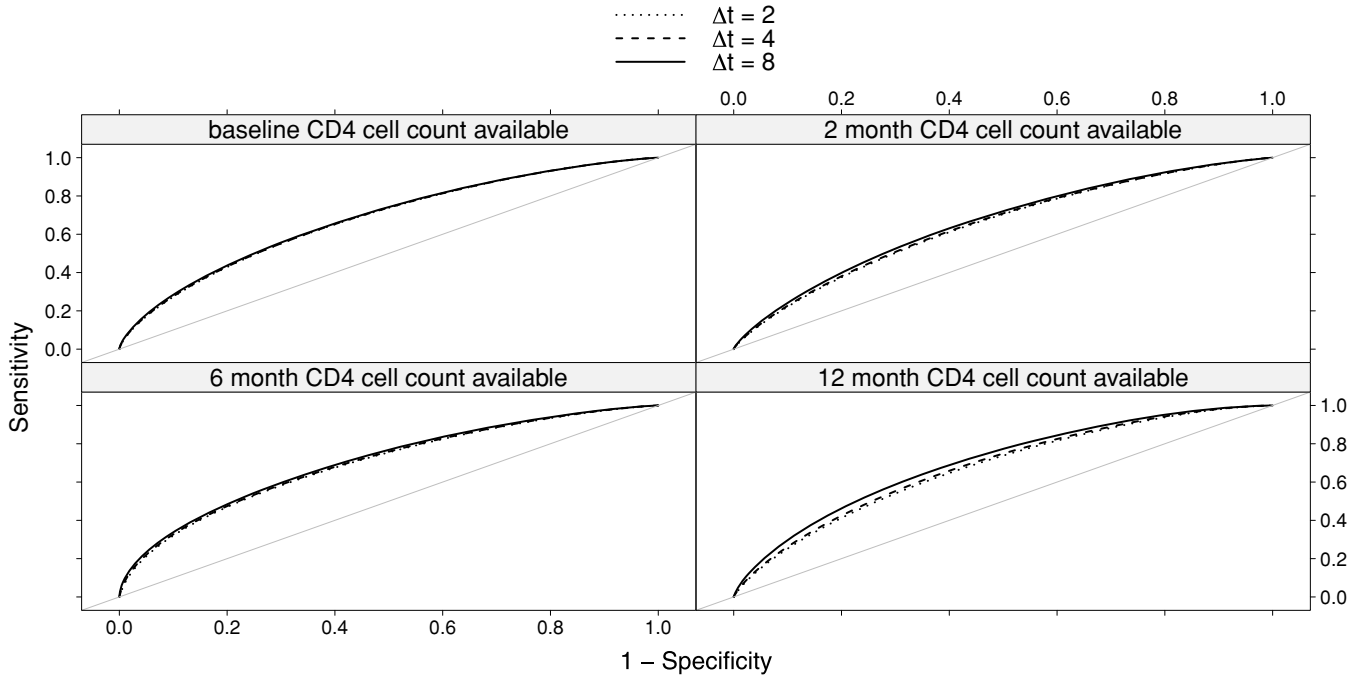


Figure 3. Time-dependent ROC curves for $\Delta t = 2$, 4, and 8 months, based on 2000 Monte Carlo samples.

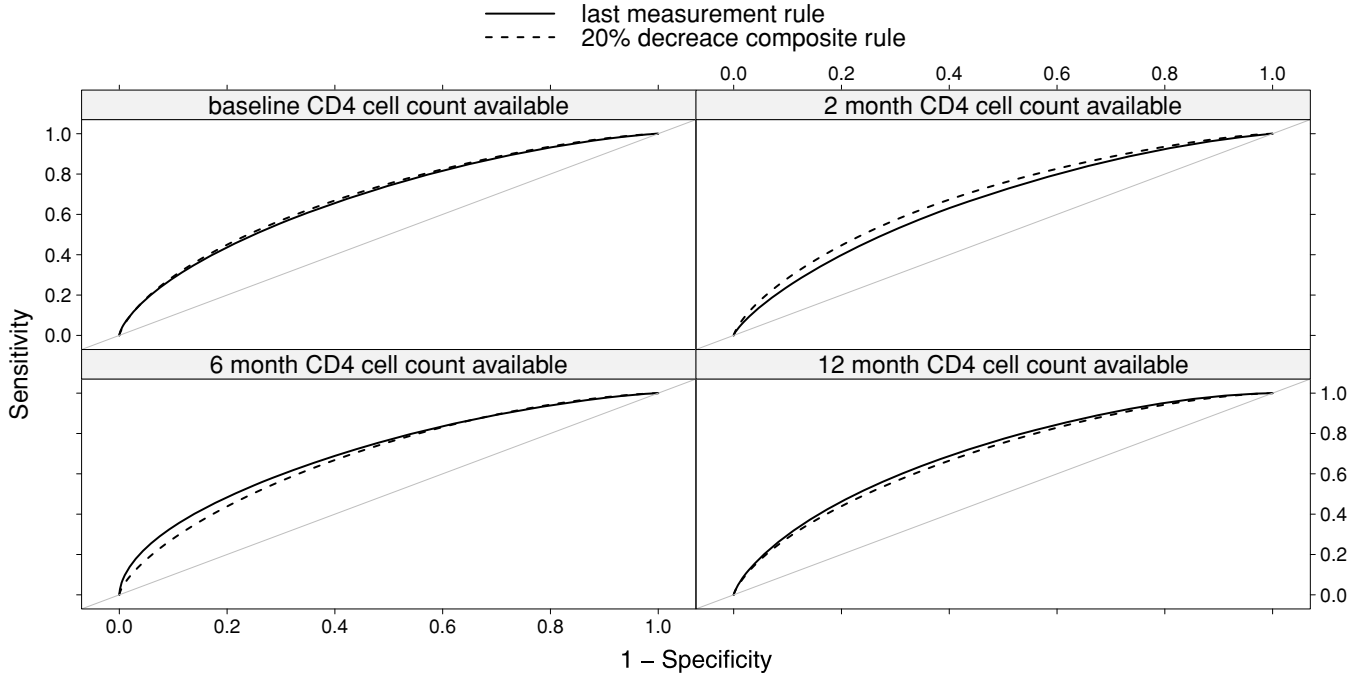


Figure 4. Time-dependent ROC curves under the simple (based on the last available measurement) and composite (based on the last two available measurements) prediction rules, for $\Delta t = 8$, based on 2000 Monte Carlo samples.

estimated $C_{dyn}^{\Delta t}$ index for all the above considered settings are presented in Table 1.

The results point to the same direction as the previous findings, namely the areas under the ROC curves for all combinations of t, Δt, prediction rule (i.e., simple and 20% decrease composite), and parameterization (i.e., $\alpha m_i(t)$ and $\alpha m_i(t) + \alpha_d m_i'(t)$) range from 0.67 to 0.72, indicating the moderate discriminative capability of the CD4 cell count for advanced HIV-infected patients. This general conclusion has been also corroborated by Goldman et al. (1996) who found that even though that the CD4 lymphocyte count is generally considered a valuable prognostic indicator for opportunistic disease or death, for the present data set it failed to prove a good surrogate marker.

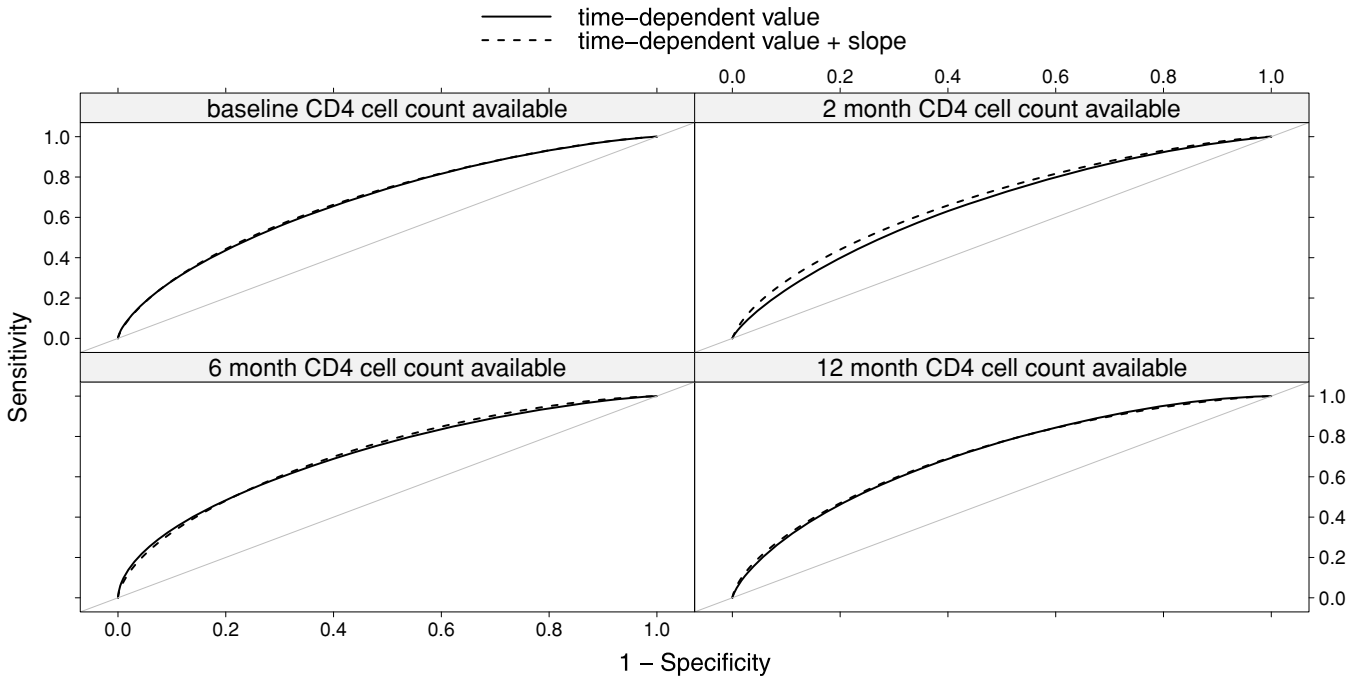


Figure 5. Time-dependent ROC curves under the time-dependent value and time-dependent value + slope parameterizations, for the simple prediction rule and for $\Delta t = 8$, based on 2000 Monte Carlo samples.

Table 1

Areas under the ROC curve and the estimated $C_{dyn}^{\Delta t}$ *index (based on* 2000 *Monte Carlo samples) for all predictions rules applied in the AIDS data set*

Prediction rule parameterization		Simple value		20% Relat. decrease value		Simple value + slope	
Δt	t	$AUC_t^{\Delta t}$	$C_{dyn}^{\Delta t}$	$AUC_t^{\Delta t}$	$C_{dyn}^{\Delta t}$	$AUC_t^{\Delta t}$	$C_{dyn}^{\Delta t}$
2	0	0.6839	0.6796	0.6846	0.6769	0.6834	0.6877
	2	0.6538		0.6802		0.6802	
	6	0.7029		0.6730		0.7022	
	12	0.6804		0.6658		0.6858	
4	0	0.6851	0.6834	0.6880	0.6823	0.6858	0.6922
	2	0.6572		0.6839		0.6828	
	6	0.7066		0.6800		0.7074	
	12	0.6886		0.6748		0.6954	
8	0	0.6885	0.6934	0.6961	0.6956	0.6912	0.7016
	2	0.6677		0.6951		0.6894	
	6	0.7146		0.6958		0.7190	
	12	0.7101		0.6956		0.7118	

6. Simulation Study

Assuming that the joint model of Section 2 is the appropriate modeling framework to capture the association between the longitudinal marker and the event process, and that the specification of the longitudinal and survival submodels is close to the truth, then the finite sample performance of the estimators for $\pi(u \mid t)$ presented in Section 3 and those for sensitivity and specificity of Section 4 will depend on the quality of the maximum likelihood estimates (MLEs) $\widehat{\theta}$ and the quality of the random effects estimates. In the joint modeling literature there have been numerous simulations studies demonstrating the excellent performance of MLEs in joint models, even in small samples—see, for instance, Tsiatis and Davidian (2004) and references therein. However, regarding the quality of the random-effects estimates little work has been done. From standard mixed models theory it is known that the empirical Bayes estimates exhibit shrinkage (Verbeke and Molenberghs, 2000; Fitzmaurice, Laird, and Ware, 2004), which especially increases as the number of repeated measurements n_i decreases. This motivated us to investigate the finite-sample performance of the two estimators (9) and (11) as a function of the follow-up time t (corresponding to different amounts of available longitudinal information per subject), and the relevant time window Δt. In particular, we simulate under the same sampling setting as in the AIDS data set, and we compare the two estimators at

different follow-up times t corresponding to different amount of information per patient, with the "gold standard" estimator $S_i\{t + \Delta t \mid \mathcal{M}_i(t + \Delta t, b_i)\}/S_i\{t \mid \mathcal{M}_i(t, b_i)\}$ that uses the true parameter and random-effect values. The details of the design of this study are presented in Web Section 2, and the results in Web Figures 6 and 7, and Web Table 3. The two main conclusions drawn from this simulation are the following. First, both estimators performed very well in comparison to the gold standard with negligible average bias. Second, for the same Δt and with increasing t we observed that there is an initial decrease in the bias followed by increase as time progresses. This is attributed to the fact that in the early follow-up period, estimators (9) and (11) are benefited from the additional longitudinal information recorded for the patients (i.e., $\widehat{b}_i$ have less shrinkage), whereas later in the follow-up the precision in affected because the hazard function is less accurately estimated. More detailed discussion of these results can be found in Web Section 2.2.

7. Concluding Remarks

The aim of this article was twofold. Namely, first, to describe how appropriate survival probabilities can be estimated in practice from the output of joint models taking into account the endogenous nature of the longitudinal marker, and second, to assess the discriminative ability of the marker. To estimate conditional survival probabilities we proposed two estimators, estimator (9) that is directly based on the empirical Bayes estimates for the random effects and the maximum likelihood estimates of the parameters, and estimator (11), which is based on a Monte Carlo approach and facilitates the estimation of standard errors and the construction of confidence intervals. To assess the predictive accuracy of the longitudinal marker we defined appropriate sensitivity and specificity measures and derived their estimates under the joint modeling framework. Based on these estimates time-dependent ROCs and AUCs can be utilized to illustrate how the accuracy of the marker evolves dynamically in time. To summarize the information over the whole follow-up period, we proposed the use of the discrimination index $\mathrm{C}_{dyn}^{\Delta t}$ that calculates a weighted average of AUCs. The estimation of these accuracy measures was based on a similar Monte Carlo simulation scheme as for estimator (11).

In our developments we have assumed that the joint model has been correctly specified. Thus, misspecification in any part of the model could affect the derived accuracy estimates. However, for mainly two reasons we feel that such concerns can be alleviated in practice. First, empirical evidence and theoretical work have shown that these models are rather robust against certain types of misspecification (Song et al., 2002; Rizopoulos et al., 2008). Second, we can always further improve the robustness of the joint model presented in Section 2 by extending it and allowing for more flexibility (using splines, kernels, or other approaches) in key components, such as the baseline risk function and the error distribution.

One further aspect in the application of the predictive accuracy measures presented in Section 4.1 that we have not discussed is overfitting, that is, opting for a very elaborate joint model may lead to an overestimation of the predictive performance of the marker. It is important to note that in the joint modeling setting overfitting could be induced in either or both the longitudinal and event time models. To account for potential overfitting and produce corrected versions of the accuracy measures, we can use a bootstrap approach, as described in Harrell, Kerry, and Mark (1996), to estimate the optimism and subtract it from the estimated values of these measures in the original data set. Moreover, in this article we have mainly focused on discrimination rather than calibration. One of the main reasons for this choice is the fact that even if a joint model is not particularly well calibrated, there are some approaches that can be utilized to improve accuracy of predictions without sacrificing discrimination (Harrell et al., 1996). However, this is not to say that calibration is of less importance. In particular, in joint models with latent classes, calibration measures have been considered by Proust-Lima and Taylor (2009), and their extension to the continuous random-effects case is currently under investigation.

All of our proposals have been implemented in `R` with the freely available package `JM` (Rizopoulos, 2010). This package can be used to fit a variety of joint models for longitudinal and time-to-event data with several different options for the baseline risk function. Moreover, both the first-order and Monte Carlo estimates of $\pi_i(u \mid t)$ are calculated using function `survfitJM()`, whereas sensitivity, specificity, and the corresponding ROC and AUC are calculated by function `rocJM()`. The package can be downloaded from the Comprehensive `R` Archive Network accessible via `http://cran.r-project.org/package=JM`, and more information about it can be found at `http://rwiki.sciviews.org/doku.php?id=packages:cran:jm`.

8. Supplementary Materials

Web Sections, Tables, and Figures are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

The author thanks the associate editor and the two anonymous referees for their helpful comments and suggestions that considerably improved the article.

References

Abrams, D., Goldman, A., Launer, C., Korvick, A., Neaton, J., Crane, L., Grodesky, M., Wakefield, S., Muth, K., Kornegay, S., Cohn, D., Harris, A., Luskin-Hawk, R., Markowitz, N., Sampson, J., Thompson, M., Deyton, L. and the Terry Beirn Community Programs for Clinical Research on AIDS. (1994). Comparative trial of didanosine and zalcitabine in patients with human immunodeficiency virus infection who are intolerant of or have failed zidovudine therapy. *New England Journal of Medicine* **330,** 657–662.

Antolini, L., Boracchi, P., and Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Statistics in Medicine* **24,** 3927–3944.

Brown, E., Ibrahim, J., and DeGruttola, V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics* **61,** 64–73.

Cox, D. and Hinkley, D. (1974). *Theoretical Statistics*. London: Chapman & Hall.

Ding, J. and Wang, J.-L. (2008). Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics* **64,** 546–556.

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

Faucett, C. and Thomas, D. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Statistics in Medicine* **15,** 1663–1685.

Fitzmaurice, G., Laird, N., and Ware, J. (2004). *Applied Longitudinal Data*. Hoboken, New Jersey: Wiley.

Garre, F., Zwinderman, A., Geskus, R., and Sijpkens, Y. (2008). A joint latent class changepoint model to improve the prediction of time to graft failure. *Journal of the Royal Statistical Society, Series A* **171,** 299–308.

Goldman, A., Carlin, B., Crane, L., Launer, C., Korvick, J., Deyton, L., and Abrams, D. (1996). Response of CD4+ and clinical consequences to treatment using ddI or ddC in patients with advanced HIV infection. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* **11,** 161–169.

Harrell, F., Callif, R., Pryor, D., Lee, K., and Rosati, R. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association* **247,** 2543–2546.

Harrell, F., Kerry, L., and Mark, D. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15,** 361–387.

Heagerty, P. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61,** 92–105.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1,** 465–480.

Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition. New York: Wiley.

Pencina, M., D'Agostino, Sr., R., D'Agostino, Jr., R., and Vasan, R. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine* **27,** 157–172.

Proust-Lima, C. and 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,** 535–549.

Rizopoulos, D. (2010). `JM`: An `R` package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software* **35** (9), 1–33.

Rizopoulos, D. and Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. To appear in *Statistics in Medicine*.

Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society, Series B* **71,** 637–654.

Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2008). Shared parameter models under random effects misspecification. *Biometrika* **95,** 63–74.

Schemper, M. and Henderson, R. (2000). Predictive accuracy and explained variation in Cox regression. *Biometrics* **56,** 249–255.

Song, X., Davidian, M., and Tsiatis, A. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics* **58,** 742–753.

Taylor, J., Yu, M., and Sandler, H. (2005). Individualized predictions of disease progression following radiation therapy for prostate cancer. *Journal of Clinical Oncology* **23,** 816–825.

Tseng, Y.-K., Hsieh, F., and Wang, J.-L. (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika* **92,** 587–603.

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

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

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

Ye, W., Lin, X., and Taylor, J. (2008). A penalized likelihood approach to joint modeling of longitudinal measurements and time-to-event data. *Statistics and Its Interface* **1,** 33–45.

Yu, M., Law, N., Taylor, J., and Sandler, H. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica* **14,** 835–832.

Yu, M., Taylor, J., and Sandler, H. (2008). Individualized prediction in prostate cancer studies using a joint longitudinal-survival-cure model. *Journal of the American Statistical Association* **103,** 178–187.

Zheng, Y. and Heagerty, P. (2007). Prospective accuracy for longitudinal markers. *Biometrics* **63,** 332–341.

Received February 2010. *Revised October* 2010.
Accepted November 2010.

Web-based Supplementary Material for "Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data"

Dimitris Rizopoulos

Department of Biostatistics, Erasmus Medical Center, PO Box 2040, 3000 CA Rotterdam, the Netherlands

1. Web AIDS Data Set

1.1 *Descriptives*

We present three illustrative descriptive plots for the AIDS data set. In particular, Web Figures 1 and 2 depict the $\text{CD4}^{1/4}$ cell count subject-specific longitudinal trajectories and smooth average longitudinal evolutions, and Web Figure 3 the Kaplan-Meier estimate of the survival functions of the two treatment arms. We observe that both groups of patients show similar variability in their longitudinal profiles. From the Kaplan-Meier estimate in Web Figure 3 it seems that the ddC group has slightly higher survival than the ddI group after the six month of follow-up. From the 467 patients, 188 have died by the end of follow-up corresponding to about 60% censoring. In total we have 1405 recorded CD4 cell count measurements, with an average of three measurements per patient (standard deviation 1.1 measurements).

1.2 *Joint Model Specification & Results*

The joint model fitted to the AIDS data set has the following form:

- Event Process

$$\begin{aligned} h_i(t \mid \mathcal{M}_i(t), w_i) &= h_0(t) \exp\big\{\gamma^\top w_i + \alpha m_i(t)\big\} \\ &= h_0(t) \exp\big\{\gamma_1 \texttt{ddI}_i + \gamma_2 \texttt{failure}_i + \gamma_3 \texttt{prevOI}_i + \\ &\quad \gamma_4 \texttt{male}_i + \alpha m_i(t)\big\}, \end{aligned}$$

email: d.rizopoulos@erasmusmc.nl

where the baseline risk function is assumed piecewise constant:

$$h_0(t) = \sum_{q=1}^{7} \xi_q I(v_{q-1} < t \leq v_q),$$

with the internal knots $v_1 < \cdots < v_6$ placed at equally spaced percentiles of the observed event times, $v_0 = 0$, v_7 is taken larger than the largest observed time, and ξ_q denotes the value of the hazard in the interval $(v_{q-1}, v_q]$, `ddI` is the dummy variable for treatment group ddI, `failure` is the dummy variable for AZT failure, `prevOI` is the dummy variable for previous opportunistic infections, and `male` is the dummy variable for males.

- Longitudinal Process

$$\begin{aligned} y_i(t) &= m_i(t) + \varepsilon_i(t) \\ &= (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t + \beta_2 \texttt{ddI}_i + \beta_3 \texttt{failure}_i + \beta_4 \texttt{prevOI}_i + \\ &\quad \beta_5 \texttt{male}_i + \beta_6 (t \times \texttt{failure}_i) + \beta_7 (t \times \texttt{prevOI}_i) + \\ &\quad \beta_8 (t \times \texttt{male}_i) + \beta_9 (t \times \texttt{ddI}_i) + \varepsilon_i(t), \;\; \varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2), \end{aligned}$$

where $y_i(t)$ denotes the CD4$^{1/4}$ cell count for subject i at time t, $A \times B$ denotes the interaction between A and B, and the random effects $\{b_{i0}, b_{i1}\}$ are assumed bivariate normal with mean 0 and covariance matrix D.

The estimated parameters and standard errors for the joint model fitted to the AIDS data set are presented in Web Table 1.

1.3 *Comparison with Naive Cox Model*

To illustrate the virtues of the joint modelling approach versus a simplistic analysis that uses the baseline CD4$^{1/4}$ cell count measurement in a Cox model (and the same baseline covariates as in the relative risk submodel presented above), we compare the conditional survival probabilities of Patient 68 who is a male belonging to the ddI group, who had showed AZT failure, had a previous opportunistic infection, he provided four CD4 cell count measurements (192, 78, 33, and 11 cells/mm^3) and he was lost to follow-up at 14.17 months, and Patient 185 who is a male belonging to the ddI group, who

had showed AZT failure, had a previous opportunistic infection, he provided three CD4 cell count measurements (0, 60, and 70 cells/mm^3) and he was lost to follow-up at 13.37 months. We note that these two patients exhibit sharply decreasing and sharply increasing CD4 cell count profiles, with considerable difference between the baseline and last CD4 cell counts. For the joint model $\pi(u \mid t)$ was calculated using the procedure described in Section 3, whereas for the Cox model using ratios of the Breslow estimates of the survival function. The results are depicted in Web Figures 4 and 5 for Patients 68 and 185, respectively. We clearly observe the under-performance of the simplistic Cox analysis that does not acknowledge for the changes in the CD4 cell count of the two patients.

2. Web Simulation Study

2.1 *Design*

We have performed a simulation study to investigate the finite sample performance of the two estimators for $\pi_i(u \mid t)$, given by Equations (9) and (11) in the main text. In particular, we compare the two estimators with the 'gold standard' estimate of $\pi_i(u \mid t)$,

$$S_i\big\{u \mid \mathcal{M}_i(u, b_i)\big\} \Big/ S_i\big\{t \mid \mathcal{M}_i(t, b_i)\big\},$$

that is based on the true (i.e., simulated) values for the random effects, and the true value for the parameters. The design of the simulation study is almost entirely motivated by the joint model fitted in the AIDS data set, presented in Web Section 1.2. The only difference is that instead of a piecewise constant baseline risk function we opted for a Weibull baseline risk $h_0(t) = \phi\xi t^{\xi-1}$, with $\phi = 0.067$ and $\xi = 0.762$. The censoring distribution was taken exponential with mean 50, such that we achieve the same percentage of censoring (i.e., about 60%) as for the AIDS data set. Under this setting we simulated 200 data sets.

For each simulated data set we fitted the joint model with exactly the same design as the one we used to simulate from, but assuming a piecewise constant baseline risk function with internal knots placed at equally spaced percentiles of the observed event times. Following for each data set we picked randomly 20 subjects, and for each of them we esti-

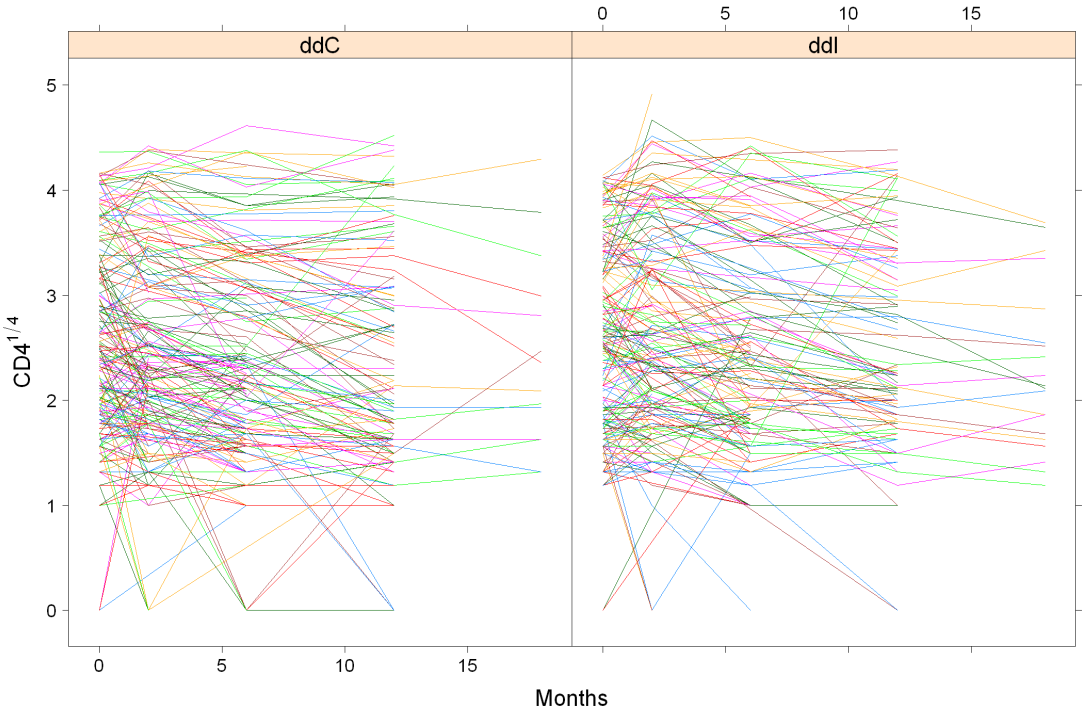
mated $\pi_i(t+\Delta t \mid t)$ using both the Empirical Bayes estimator (9), the Monte Carlo estimator (11), and the gold standard estimator $S_i\{t+\Delta t \mid \mathcal{M}_i(t+\Delta t, b_i)\}/S_i\{t \mid \mathcal{M}_i(t, b_i)\}$, for $t = 2$, 6, 12, and 18, and various values for Δt. The different follow-up times t have been chosen to reflect an increasing number of repeated measurements per subject n_i, which is expected to result in a decreasing degree of shrinkage in the estimation of the random effects. This procedure yielded about 45000 estimates of $\pi_i(u \mid t)$ for each estimator. For each simulated data set and for estimator (11) we used the median $\hat{\pi}_i(u \mid t) = \text{median}\{\pi_i^{(l)}(u \mid t), l = 1, \ldots, L\}$ over $L = 200$ Monte Carlo samples.

2.2 *Results*

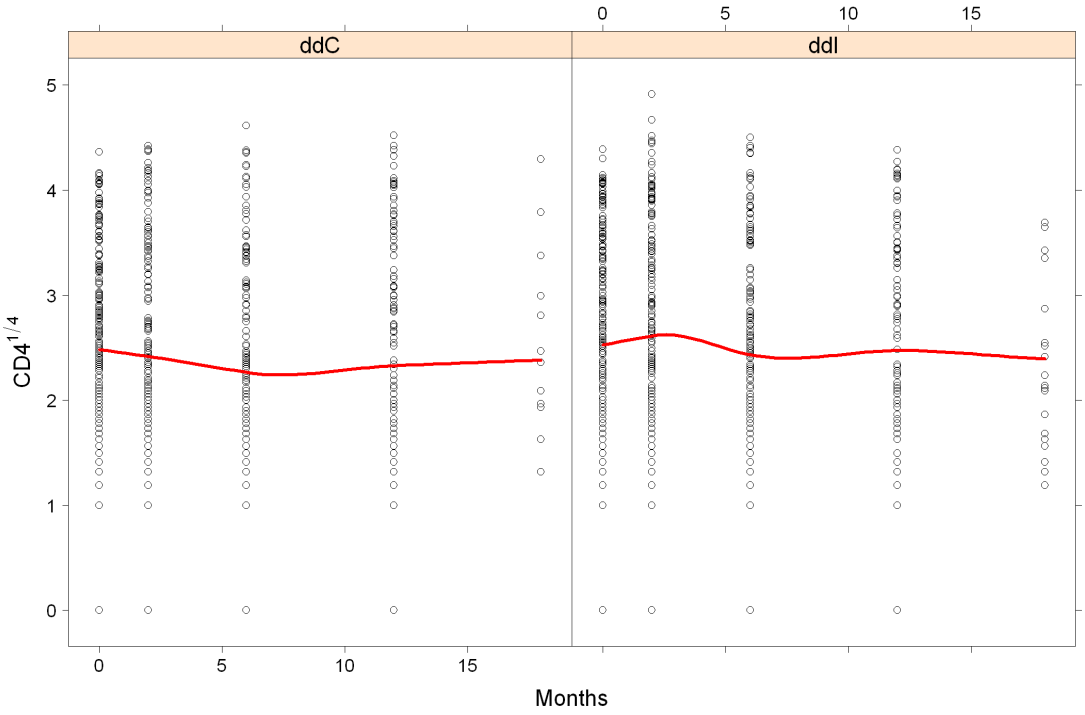
Since we are primarily interested in agreement between the gold standard estimator and the two estimators (9) and (11), we compared the 45000 estimates of $\pi_i(u \mid t)$ using Bland & Altman plots. Moreover, since the survival probabilities are constrained in the $[0, 1]$ interval, we expect more variability around 0.5 while less near the boundaries. Therefore, in order to stabilize the variance we present all results in the $\log\{x/(1-x)\}$ scale. Web Figures 6 and 7 depict the comparison of the two estimators over the whole follow-up period. For each plot the limits of agreement were constructed as plus/minus two times the sample standard deviation of the differences between the two estimators and the gold standard. For $t = 2$, 6, 12, and 18 months, and $\Delta t = 2$, 4, and 8 months, the estimated bias and limits of agreement are presented in Web Table 3.

Before discussing the results we should note that we should always expect some degree of shrinkage in both estimators, since this is an inherent characteristic of random-effects models. Focusing on the whole follow-up period in Web Figures 6 and 7, we observe that for both estimators the average bias is negligible, and that the expected number of comparisons (i.e., about 95%) is located well within the limits of agreement. The performance of the two estimators is almost indistinguishable, with the Monte Carlo estimator performing slightly better (i.e., less estimates outside the limits of agreement). Following, we concentrate on different values for Δt for fixed t. From Web Table 3 it is evident that both the bias and the variability in the differences between both estimators and the gold standard increase analogously with Δt – again, no considerable differences

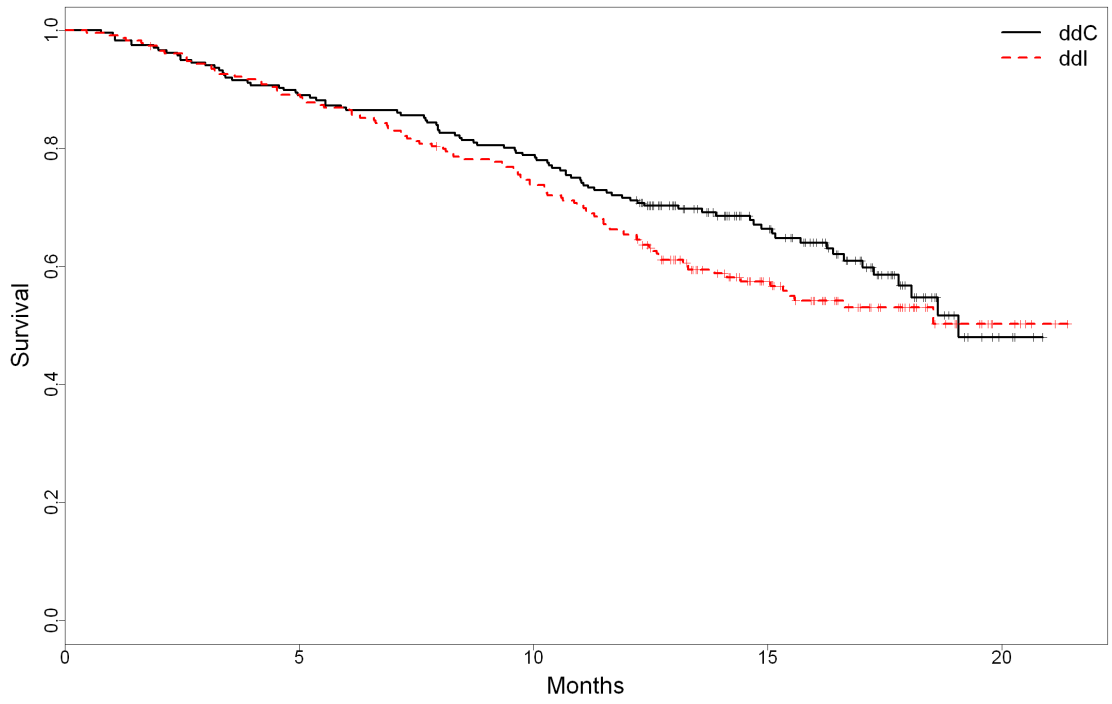
are observed between the behaviour of the two estimators. Taking into account the shape of the survival function, this behaviour is explained by the fact that, small differences between the true random effect values and their estimates are expected to translate to greater differences between the corresponding survival curves as time progresses (i.e., for greater Δt). Finally, we focus on increasing follow-up time t for the same time window Δt. Observing the results from Web Table 3, a counterintuitive relation emerges, namely, the bias decreases from $t = 2$ months to $t = 6$, but then increase again at $t = 12$ to the same levels of $t = 2$, and even more at $t = 18$. To understand why is this happening, we need to note that there is a 'competition' for information between the longitudinal and event time processes. More specifically, as time progresses, we collect more measurements for each subject and therefore we decrease the degree of shrinkage. However, for later stages of the follow-up, and especially for settings in which many events occur early (as in our simulation set-up), there are fewer patients left in the study, which means that the estimated hazard at these time points will have greater variability. The performance of both estimators (9) and (11) versus the gold standard (which uses the true values for both parameters and random effects) will depend on the combined performance of the MLEs and the random-effects estimates, and therefore will be affected when either of the two is less accurately estimated.



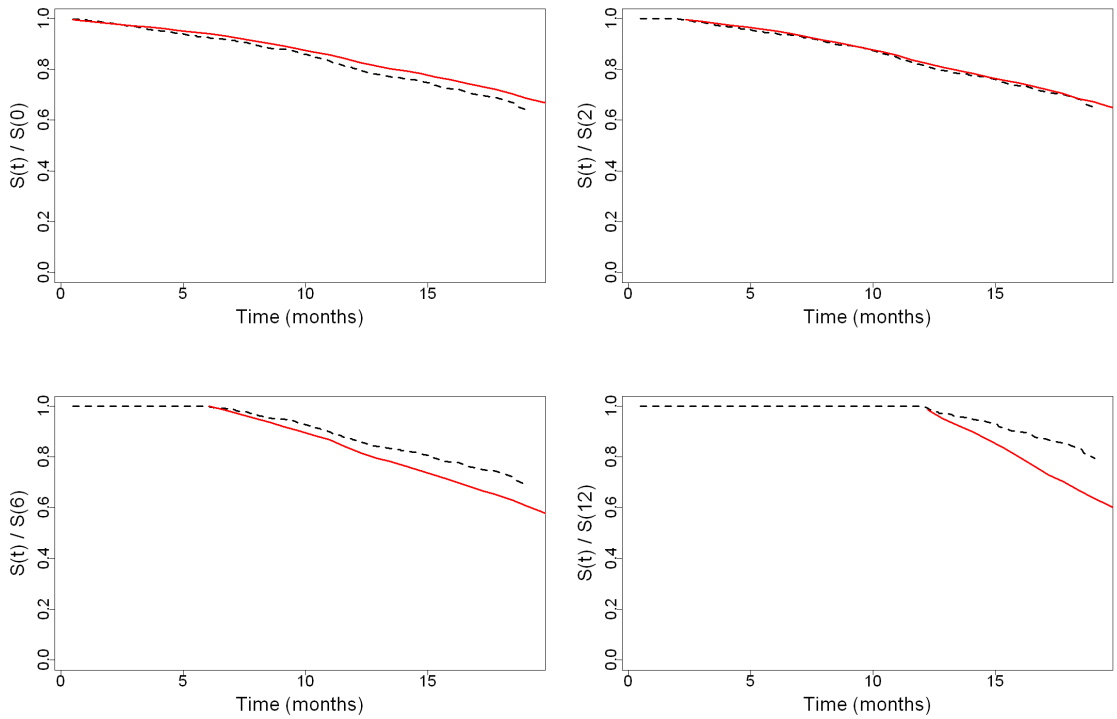
Web Figure 1. Subject-specific longitudinal trajectories for the $\sqrt{\text{CD4}}$ cell count for the ddI and ddC groups, separately.



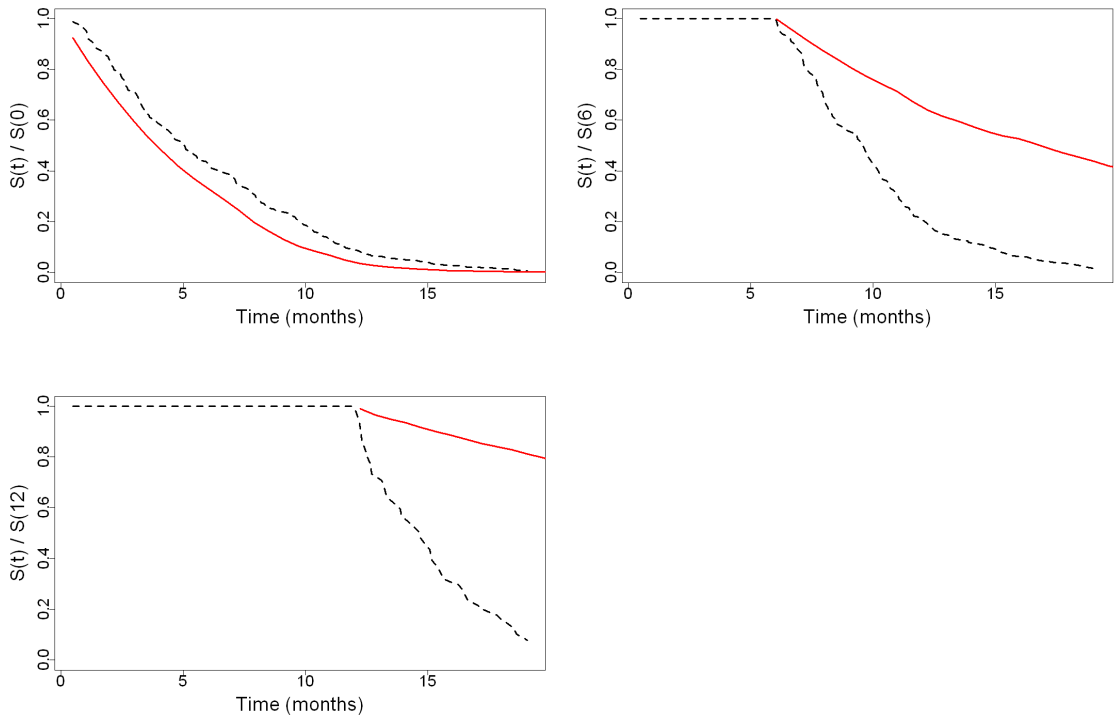
Web Figure 2. Smooth average longitudinal evolutions for the $\sqrt{\text{CD4}}$ cell count for the ddI and ddC groups, separately.



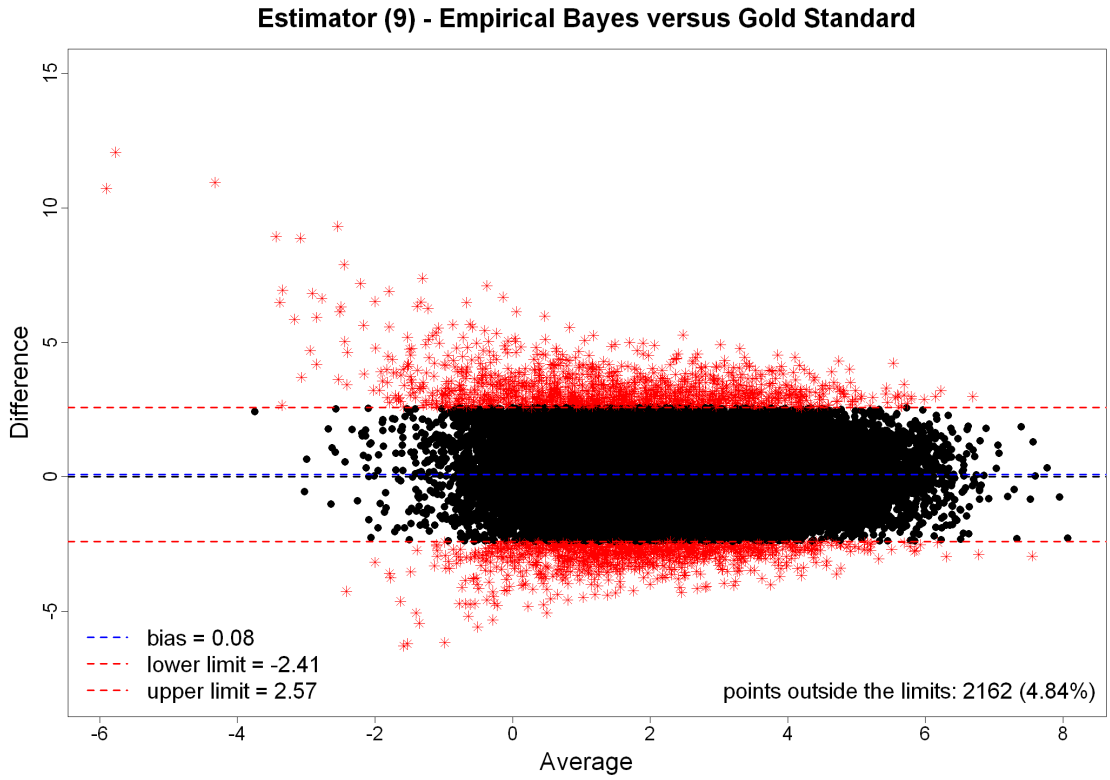
Web Figure 3. Kaplan-Meier estimates of the survival functions of the ddI and ddC treatment groups.



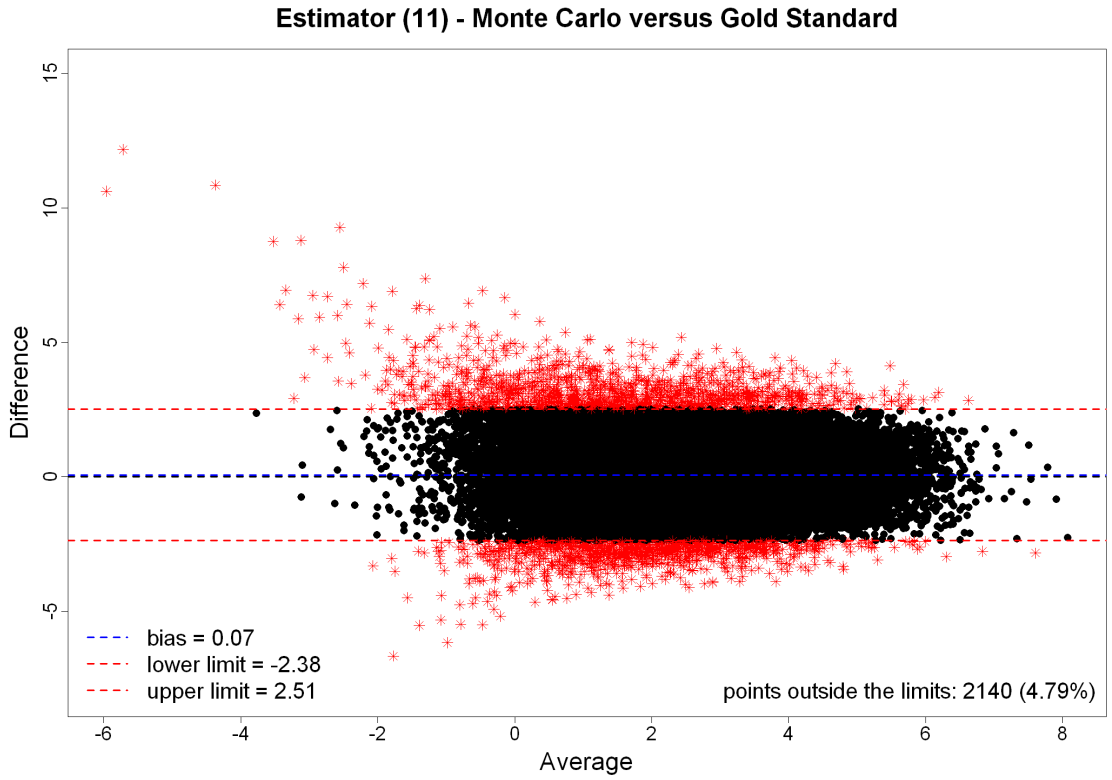
Web Figure 4. Conditional survival probabilities from the Cox (dashed line) and joint model (solid line) for Patient 68.



Web Figure 5. Conditional survival probabilities from the Cox (dashed line) and joint model (solid line) for Patient 185.



Web Figure 6. Bland-Altman plot for the comparison of the first order estimator of $\pi_i(u \mid t)$ with the gold standard estimator over the whole follow-up period. Simulation results on the logit scale based on 200 data sets. The blue dashed line denotes the average bias. The red dashed lines denote the limits of agreement.



Web Figure 7. Bland-Altman plot for the comparison of the Monte Carlo estimator of $\pi_i(u \mid t)$ with the gold standard estimator over the whole follow-up period. Simulation results on the logit scale based on 200 data sets. The blue dashed line denotes the average bias. The red dashed lines denote the limits of agreement.

Web Table 1

Parameter estimates and standard errors for the joint model fitted to the AIDS data set. The parameters D_{11}, D_{22}, and D_{12} correspond to the variance of the random intercepts b_{i0}, the variance of the random slopes b_{i1}, and their covariance, respectively.

Event Process			Longitudinal Process		
	est.	std. err.		est.	std. err.
ddI	0.3450	0.1526	intercept	2.9396	0.1145
failure	0.1083	0.1689	time	−0.0323	0.0122
prevOI	0.6316	0.2409	ddI	0.0277	0.0727
male	−0.3524	0.2555	failure	−0.0309	0.0852
α	−0.9769	0.1263	prevOI	−0.8758	0.0818
$\log(\xi_1)$	−2.1345	0.4182	male	0.1612	0.1215
$\log(\xi_2)$	−1.7771	0.4086	time × failure	0.0009	0.0085
$\log(\xi_3)$	−1.4316	0.4339	time × prevOI	−0.0107	0.0080
$\log(\xi_4)$	−1.9657	0.4940	time × male	−0.0029	0.0121
$\log(\xi_5)$	−1.8453	0.4737	time × ddI	0.0050	0.0064
$\log(\xi_6)$	−1.7844	0.5310	σ	0.3711	0.0384
$\log(\xi_7)$	−1.7766	0.6253			
			Random Effects		
			D_{11}	0.5926	0.0449
			D_{12}	−0.0007	0.0023
			D_{22}	0.0013	0.0003

Web Table 2

Parameter estimates and standard errors for the joint model fitted to the AIDS data set, under the time-dependent value + derivative parameterization. The parameters D_{11}, D_{22}, and D_{12} correspond to the variance of the random intercepts b_{i0}, the variance of the random slopes b_{i1}, and their covariance, respectively.

Event Process			Longitudinal Process		
	est.	std. err.		est.	std. err.
ddI	0.3527	0.1563	intercept	2.9396	0.1136
failure	0.1172	0.1736	time	−0.0327	0.0122
prevOI	0.6188	0.2440	ddI	0.0257	0.0726
male	−0.3724	0.2639	failure	−0.0352	0.0852
α	−0.9769	0.1275	prevOI	−0.8732	0.0822
α_d	−1.7600	4.5497	male	0.1660	0.1201
$\log(\xi_1)$	−2.1939	0.4656	time × failure	0.0008	0.0085
$\log(\xi_2)$	−1.8441	0.4637	time × prevOI	−0.0113	0.0081
$\log(\xi_3)$	−1.4965	0.4833	time × male	−0.0027	0.0121
$\log(\xi_4)$	−2.0288	0.5367	time × ddI	0.0047	0.0064
$\log(\xi_5)$	−1.9049	0.5142	σ	0.3710	0.0384
$\log(\xi_6)$	−1.8357	0.5616			
$\log(\xi_7)$	−1.8177	0.6462	Random Effects		
			D_{11}	0.5926	0.0449
			D_{12}	−0.0005	0.0027
			D_{22}	0.0013	0.0003

Web Table 3

Bias and lower and upper limits of agreement in the Bland-Altman plots for $t = 2$, 6, 12 and 18 months, and $\Delta t = 2$, 4 and 8 months. The top results correspond to the Empirical Bayes estimator, and the bottom ones to the Monte Carlo estimator. Simulation results on the logit scale based on 200 data sets.

	$\Delta t = 2$	$\Delta t = 4$	$\Delta t = 8$
	bias (low; upp)	bias (low; upp)	bias (low; upp)
$t = 2$	0.061 (−2.02; 2.15)	0.053 (−2.07; 2.18)	0.033 (−2.18; 2.25)
$t = 6$	0.001 (−2.15; 2.15)	−0.008 (−2.21; 2.19)	−0.029 (−2.34; 2.28)
$t = 12$	−0.052 (−2.27; 2.17)	−0.062 (−2.35; 2.23)	−0.086 (−2.55; 2.38)
$t = 18$	−0.280 (−2.71; 2.15)	−0.298 (−2.82; 2.23)	−0.341 (−3.12; 2.44)
$t = 2$	0.058 (−2.00; 2.11)	0.058 (−2.03; 2.15)	0.042 (−2.14; 2.22)
$t = 6$	0.001 (−2.11; 2.11)	−0.006 (−2.16; 2.15)	−0.016 (−2.28; 2.25)
$t = 12$	−0.054 (−2.23; 2.13)	−0.064 (−2.31; 2.18)	−0.080 (−2.49; 2.33)
$t = 18$	−0.272 (−2.66; 2.12)	−0.289 (−2.77; 2.19)	−0.329 (−3.06; 2.41)