Dynamic Predictions with Time-Dependent Covariates in Survival Analysis using Joint Modeling and Landmarking

Dimitris Rizopoulos[1,*], Geert Molenberghs[2] and Emmanuel M.E.H. Lesaffre[2,1]

[1] Department of Biostatistics, Erasmus Medical Center, the Netherlands

[2] Interuniversity Institute for Biostatistics and statistical Bioinformatics, KU Leuven & Universiteit Hasselt, Belgium

Abstract

A key question in clinical practice is accurate prediction of patient prognosis. To this end, nowadays, physicians have at their disposal a variety of tests and biomarkers to aid them in optimizing medical care. These tests are often performed on a regular basis in order to closely follow the progression of the disease. In this setting it is of interest to optimally utilize the recorded information and provide medically relevant summary measures, such as survival probabilities, that will aid in decision making. In this work we present and compare two statistical techniques that provide dynamically-updated estimates of survival probabilities, namely landmark analysis and joint models for longitudinal and time-to-event data. Special attention is given to the functional form linking the longitudinal and event time processes, and to measures of discrimination and calibration in the context of dynamic prediction.

Keywords: Calibration; Discrimination; Prognostic Modeling; Risk Prediction; Random Effects.

*Correspondance to: Department of Biostatistics, Erasmus University Medical Center, PO Box 2040, 3000 CA Rotterdam, the Netherlands. E-mail address: d.rizopoulos@erasmusmc.nl.

1 Introduction

Nowadays there is great interest in accurate risk assessment for prevention and treatment of disease. Physicians use risk scores to reach appropriate decisions, such as prescribing treatment, or extra medical tests or suggesting alternative therapies. Patients who are informed about their health risk often decide to adjust their lifestyles to mitigate it. Risk scores are typically based on several factors that describe the patients' physical condition and risk factors, such as age, sex, BMI, smoking, genetic predisposition, and the results of medical tests. In this work we focus on the use of the results of such tests and more specifically on biomarkers. The majority of prognostic models in the medical literature utilize only a small fraction of the available biomarker information. In particular, even though biomarkers are measured repeatedly over time, risk scores are typically based on the last available biomarker measurement. It is evident that such an approach discards valuable information because it does not take into account that the rate of change in the biomarker levels is not only different from patient to patient but also dynamically changes over time for the same patient. Hence, it is medically relevant to investigate whether repeated measurements of a biomarker can provide a better understanding of disease progression and a better prediction of the risk for the event of interest than a single biomarker measurement.

In line with the previous arguments, the motivation for this research comes from a study conducted by the Department of Cardio-Thoracic Surgery of the Erasmus Medical Center in the Netherlands. This study includes 285 patients who received a human tissue valve in the aortic position in the hospital from 1987 until 2008 (Bekkers et al. 2011). Aortic allograft implantation has been widely used for a variety of aortic valve or aortic root diseases. Major advantages ascribed to allografts are the excellent hemodynamic characteristics as a valve substitute; the low rate of thrombo-embolic complications, and, therefore, absence of the need for anticoagulant treatment; and the resistance to endocarditis. A major disadvantage of using human tissue valves, however is the susceptibility to degeneration and the concomitant need for re-interventions. The durability of a cryopreserved aortic allograft is age-dependent, leading to a high lifetime risk of re-operation, especially for young patients.

Re-operations on the aortic root are complex, with substantial operative risks, and mortality rates in the range 4–12%. It is therefore of great interest for cardiologists and cardio-thoracic surgeons to have at their disposal an accurate prognostic tool that will inform them about the future prospect of a patient with a human tissue valve in order to optimize medical care, carefully plan re-operation and minimize valve-related morbidity and mortality.

From the statistical analysis viewpoint the challenge is to utilize a technique capable of updating estimates of survival probabilities for a new patient as additional longitudinal information is recorded. An early approach in solving this problem has been landmarking (Anderson et al. 1983; van Houwelingen 2007; van Houwelingen and Putter 2011). The basic idea behind landmarking is to obtain survival probabilities from a Cox model fitted to the patients from the original dataset who are still at risk at the time point of interest (e.g., the last time point we know that the new patient was still alive). A relatively newer method for producing dynamic predictions of survival probabilities is based on the class of joint models for longitudinal and time-to-event data (Yu et al. 2008; Proust-Lima and Taylor 2009; Rizopoulos 2011, 2012). In these models we have a complete specification of the joint distribution of the longitudinal response and the event times based on which the predictions in question can be derived. The main aim of this paper is to further study and contrast these two approaches. In particular, we show how survival probabilities are obtained under each method and what the differences are in the underlying assumptions. In addition, we focus on the functional relationship between the longitudinal and event time processes, and how this may affect predictions. To assess the quality of the derived predictions from the two approaches we present different measures of discrimination and calibration, suitably adjusted to the context of longitudinal biomarkers.

The rest of the paper is organized as follows. Section 2 formally describes the context of dynamic predictions and presents the landmarking and joint modeling approaches. Section 3 presents measures of discrimination and calibration adapted to the dynamic predictions setting. Section 4 illustrates the use of joint modeling and landmarking in the Aortic Valve dataset and Section 5 refers to the results of a simulation study. Finally, Section 6 concludes

the paper.

2 Dynamic Individualized Predictions

Following the discussion in Section 1 and the motivation from the Aortic Valve dataset, we present here the two frameworks for deriving dynamic individualized predictions. Let $\mathcal{D}_n = \{T_i, \delta_i, \boldsymbol{y}_i; i = 1, \ldots, n\}$ denote a sample from the target population, where $T_i = \min(T_i^*, C_i)$ denotes the observed event time for the i-th subject ($i = 1, \ldots, n$), with T_i^* denoting the true event time, C_i the censoring time, and $\delta_i = I(T_i^* \leq C_i)$ the event indicator, with $I(\cdot)$ being the indicator function that takes the value 1 when $T_i^* \leq C_i$, and 0 otherwise. In addition, we let $\boldsymbol{y}_i$ denote the $n_i \times 1$ longitudinal response vector for the i-th subject, with element y_{il} denoting the value of the longitudinal outcome taken at time point t_{il}, $l = 1, \ldots, n_i$.

We are interested in deriving predictions for a new subject j from the same population that has provided a set of longitudinal measurements $\mathcal{Y}_j(t) = \{y_j(t_{jl}); 0 \leq t_{jl} \leq t, l = 1, \ldots, n_j\}$. The fact that biomarker measurements have been recorded up to t, implies survival of this subject up to this time point, meaning that it is more relevant to focus on the conditional subject-specific predictions, given survival up to t. In particular, for any time $u > t$ we are interested in the probability that this new subject j will survive at least up to u, i.e.,

$$\pi_j(u \mid t) = \Pr(T_j^* \geq u \mid T_j^* > t, \mathcal{Y}_j(t), \mathcal{D}_n).$$

The time-dynamic nature of $\pi_j(u \mid t)$ is evident because when new information is recorded for patient j at time $t' > t$, we can update these predictions to obtain $\pi_j(u \mid t')$, and therefore proceed in a time-dynamic manner.

2.1 Joint Modeling

In the framework of joint models for longitudinal and time-to-event data we have a complete specification of the joint distribution of the two outcomes (Henderson et al. 2000; Tsiatis and Davidian 2004; Rizopoulos 2012). For the longitudinal biomarker measurements

mixed-effects models are typically employed to describe the subject-specific longitudinal trajectories. For simplicity of exposition and because the marker that we are going to use for the Aortic Valve dataset, namely the aortic gradient, is a continuous one, we focus here on linear mixed-effects models,

$$\begin{aligned} y_i(t) &= m_i(t) + \varepsilon_i(t) = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_i + \varepsilon_i(t), \\ \boldsymbol{b}_i &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}), \quad \varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2), \end{aligned} \tag{1}$$

where $y_i(t)$ denotes the observed value of the longitudinal outcome at any particular time point t, $\boldsymbol{x}_i(t)$ and $\boldsymbol{z}_i(t)$ denote the design vectors for the fixed-effects $\boldsymbol{\beta}$ and for the random effects $\boldsymbol{b}_i$, respectively, and $\varepsilon_i(t)$ the corresponding error terms that are assumed independent of the random effects, and $\text{cov}\{\varepsilon_i(t), \varepsilon_i(t')\} = 0$ for $t' \neq t$. The fixed and random effects design vectors are assumed to contain a mix of baseline and time-varying covariates. For the survival process, we assume that the risk for an event depends on the 'true' and unobserved value of the marker at time t (i.e., excluding the measurement error), denoted by $m_i(t)$ in (1). More specifically, we have

$$\begin{aligned} h_i(t \mid \mathcal{M}_i(t), \boldsymbol{w}_i) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr\{t \leq T_i^* < t + \Delta t \mid T_i^* \geq t, \mathcal{M}_i(t), \boldsymbol{w}_i\} \\ &= h_0(t) \exp\big\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \alpha m_i(t)\big\}, \quad t > 0, \end{aligned} \tag{2}$$

where $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$ denotes the history of the true unobserved longitudinal process up to t, $h_0(\cdot)$ denotes the baseline hazard function, and $\boldsymbol{w}_i$ is a vector of baseline covariates with corresponding regression coefficients $\boldsymbol{\gamma}$. Parameter α quantifies the association between the true value of the marker at t and the hazard for an event at the same time point. Estimation of the joint model's parameters can be based either on maximum likelihood or a Bayesian approach using Markov chain Monte Carlo algorithms. The likelihood of the model is derived under the conditional independence assumptions that given the random effects, both the longitudinal and event time process are assumed independent, and the longitudinal responses of each subject are assumed independent. More details regarding the estimation

and properties of joint models can be found in Rizopoulos (2012) and Ibrahim et al. (2001, Chapter 7).

Under this framework, estimation of $\pi_j(u \mid t)$ can be based on (asymptotic) Bayesian arguments and the corresponding posterior predictive distribution:

$$\pi_j^{JM}(u \mid t) = \int \Pr(T_j^* \geq u \mid T_j^* > t, \mathcal{Y}_j(t), \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D}_n)\, d\boldsymbol{\theta}, \tag{3}$$

where $\boldsymbol{\theta}$ denotes the vector of all model parameters. The calculation of the first term of the integrand utilizes the aforementioned conditional independence assumptions. In particular, the first term of the integrand of $\pi_j^{JM}(u \mid t)$ can be written as:

$$\begin{aligned}
\Pr(T_j^* \geq u \mid T_j^* > t, \mathcal{Y}_j(t), \boldsymbol{\theta}) &= \int \Pr(T_j^* \geq u \mid T_j^* > t, \boldsymbol{b}_j, \boldsymbol{\theta})\, p(\boldsymbol{b}_j \mid T_j^* > t, \mathcal{Y}_j(t), \boldsymbol{\theta})\, d\boldsymbol{b}_j \\
&= \int \frac{S_j\big\{u \mid \mathcal{M}_j(u, \boldsymbol{b}_j), \boldsymbol{\theta}\big\}}{S_j\big\{t \mid \mathcal{M}_j(t, \boldsymbol{b}_j), \boldsymbol{\theta}\big\}}\, p(\boldsymbol{b}_j \mid T_j^* > t, \mathcal{Y}_j(t), \boldsymbol{\theta})\, d\boldsymbol{b}_j,
\end{aligned}$$

where

$$S_j\big\{t \mid \mathcal{M}_j(t, \boldsymbol{b}_j), \boldsymbol{\theta}\big\} = \exp\left\{ \int_0^t h_0(s) \exp\{\boldsymbol{\gamma}^\top \boldsymbol{w}_j + \alpha m_j(s)\} \right\} ds,$$

denotes the subject-specific survival function.

Combining these equations with the maximum likelihood estimates or with the MCMC sample from the posterior distribution of the parameters for the original data $\mathcal{D}_n$, we can devise a simple simulation scheme to obtain a Monte Carlo estimate of $\pi_j(u \mid t)$. More details can be found in Yu et al. (2008) and Rizopoulos (2011, 2012).

2.2 Standard Landmarking

The landmarking approach provides an estimate of $\pi_j(u \mid t)$ by selecting the subjects at risk at t from the original dataset $\mathcal{D}_n$, and using these to derive predictions. More formally, let $\mathcal{R}(t) = \{i : T_i > t\}$ denote the risk set, including all subjects who were not censored or dead by the landmark time t. Then a Cox model is fitted to these subjects by resetting time to

zero being the landmark time, i.e.,

$$\begin{aligned} h_i(u) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr\big\{u \leq T_i^* < u + \Delta t \mid T_i^* \geq u, \mathcal{Y}_i(t)\big\} \\ &= h_0(u) \exp\big\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \alpha \tilde{y}_i(t)\big\}, u > t, \end{aligned}$$

where the baseline hazard function $h_0(\cdot)$ is assumed completely unspecified, $\boldsymbol{w}_i$ denotes a vector of baseline covariates, and $\tilde{y}_i(t)$ denotes the last available longitudinal response that also enters into the model as an ordinary baseline covariate. Having fitted this Cox model, an estimate of $\pi_j(u \mid t)$ is simply obtained by means of the Breslow estimator:

$$\hat{\pi}_j^{LM}(u \mid t) = \exp\Big[-\widehat{H}_0(u) \exp\{\widehat{\boldsymbol{\gamma}}^\top \boldsymbol{w}_j + \hat{\alpha} \tilde{y}_j(t)\}\Big], \tag{4}$$

where

$$\widehat{H}_0(u) = \sum_{i \in \mathcal{R}(t)} \frac{I(T_i \leq u)\delta_i}{\sum_{\ell \in \mathcal{R}(u)} \exp\{\widehat{\boldsymbol{\gamma}}^\top \boldsymbol{w}_\ell + \hat{\alpha} \tilde{y}_\ell(t)\}}.$$

van Houwelingen (2007) and Zheng and Heagerty (2005) discuss several extensions of this approach that have greater flexibility by allowing the regression coefficient α to depend on time, and also, allowing the baseline hazard to be not only a function of the time since the last measurement $u - t$, but also a function of the measurement time t, relaxing thus the proportional hazards assumption.

2.3 Mixed Model Landmarking

Compared to the joint modeling approach, the standard landmark approach does not account for the fact that the longitudinal outcome that is included as a time-varying covariate in the Cox model contains measurement error. In addition, by only including the last available longitudinal measurement in the Cox model, the landmark approach essentially does not utilize the measurements recorded earlier. Furthermore, the standard landmark approach assumes that the event risk depends on the last observed marker value, neglecting the impact of the time elapsed between the longitudinal measurements and the landmark time. To

account for these issues, we could make a compromise between the joint modeling approach and landmarking. In particular, for any landmark time t, we could fit a mixed-effects model of the form (1) in the adjusted risk set $R(t)$. Under the fitted mixed model, we subsequently derive the empirical Bayes estimates of the random effects of the subjects at risk, namely,

$$\hat{\boldsymbol{b}}_i = \arg\max_b \Big\{ \sum_l \log p(y_i(t_{il}) \mid \boldsymbol{b}; \hat{\boldsymbol{\theta}}) + \log p(\boldsymbol{b}; \hat{\boldsymbol{\theta}}) \Big\},$$

where $\hat{\boldsymbol{\theta}}$ denotes the vector of maximum likelihood estimates from the mixed model. Based on these empirical Bayes estimates and $\hat{\boldsymbol{\theta}}$ we can obtain an estimate of $m_i(t)$ denoted by $\hat{m}_i(t)$. Following, in the landmark dataset we fit the Cox model

$$h_i(u) = h_0(u) \exp\big\{ \boldsymbol{\gamma}^\top \boldsymbol{w}_i + \alpha \hat{m}_i(t) \big\}, u > t.$$

Using a similar calculation, we can derive the empirical Bayes estimates of the random effects for subject j for whom we want to calculate predictions. Finally, following the same procedure as in the standard landmark approach, we combine the derived term $\hat{m}_j(t)$ with the Breslow estimator of the cumulative baseline hazard and the estimated regression coefficients from the Cox model to obtain the predictions:

$$\hat{\pi}_j^{LMmixed}(u \mid t) = \exp\Big[-\widehat{H}_0(u) \exp\{ \widehat{\boldsymbol{\gamma}}^\top \boldsymbol{w}_j + \hat{\alpha} \hat{m}_j(t) \} \Big]. \tag{5}$$

2.4 Heuristic Comparison between Landmarking and Joint Modeling

The previous two sections illustrated that both landmarking and joint modeling can be utilized to derive dynamically updated estimates of conditional survival probabilities $\pi_j(u \mid t)$. The landmark approach can be more easily implemented in practice because it only requires fitting a standard Cox model, whereas joint models require specialized software (Rizopoulos 2010, 2012). In addition, joint models seem to make more modeling assumptions than the

landmark approach, which poses a concern regarding how misspecification of these assumptions may affect predictions. On the other hand, the landmark approach uses less information than joint modeling (i.e., only the subjects at risk at the landmark time), and hence may be less efficient.

Two key concepts that relate to the prediction problem we investigate here are first, the nature of the longitudinal outcome when seen as time-varying covariate for the survival model, and, second, the consistency of the predictions we obtain. With respect to the former concept, the exo-/endogeneity of the longitudinal outcome can be discerned using the definition provided by Kalbfleisch and Prentice (2002, Section 6.3). More specifically, a covariate process is defined exogenous when the following condition is satisfied:

$$\Pr\big\{s \leq T_i^* < s + ds \mid T_i^* \geq s, \mathcal{Y}_i(s)\big\} = \Pr\big\{s \leq T_i^* < s + ds \mid T_i^* \geq s, \mathcal{Y}_i(t)\big\},$$

for all s, t such that $0 < s \leq t$, and $ds \to 0$. This definition implies that $y_i(\cdot)$ is associated with the rate of failures over time, but its future path up to any time $t > s$ is not affected by the occurrence of failure at time s. This will hold for time-varying covariates, such as environmental factors or the period of the year (e.g., summer or winter), but it will not hold for the most often used time-varying covariates, such as biomarkers or patient parameters. With respect to landmarking and joint modeling, the fact that the landmark approach does not specify the joint distribution of the longitudinal outcome and the event process implies that it will only be valid for exogenous covariates. On the other hand, the joint modeling approach directly translates to an endogenous time-varying covariate. To see this we study the posterior distribution of the random effects under a joint model:

$$p(\boldsymbol{b}_i \mid \mathcal{Y}_i(t), T_i, \delta_i) \propto p(\mathcal{Y}_i(t) \mid \boldsymbol{b}_i)\{h_i(T_i \mid \boldsymbol{b}_i)^{\delta_i} S_i(T_i \mid \boldsymbol{b}_i)\}p(\boldsymbol{b}_i),$$

from which we see that occurrence of an event (i.e., $\delta_i = 1$) alters the random effects estimates we will obtain, and as a result this will also alter the estimated future path of the longitudinal outcome. As an illustration of this issue, we show in Figure 1 the estimates of the linear

random slopes from a linear mixed model and a joint model fitted to the Aortic Valve dataset.

[Figure 1 about here.]

From this figure it is evident that the joint model returns larger random slopes estimates for patients with an event than the mixed model, and analogously, smaller random slopes estimates for patients who were event-free.

The second concept has been formalized by Jewell and Nielsen (1993). In particular, they defined that a valid prediction function must satisfy the consistency condition:

$$h_i(t+s \mid \mathcal{Y}_i(t)) = E\{h_i(0 \mid \mathcal{Y}_i(t+s)) \mid \mathcal{Y}_i(t)\},$$

where the expectation is taken with respect to the sample path of the longitudinal outcome from t to $t+s$. This condition implies that prediction at time $t+s$ should be obtainable from the prediction we have available at t by integrating out over the probability distribution of the longitudinal outcome in the interval $(t, t+s)$. As Jewell and Nielsen (1993) explain, to satisfy the consistency condition, a prediction function needs to be derived from the joint distribution of the longitudinal and time-to-event outcomes. Hence, both the standard landmark approach and the mixed model landmark approach will not satisfy this condition, because, once again, they do not model the joint distribution of the longitudinal and event time processes. On the other hand, a correctly specified joint model, will provide consistent predictions.

3 Measuring Predictive Performance

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 focused on calibration, i.e., how well the model predicts the observed data (Schemper and Henderson 2000; Gerds and Schumacher 2006) or discrimination, i.e., how well can the model discriminate between patients that had the event from patients that did not (Harrell et al. 1996;

Pencina et al. 2008). In the following we present discrimination and calibration measures suitably adapted to the dynamic prediction setting. It should be noted that these measures require in essence an estimate of $\pi_j(u \mid t)$, and therefore they are applicable under both landmarking and joint modeling. In the following we will use the term $\hat{\pi}_j(u \mid t)$ to generically denote either the Monte Carlo estimate of (3), (4) or (5).

Standard approaches for estimating these measures in the presence of censoring are based on inverse probability weighted estimators using Kaplan-Meier-type nonparametric estimators for the censoring distribution (Uno et al. 2007; Parast et al. 2012; Parast and Cai 2013). For joint models such estimators have been developed by Blanche et al. (2015). However, in the context we consider here, it may very often be the case that the probability of censoring depends on the observed longitudinal responses (e.g., a doctor decides to exclude a patient from the study based on his biomarker values). In addition, most often in clinical practice, censoring may also depend on other baseline covariates. In these cases, the aforementioned estimators may produce biased estimates of predictive accuracy because they do not account for these features. Hence, the predictive accuracy measures we present below correct for censoring using model-based estimators of the censoring distribution. To account for misspecification of the fitted model, we suggest that cross-validated versions of these measures are considered in practice.

3.1 Discrimination

To take into account the dynamic nature of the longitudinal marker in discriminating between subjects, we focus on a time interval of medical relevance within which the occurrence of events is of interest. In this setting, a useful property of the model would be to successfully discriminate between patients who are going to experience the event within this time frame from patients who will not. To put this formally, as before, we assume that we have collected longitudinal measurements $\mathcal{Y}_j(t) = \{y_j(t_{jl}); 0 \leq t_{jl} \leq t, l = 1, \ldots, n_j\}$ up to time point t for subject j. 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. Under the assumed model and the methodology presented in Section 2, we can define a prediction rule using $\pi_j(t + \Delta t \mid t)$ that takes into account the available longitudinal measurements $\mathcal{Y}_j(t)$. In particular, for any value c in $[0, 1]$ we can term subject j as a case if $\pi_j(t + \Delta t \mid t) \leq c$ (i.e., occurrence of the event) and analogously as a control if $\pi_j(t+\Delta t \mid t) > c$. For a randomly chosen pair of subjects $\{i, j\}$, in which both subjects have provided measurements up to time t, the discriminative capability of the assumed model can be assessed by the area under the receiver operating characteristic curve (AUC), which is obtained for varying c and equals:

$$\mathrm{AUC}(t, \Delta t) = \Pr\big[\pi_i(t + \Delta t \mid t) < \pi_j(t + \Delta t \mid t) \mid \{T_i^* \in (t, t + \Delta t]\} \cap \{T_j^* > t + \Delta t\}\big],$$

that is, if subject i experiences the event within the relevant time frame whereas subject j does not, then we would expect the assumed model to assign higher probability of surviving longer than $t + \Delta t$ for the subject who did not experience the event.

Estimation of $\mathrm{AUC}(t, \Delta t)$ is directly based on its definition, namely by appropriately counting the concordant pairs of subjects. More specifically, we have the decomposition:

$$\widehat{\mathrm{AUC}}(t, \Delta t) = \widehat{\mathrm{AUC}}_1(t, \Delta t) + \widehat{\mathrm{AUC}}_2(t, \Delta t) + \widehat{\mathrm{AUC}}_3(t, \Delta t) + \widehat{\mathrm{AUC}}_4(t, \Delta t).$$

The first term $\widehat{\mathrm{AUC}}_1(t, \Delta t)$ refers to the pairs of subjects who are comparable (i.e., their observed event times can be ordered),

$$\Omega_{ij}^{(1)}(t) \;\; = \;\; \big[\{T_i \in (t, t + \Delta t]\} \cap \{\delta_i = 1\}\big] \cap \{T_j > t + \Delta t\},$$

where $i, j = 1, \ldots, n$ with $i \neq j$. For such comparable subjects i and j, we can estimate and compare their survival probabilities $\pi_i(t + \Delta t \mid t)$ and $\pi_j(t + \Delta t \mid t)$, based on the methodology presented in Section 2. This leads to a natural estimator for $\mathrm{AUC}_1(t, \Delta t)$ as

the proportion of concordant subjects out of the set of comparable subjects at time t:

$$\widehat{\mathrm{AUC}}_1(t, \Delta t) = \frac{\sum_{i=1}^{n} \sum_{j=1; j \neq i}^{n} I\{\hat{\pi}_i(t + \Delta t \mid t) < \hat{\pi}_j(t + \Delta t \mid t)\} \times I\{\Omega_{ij}^{(1)}(t)\}}{\sum_{i=1}^{n} \sum_{j=1; j \neq i}^{n} I\{\Omega_{ij}^{(1)}(t)\}},$$

where $I(\cdot)$ denotes the indicator function. Analogously, $\widehat{\mathrm{AUC}}_2(t, \Delta t)$, $\widehat{\mathrm{AUC}}_3(t, \Delta t)$ and $\widehat{\mathrm{AUC}}_4(t, \Delta t)$ refer to the pairs of subjects who due to censoring cannot be compared, namely

$$\begin{aligned}
\Omega_{ij}^{(2)}(t) &= \big[\{T_i \in (t, t + \Delta t]\} \cap \{\delta_i = 0\}\big] \cap \{T_j > t + \Delta t\}, \\
\Omega_{ij}^{(3)}(t) &= \big[\{T_i \in (t, t + \Delta t]\} \cap \{\delta_i = 1\}\big] \cap \big[\{T_i < T_j \leq t + \Delta t\} \cap \{\delta_j = 0\}\big], \\
\Omega_{ij}^{(4)}(t) &= \big[\{T_i \in (t, t + \Delta t]\} \cap \{\delta_i = 0\}\big] \cap \big[\{T_i < T_j \leq t + \Delta t\} \cap \{\delta_j = 0\}\big],
\end{aligned}$$

with again $i, j = 1, \ldots, n$ with $i \neq j$. Subjects in these sets contribute to the overall AUC appropriately weighted with the probability that they would be comparable, i.e.,

$$\widehat{\mathrm{AUC}}_m(t, \Delta t) = \frac{\sum_{i=1}^{n} \sum_{j=1; j \neq i}^{n} I\{\hat{\pi}_i(t + \Delta t \mid t) < \hat{\pi}_j(t + \Delta t \mid t)\} \times I\{\Omega_{ij}^{(m)}(t)\} \times \hat{\nu}_{ij}^{(m)}}{\sum_{i=1}^{n} \sum_{j=1; j \neq i}^{n} I\{\Omega_{ij}^{(m)}(t)\} \times \hat{\nu}_{ij}^{(m)}},$$

with $m = 2, 3, 4$ and $\hat{\nu}_{ij}^{(2)} = 1 - \hat{\pi}_i(t + \Delta t \mid T_i)$, $\hat{\nu}_{ij}^{(3)} = \hat{\pi}_j(t + \Delta t \mid T_j)$ and $\hat{\nu}_{ij}^{(4)} = \{1 - \hat{\pi}_i(t + \Delta t \mid T_i)\} \times \hat{\pi}_j(t + \Delta t \mid T_j)$.

3.2 Calibration

The assessment of the accuracy of predictions of survival models is typically based on the expected error of predicting future events. In our setting, and again taking into account the dynamic nature of the longitudinal outcome, it is of interest to predict the occurrence of events at $u > t$ given the information we have recorded up to time t (Schoop et al. 2011).

This gives rise to expected prediction error:

$$\mathrm{PE}(u \mid t) = E\big[L\{N_i(u) - \pi_i(u \mid t)\}\big],$$

where $N_i(t) = I(T_i^* > t)$ is the event status at time t, $L(\cdot)$ denotes a loss function, such as the absolute or square loss, and the expectation is taken with respect to the distribution of the event times. An estimate of $\mathrm{PE}(u \mid t)$ that accounts for censoring has been proposed by Henderson et al. (2002):

$$\begin{aligned}\widehat{\mathrm{PE}}(u \mid t) = \{n(t)\}^{-1} \sum_{i:T_i \geq t} & I(T_i \geq u) L\{1 - \hat{\pi}_i(u \mid t)\} + \delta_i I(T_i < u) L\{0 - \hat{\pi}_i(u \mid t)\} \\ & + (1 - \delta_i) I(T_i < u) \Big[\hat{\pi}_i(u \mid T_i) L\{1 - \hat{\pi}_i(u \mid t)\} + \{1 - \hat{\pi}_i(u \mid T_i)\} L\{0 - \hat{\pi}_i(u \mid t)\}\Big],\end{aligned}$$

where $n(t)$ denotes the number of subjects at risk at time t. The first two terms in the sum correspond to patients who were alive after time u and dead before u, respectively; the third term corresponds to patients who were censored in the interval $[t, u]$. Using the longitudinal information up to time t, $\mathrm{PE}(u \mid t)$ measures the predictive accuracy at the specific time point u. The estimated prediction error $\widehat{\mathrm{PE}}(u \mid t)$ can be used to provide a measure of explained variation between nested models. Assuming model M_1 is nested in model M_2, we can compute how much the extra structure in M_2 improves accuracy by

$$R^2_{PE}(u \mid t; M_1, M_2) = 1 - \widehat{\mathrm{PE}}_{M_2}(u \mid t) \Big/ \widehat{\mathrm{PE}}_{M_1}(u \mid t).$$

4 Analysis of the Aortic Valve Dataset

We return to the Aortic Valve dataset introduced in Section 1. Our aim is to use the existing data and provide accurate predictions of re-operation-free survival for future patients from the same population, utilizing their baseline information, namely age, gender, BMI and the type of operation they underwent, and their recorded aortic gradient levels. In our study, a total of 77 (27%) patients received a sub-coronary implantation (SI) and the remaining

208 patients a root replacement (RR). These patients were followed prospectively over time with annual telephone interviews and biennial standardized echocardiographic assessment of valve function until July 8, 2010. Echo examinations were scheduled at 6 months and 1 year postoperatively, and biennially thereafter, and at each examination, echocardiographic measurements of aortic gradient (mmHg) were taken. By the end of follow-up, 1262 aortic gradient measurements were recorded with an average of 4.3 measurements per patient (s.d. 2.4 measurements), 59 (20.7%) patients had died, and 73 (25.6%) patients required a re-operation on the allograft. The composite event, re-operation or death, was observed for 125 (43.9%) patients, and the corresponding Kaplan-Meier estimator for the two intervention groups is shown in Figure 2.

[Figure 2 about here.]

We can observe minimal differences in the re-operation-free survival rates between sub-coronary implantation and root replacement, with only a slight advantage of sub-coronary implantation towards the end of the follow-up. For the longitudinal process and because aortic gradient exhibits right skewness, we will proceed in our analysis using the square root transform of this outcome. Figure 3 depicts the subject-specific longitudinal profiles of the square root aortic gradient for the two intervention groups.

[Figure 3 about here.]

We observe considerable variability in the shapes of these trajectories, but there are no systematic differences apparent between the two groups.

We start by defining a set of joint models based on which predictions will be calculated. For the longitudinal process we allow a flexible specification of the subject-specific square root aortic gradient trajectories using cubic splines of time. More specifically, the linear mixed model takes the form

$$\begin{aligned} y_i(t) &= \beta_1 \texttt{SI}_i + \beta_2 \texttt{RR}_i + \sum_{k=1}^{3} \beta_{2k+1}\{B_k(t,\lambda) \times \texttt{SI}_i\} + \beta_{2k+2}\{B_k(t,\lambda) \times \texttt{RR}_i\} \\ &\quad + b_{i0} + \sum_{k=1}^{3} b_{ik} B_k(t,\lambda) + \varepsilon_i(t), \end{aligned}$$

where $B_n(t, \lambda)$ denotes the B-spline basis for a natural cubic spline with boundary knots at 0.5 and 13 years and two internal knots at $\lambda = 2.5$ and 6 years, `SI` and `RR` are the dummy variables for the sub-coronary implantation and root replacement groups, respectively, $\varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$ and for the random effects $\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$. Recent research in the field of joint models has shown that different characteristics of the longitudinal profiles may be more strongly associated with the risk of an event (Rizopoulos et al. 2014; Rizopoulos 2016). Following this work, for the survival process we consider three relative risk models, each positing a different association structure between the two processes, namely:

$$
\begin{aligned}
M_1: &\quad h_i(t) = h_0(t) \exp\big\{\gamma_1 \texttt{RR}_i + \gamma_2 \texttt{Age}_i + \gamma_3 \texttt{Female}_i + \alpha_1 m_i(t)\big\},\\
M_2: &\quad h_i(t) = h_0(t) \exp\big\{\gamma_1 \texttt{RR}_i + \gamma_2 \texttt{Age}_i + \gamma_3 \texttt{Female}_i + \alpha_1 m_i(t) + \alpha_2 m_i'(t)\big\}, \quad m_i'(t) = \frac{dm_i(t)}{dt}\\
M_3: &\quad h_i(t) = h_0(t) \exp\Big\{\gamma_1 \texttt{RR}_i + \gamma_2 \texttt{Age}_i + \gamma_3 \texttt{Female}_i + \alpha_1 \int_0^t m_i(s) ds\Big\},
\end{aligned}
$$

where the baseline hazard is approximated with penalized B-splines, i.e.,

$$
\log h_0(t) = \gamma_{h_0,0} + \sum_{q=1}^{Q} \gamma_{h_0,q} B_q(t, \boldsymbol{v}),
$$

with $\boldsymbol{v}$ denoting 13 internal knots placed at the corresponding percentiles of the observed event times, and `Female` denotes the dummy variable for females. All computations have been performed in R (version 3.3.2) using package **JMbayes** (Rizopoulos 2016) Tables 1 and 2 in the supplementary material show estimates and the corresponding 95% credible intervals for the parameters in the longitudinal and survival submodels, respectively.

To assess the predictive ability of aortic gradient, but also to compare joint models with the landmark approach, we consider three follow-up times, namely $t^* = 5.5, 7.5$, and 9.5 years, and a medically relevant window of $\Delta t = 2$ years. For each of the three follow-up times we constructed the versions of the database with the patients at risk at the corresponding

follow-up. Next, in these databases we fitted the Cox models:

$$
\begin{aligned}
M_5: \quad & h_i(u) = h_0(u)\exp\big\{\gamma_1 \texttt{RR}_i + \gamma_2 \texttt{Age}_i + \gamma_3 \texttt{Female}_i + \alpha_1 \tilde{y}_i(t^*)\big\},\\
M_6: \quad & h_i(u) = h_0(u)\exp\big\{\gamma_1 \texttt{RR}_i + \gamma_2 \texttt{Age}_i + \gamma_3 \texttt{Female}_i \\
& \qquad\qquad + \alpha_1 \tilde{y}_i(t^*) + \alpha_2 \tilde{y}'_i(t^*)\big\},\\
M_7: \quad & h_i(u) = h_0(u)\exp\Big\{\gamma_1 \texttt{RR}_i + \gamma_2 \texttt{Age}_i + \gamma_3 \texttt{Female}_i \\
& \qquad\qquad + \alpha_1 \sum_{s=0}^{t^*} y_i(s)\Delta s\Big\},
\end{aligned}
$$

where $u > t^*$, variable $\tilde{y}_i(t^*)$ denotes the last available square root aortic gradient value of each patient before year t^*, $\tilde{y}'_i(t^*)$ denotes the slope defined from the last two available measurements, and $\sum_{0\leq s\leq t^*} y_i(s)\Delta s$ denotes the area under the step function defined from the observed square root aortic gradient measurements up to t^* years. The parameter estimates and confidence intervals of these Cox models are presented in Table 3 in the supplementary material.

Furthermore, for the same follow-up times we also calculated predictions using the mixed-model landmark approach presented in Section 2.3. In particular, we fitted in each of the landmark datasets a linear mixed model for the square root transform aortic gradient with the same specification as the longitudinal submodel from the joint models presented above. From these mixed models and utilizing the empirical Bayes estimates of the random effects, we obtain fully specification of the subject-specific profile $\hat{m}_i(t)$, and based on that we fitted the Cox models

$$
\begin{aligned}
M_8: \quad & h_i(u) = h_0(u)\exp\big\{\gamma_1 \texttt{RR}_i + \gamma_2 \texttt{Age}_i + \gamma_3 \texttt{Female}_i + \alpha_1 \hat{m}_i(t^*)\big\},\\
M_9: \quad & h_i(u) = h_0(u)\exp\big\{\gamma_1 \texttt{RR}_i + \gamma_2 \texttt{Age}_i + \gamma_3 \texttt{Female}_i \\
& \qquad\qquad + \alpha_1 \hat{m}_i(t^*) + \alpha_2 \hat{m}'_i(t^*)\big\},\\
M_{10}: \quad & h_i(u) = h_0(u)\exp\Big\{\gamma_1 \texttt{RR}_i + \gamma_2 \texttt{Age}_i + \gamma_3 \texttt{Female}_i \\
& \qquad\qquad + \alpha_1 \int_0^u \hat{m}_i(s)ds\Big\}.
\end{aligned}
$$

The parameter estimates and confidence intervals of these Cox models are also presented in Table 3 in the supplementary material.

We evaluate both discrimination and calibration capabilities of the aforementioned models using the predictive accuracy measures presented in Section 3. For the calculation of the prediction error we use the quadratic loss function, i.e., $L(x) = x^2$. To account for potential over-fitting in the calculation of these measures we utilize a repeated 5-fold cross-validation procedure. More specifically, 20 times we have randomly split the original dataset in five folds. For each of the 20 repetitions, we performed a cross-validation procedure, in which each model was fitted 5 times in 4/5 of the patients in the training dataset, and the accuracy measures were calculated on 1/5 of the patients in the test dataset. The results from the repeated cross-validation procedure are presented in Table 1.

[Table 1 about here.]

With respect to the discrimination capability, we observe that the value, and value + slope functional forms show higher AUCs than the cumulative effect parameterization. With respect to the prediction error we observe minimal differences between the different functional forms. A comparison between the landmark approaches and joint modeling in this particular dataset shows that joint models perform marginally better in terms of both the prediction error and discrimination.

5 Simulations

5.1 Design

We performed a series of simulations to compare landmarking with joint models in the context of dynamic predictions. The design of our simulation study is motivated by the set of models we fitted to the Aortic Valve dataset in Section 4. In particular, we assume 1000 patients who have been followed-up for a period of 15 years, and were planned to provide longitudinal measurements at baseline and afterwards at 14 random follow-up times. For

the longitudinal process, and similarly to the model fitted in the Aortic Valve dataset, we used B-splines of time with two internal knots placed at 2.5 and 6 years, and boundary knots placed at 0.5 and 13 years, i.e., the form of the model is as follows

$$\begin{aligned} y_i(t) &= \beta_1 \texttt{Trt0}_i + \beta_2 \texttt{Trt1}_i + \sum_{k=1}^{3} \beta_{2k+1}\{B_k(t,\lambda) \times \texttt{Trt0}_i\} + \beta_{2k+2}\{B_k(t,\lambda) \times \texttt{Trt1}_i\} \\ &\quad + b_{i0} + \sum_{k=1}^{3} b_{ik} B_k(t,\lambda) + \varepsilon_i(t), \end{aligned}$$

where $B_n(t, \boldsymbol{\lambda})$ denotes the B-spline basis for a natural cubic spline with $\boldsymbol{\lambda} = (0.5, 2.5, 6, 13)$, `Trt0` and `Trt1` are the dummy variables for the two treatment groups, $\varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$ and $\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$. Figure 1 in the supplementary material gives a visual impression of the subject-specific profiles under the posited model.

For the survival process, we have assumed three scenarios, each one corresponding to a different functional form for the association structure between the two processes:

$$\begin{aligned} \text{Scenario I:} \quad & h_i(t) = h_0(t) \exp\big\{\gamma_1 \texttt{Trt1}_i + \alpha_1 m_i(t)\big\}, \\ \text{Scenario II:} \quad & h_i(t) = h_0(t) \exp\big\{\gamma_1 \texttt{Trt1}_i + \alpha_1 m_i(t) + \alpha_2 m_i'(t)\big\}, \\ \text{Scenario III:} \quad & h_i(t) = h_0(t) \exp\big\{\gamma_1 \texttt{Trt1}_i + \alpha_1 \int_0^t m_i(s) ds\big\}, \end{aligned}$$

with $h_0(t) = \exp(\gamma_0)\sigma_t t^{\sigma_t - 1}$, i.e., the Weibull baseline hazard. The values for the regression coefficients in the longitudinal and survival submodels, the variance of the error terms of the mixed model, the covariance matrix for the random effects, and the scale of the Weibull baseline risk function are given in Section 2.1 of the supplementary material, and have been chosen such that the distribution of the event times and the distribution of the follow-up longitudinal measurements were comparable across scenarios. To better reflect clinical practice, the censoring distribution is allowed to depend on treatment. In particular, for `Trt0` censoring times were simulated from a uniform distribution in the interval $(0, 10)$, and for `Trt1` from a uniform distribution in the interval $(0, 14)$. This resulted in about 45%

censoring in each scenario. For each scenario 1000 datasets were simulated.

5.2 Analysis

Mimicking the real-life use of a prognostic model, and to assess any potential over-fitting issues, the comparison between the landmark and joint modeling approaches is based on subjects who were not used in fitting the corresponding models. More specifically, under each scenario and for each simulated dataset, we randomly excluded 500 subjects that were not used in fitting the models based on which the dynamic survival probabilities are computed. Evaluation of predictive ability was performed at the same three follow-up times we considered for the Aortic Valve dataset, namely $t^* = 5.5, 7.5$, and 9.5 years, and a medically relevant window of $\Delta t = 2$ years. Furthermore, we assess two aspects of model misspecification, namely, misspecification of the functional form that relates the longitudinal and survival outcomes, and misspecification of the functional form of the time effect in the longitudinal model. More specifically, for each simulated dataset we fitted:

- Three joint models, with the correct specification of the longitudinal submodel, and survival submodels:

$$\begin{aligned} h_i(t) &= h_0(t)\exp\big\{\gamma_1 \texttt{Trt1}_i + \alpha_1 m_i(t)\big\}, \\ h_i(t) &= h_0(t)\exp\big\{\gamma_1 \texttt{Trt1}_i + \alpha_1 m_i(t) + \alpha_2 m_i'(t)\big\}, \\ h_i(t) &= h_0(t)\exp\Big\{\gamma_1 \texttt{Trt1}_i + \alpha_1 \int_0^t m_i(s)ds\Big\}, \end{aligned}$$

  and, another three joint models with the same survival submodels as above, and a misspecified longitudinal submodel assuming linear longitudinal evolutions instead of splines, i.e.,

$$y_i(t) = \beta_1 \texttt{Trt0}_i + \beta_2 \texttt{Trt1}_i + \beta_3\{t \times \texttt{Trt0}_i\} + \beta_3\{t \times \texttt{Trt1}_i\} + \varepsilon_i(t).$$

- For the standard landmark approach and for each of the three follow-up times t^* we

constructed the versions of the database with the patients at risk at the corresponding follow-up. Following, in these databases we fitted the Cox models:

$$\begin{aligned} h_i(u) &= h_0(u)\exp\big\{\gamma \texttt{Trt1}_i + \alpha_1 \tilde{y}_i(t^*)\big\}, \\ h_i(u) &= h_0(u)\exp\big\{\gamma \texttt{Trt1}_i + \alpha_1 \tilde{y}_i(t^*) + \alpha_2 \tilde{y}'_i(t^*)\big\}, \\ h_i(u) &= h_0(u)\exp\Big\{\gamma \texttt{Trt1}_i + \alpha_1 \sum_{s=0}^{t^*} y_i(s)\Delta s\Big\}, \end{aligned}$$

where $u > t^*$, the definitions of $\tilde{y}_i(t^*)$ and $\tilde{y}'_i(t^*)$ are the same as in Section 4.

- For the same follow-up times we also calculated predictions using the mixed-model landmark approach. In particular, in each of the landmark datasets we fitted the correctly-specified linear mixed model, and utilizing the empirical Bayes estimates of the random effects, we obtain a full specification of the subject-specific profile $\hat{m}_i(t)$. Using these terms we fitted the Cox models

$$\begin{aligned} h_i(u) &= h_0(u)\exp\big\{\gamma \texttt{Trt1}_i + \alpha_1 \hat{m}_i(t^*)\big\}, \\ h_i(u) &= h_0(u)\exp\big\{\gamma \texttt{Trt1}_i + \alpha_1 \hat{m}_i(t^*) + \alpha_2 \hat{m}'_i(t^*)\big\}, \\ h_i(u) &= h_0(u)\exp\Big\{\gamma \texttt{Trt1}_i + \alpha_1 \int_0^u \hat{m}_i(s)ds\Big\}. \end{aligned}$$

  In addition, similarly to joint models above, predictions from the mixed-model landmark approach were also calculated from the misspecified longitudinal model that assumes linear time evolutions.

For all the aforementioned combinations we evaluated both discrimination and calibration capabilities using the predictive accuracy measures presented in Section 3. Similarly to the Aortic Valve dataset, for all comparisons for the AUC we set $\Delta t = 2$ and for PE $u = t+2$. In addition, for the calculation of the prediction error we use again the quadratic loss function. However, in order to obtain a more objective comparison between the two frameworks, the censoring weights $\nu_{ij}^{(m)}, m = 2, 3, 4$ in the calculation of the AUC, and $\pi_i(u \mid T_i)$ in the calculation of PE are based on the true model under each scenario, using the true parameter

values and true values of the random effects.

5.3 Results

The results from the simulation study are presented in Figures 2–13 in Section 2.2 of the supplementary material. Based on these results we can make the following observations: First, when there is no misspecification of the functional form of the time effect in the longitudinal submodel, in the majority of the cases the joint modeling approach seems to outperform the landmark approaches. However, when there is a strong misspecification of the time effect, the differences become smaller, and in a few cases the landmark approaches outperformed joint models. Second, misspecification of the functional form that links the two processes (i.e., current value, current value & current slopes and cumulative effect) did not seem to particularly influence the relative performance between the joint modeling and landmark approaches. The previous two observations are with respect to both the AUCs and the prediction errors.

6 Discussion

In this work we have contrasted and compared two popular approaches, namely landmarking and joint modeling, for producing dynamically-updated predictions of survival probabilities with time-dependent covariates. Landmarking can effortlessly be implemented in practice but to be valid it implies strong assumptions about the path of the time-varying covariate, which may be unrealistic for longitudinal biomarker measurements. On the other hand, joint modeling allows for greater flexibility in the attributes of time-dependent covariate process, but requires more modeling assumptions to achieve this and is generally more computationally intensive. Our simulation study and the analysis of the motivating Aortic Valve dataset have shown that, in general, there is a gain from considering the joint modeling approach instead of landmarking. However, it should be mentioned that only the standard landmarking approach and its mixed models extensions have been considered due to the fact that this

versions of the technique are predominantly used in practice. In the literature several other extensions of landmarking have been proposed that may alleviate some of the shortcomings of the standard landmarking but at the expense of extra computational complexity (van Houwelingen and Putter 2011; Parast et al. 2012; Parast and Cai 2013; Nicolaie et al. 2013). Similarly, several extensions of joint models have been proposed in the literature to further improved predictions from joints models (Rizopoulos et al. 2014; Andrinopoulou et al. 2015; Andrinopoulou and Rizopoulos 2016; Njeru Njagi et al. 2016).

Regarding the software implementation of the methodology presented in the paper, the landmark approach is readily available in all statistical software that fit Cox models. The fitting of joint models, the derivation of dynamic predictions (for the survival and longitudinal outcomes) and the calculation of the calibration and discrimination measures presented in Section 3 are implemented in the freely available R packages **JM** (Rizopoulos 2010, 2012) and **JMbayes** (Rizopoulos 2016), which can be downloaded from CRAN at `http://cran.r-project.org/package=JM` and `http://cran.r-project.org/package=JMbayes`, respectively. Code to replicate the analysis in the paper is available on the GitHub repository `https://github.com/drizopoulos/jm_and_lm`.

Supplementary material

Supplementary material referenced in Section 5 are available in `SuppCompParam.pdf`.

Acknowledgements

The first author would like to acknowledge support by the Netherlands Organization for Scientific Research VIDI grant nr. 016.146.301.

Conflict of Interest *The authors have declared no conflict of interest.*

References

Anderson, J., Cain, K., and Gelber, R. (1983). Analysis of survival by tumor response. *Journal of Clinical Oncology* **1,** 710–719.

Andrinopoulou, E. and Rizopoulos, D. (2016). Bayesian shrinkage approach for a joint model of longitudinal and survival outcomes assuming different association structures. *Statistics in Medicine* **35,** 4813–4823.

Andrinopoulou, E., Rizopoulos, D., Takkenberg, J., and Lesaffre, E. (2015). Combined dynamic predictions using joint models of two longitudinal outcomes and competing risk data. *Statistical Methods in Medical Research* page doi: 10.1177/0962280215588340.

Bekkers, J., Klieverik, L., Raap, G., Takkenberg, J., and Bogers, A. (2011). Re-operations for aortic allograft root failure: Experience from a 21-year single-center prospective follow-up study. *European Journal of Cardio-Thoracic Surgery* **40,** 35–42.

Blanche, P., Proust-Lima, C., Loubére, L., Berr, C., Dartigues, J., and Jacqmin-Gadda, H. (2015). Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics* **71,** 102–113.

Gerds, T. and Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal* **48,** 1029 – 1040.

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.

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

Henderson, R., Diggle, P., and Dobson, A. (2002). Identification and efficacy of longitudinal markers for survival. *Biostatistics* **3,** 33–50.

Ibrahim, J., Chen, M., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer-Verlag, New York.

Jewell, N. and Nielsen, J. (1993). A framework for consistent prediction rules based on markers. *Biometrika* **80,** 153–164.

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

Nicolaie, M., van Houwelingen, J., de Witte, T., and Putter, H. (2013). Dynamic prediction by landmarking in competing risks. *Statistics in Medicine* **32,** 2031–2047.

Njeru Njagi, E., Molenberghs, G., Rizopoulos, D., Verbeke, G., Kenward, M., Dendale, P., and Willekens, K. (2016). A flexible joint modelling framework for longitudinal and time-to-event data with overdispersion. *Statistical Methods in Medical Research* **25,** 1661–1676.

Parast, L. and Cai, T. (2013). Landmark risk prediction of residual life for breast cancer survival. *Statistics in Medicine* **32,** 3459–3471.

Parast, L., Cheng, S.-C., and Cai, T. (2012). Landmark prediction of long term survival incorporating short term event time information. *Journal of the American Statistical Association* **107,** 1492–1501.

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. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67,** 819–829.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data, with Applications in R.* Chapman & Hall/CRC, Boca Raton.

Rizopoulos, D. (2016). The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *Journal of Statistical Software* **72 (7),** 1–45.

Rizopoulos, D., Hatfield, L., Carlin, B., and Takkenberg, J. (2014). Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging. *Journal of the American Statistical Association* **109,** 1385–1397.

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

Schoop, R., Schumacher, M., and Graf, E. (2011). Measures of prediction error for survival data with longitudinal covariates. *Biometrical Journal* **53,** 275–293.

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

Uno, H., Cai, T., Tian, L., and Wei, L. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association* **102,** 527–537.

van Houwelingen, H. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics* **34,** 70–85.

van Houwelingen, H. and Putter, H. (2011). *Dynamic Prediction in Clinical Survival Analysis.* Chapman & Hall/CRC, Boca Raton.

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. (2005). Partly conditional survival models for longitudinal data. *Biometrics* **61,** 379–391.

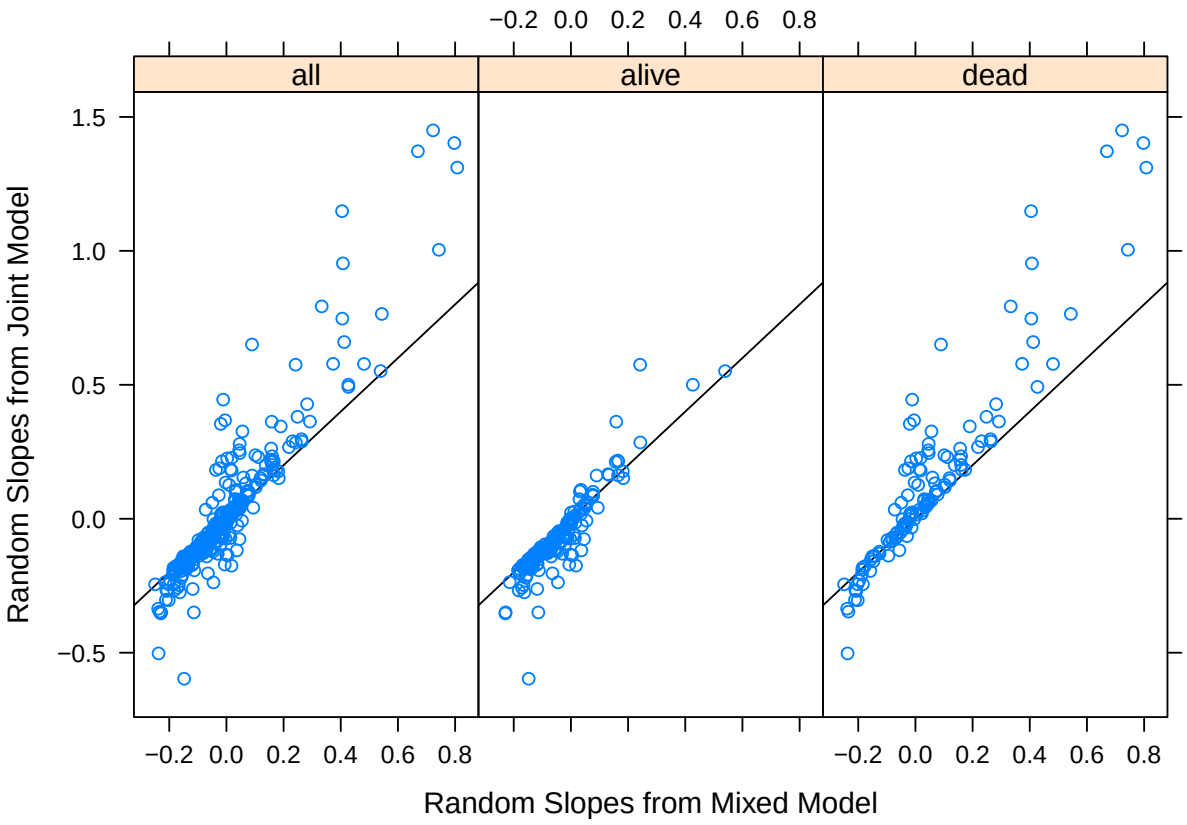

Figure 1: Estimated random slopes from a linear mixed and a joint model split according to event status.

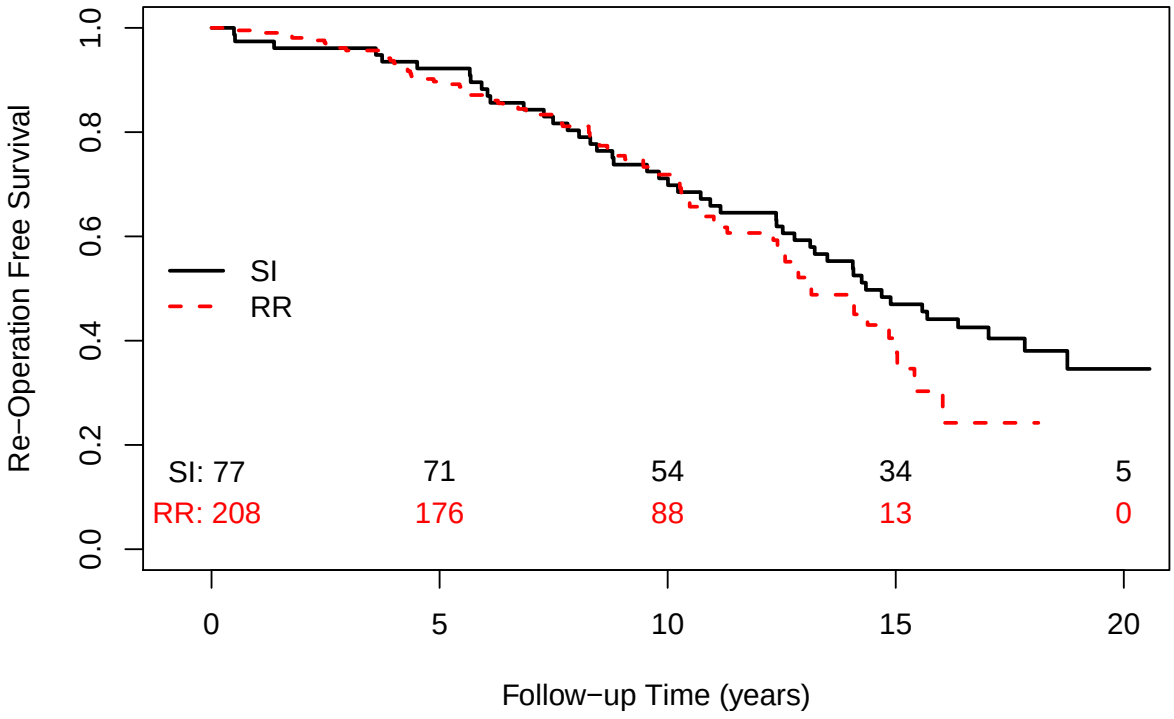

Figure 2: Kaplan-Meier estimates of the survival functions for re-operation-free survival for the sub-coronary implantation (SI) and root replacement (RR) groups.

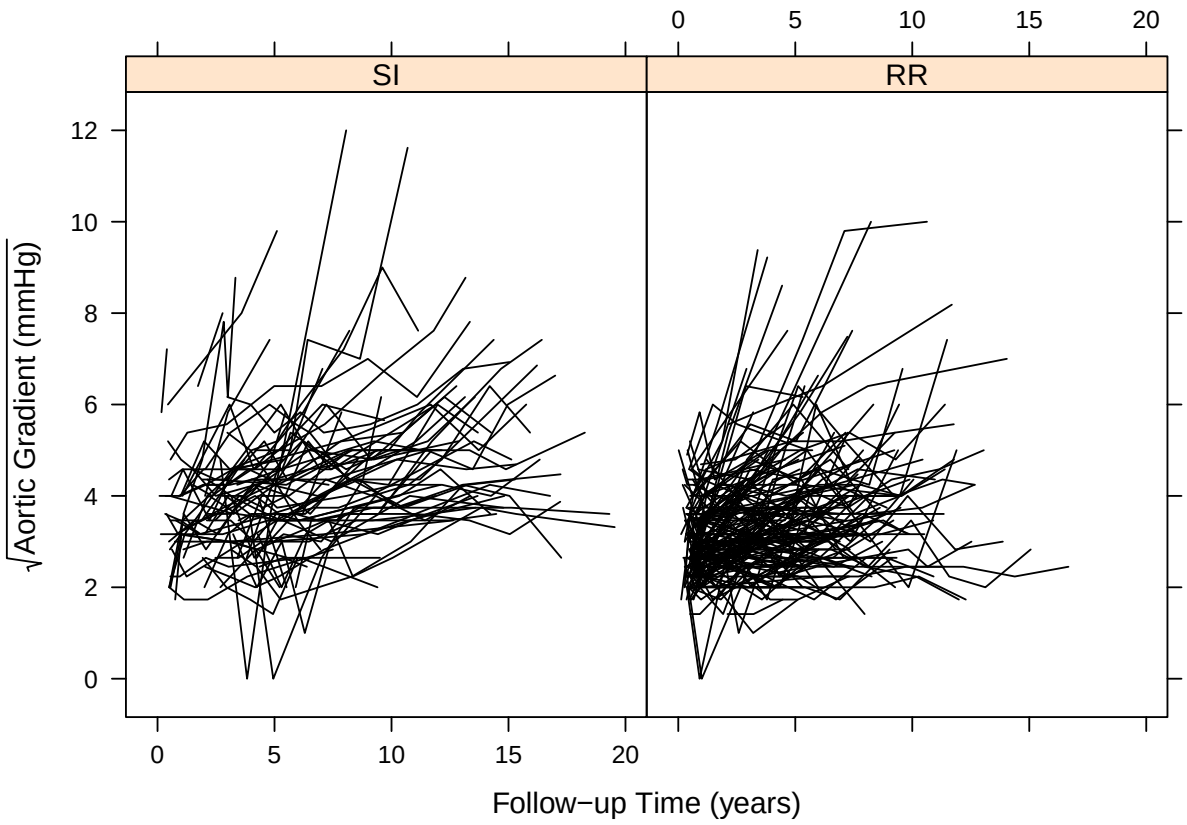

Figure 3: Subject-specific profiles for the square root aortic gradient separately for the sub-coronary implantation (SI) and root replacement (RR) groups.

Table 1: Cross-validated AUC and prediction error using the quadratic loss functions for the joint models, the landmarking approach and the landmarking combined with a linear mixed model for different functional forms. The accuracy measures are estimated for the follow-up times 5.5, 7.5 and 9.5 years, and a medically-relevant window of 2 years. The results are based on a 5-fold cross-validation repeated 20 times.

		AUC($t^*, 2$)			PE($t^* + 2 \mid t^*$)		
functional form	t^*	JM	LM	LMmixed	JM	LM	LMmixed
value	5.5	0.602	0.609	0.592	0.076	0.078	0.079
value+slope	5.5	0.605	0.557	0.516	0.076	0.079	0.078
cumulative	5.5	0.558	0.473	0.554	0.077	0.080	0.080
value	7.5	0.598	0.496	0.508	0.100	0.104	0.104
value+slope	7.5	0.591	0.462	0.500	0.100	0.104	0.104
cumulative	7.5	0.600	0.504	0.499	0.096	0.104	0.104
value	9.5	0.607	0.604	0.564	0.134	0.144	0.144
value+slope	9.5	0.618	0.611	0.594	0.132	0.147	0.141
cumulative	9.5	0.563	0.537	0.545	0.129	0.143	0.143