Individual dynamic prediction of clinical endpoint from large dimensional longitudinal biomarker history: a landmark approach

Anthony Devaux, Robin Genuer, Karine Pérès, Cécile Proust-Lima

Individual dynamic prediction of clinical endpoint from large dimensional longitudinal biomarker history: a landmark approach

Anthony Devaux[1,*], Robin Genuer[1,2], Karine Pérès[1], and Cécile Proust-Lima[1]

[1]INSERM, BPH, U1219, University of Bordeaux, Bordeaux, France
[2]INRIA Bordeaux Sud-Ouest, Talence, France
[*]Email: anthony.devaux@u-bordeaux.fr

January 2022

Abstract: The individual data collected throughout patient follow-up constitute crucial information for assessing the risk of a clinical event, and eventually for adapting a therapeutic strategy. Joint models and landmark models have been proposed to compute individual dynamic predictions from repeated measures to one or two markers. However, they hardly extend to the case where the patient history includes much more repeated markers. Our objective was thus to propose a solution for the dynamic prediction of a health event that may exploit repeated measures of a possibly large number of markers. We combined a landmark approach extended to endogenous markers history with machine learning methods adapted to survival data. Each marker trajectory is modeled using the information collected up to the landmark time, and summary variables that best capture the individual trajectories are derived. These summaries and additional covariates are then included in different prediction methods adapted to survival data, namely regularized regressions and random survival forests, to predict the event from the landmark time. We also show how predictive tools can be combined into a superlearner. The performances are evaluated by cross-validation using estimators of Brier Score and the area under the Receiver Operating Characteristic curve adapted to censored data. We demonstrate in a simulation study the benefits of machine learning survival methods over standard survival models, especially in the case of numerous and/or nonlinear relationships between the predictors and the event. We then applied the methodology in two prediction contexts: a clinical context with the prediction of death in primary biliary cholangitis, and a public health context with age-specific prediction of death in the general elderly population. Our methodology, implemented in R, enables the prediction of an event using the entire longitudinal patient history, even when the number of repeated markers is large. Although introduced with mixed models for the repeated markers and methods for a single right censored time-to-event, the technique can be used with any other appropriate modeling technique for the markers and can be easily extended to competing risks setting.

Keywords: Individual prediction; Landmark; Longitudinal data; Survival data; Machine learning methods

1 Background

A central issue in health care is to quantify the risk of disease, disease progression or death at the individual level, for instance to initiate or adapt a treatment strategy as soon as possible. To achieve this goal, the information collected at a given time (at diagnosis or at the first visit) is often not sufficient and repeated measurements of markers are essential. For example, repeated prostate specific antigen (PSA) data are highly predictive of the risk of prostate cancer recurrence [1, 2, 3], and markers such as diabetic status or blood pressure level over time are crucial in predicting the risk of cardiovascular disease [4, 5]. Including longitudinal information into the prediction of a clinical event defines the framework for individual dynamic predictions [1, 6, 7]. In some contexts, a single marker may be sufficient to predict the occurrence of the event (e.g., in prostate cancer with PSA) but often the complete patient history with possibly many repeated markers should be exploited (see Figure 1). Yet, statistical developments for individual prediction of event have so far either focused on the repeated nature of the information or on its large dimension.

When using repeated information to develop dynamic prediction tools, two approaches are commonly used: joint models [1, 6] and landmark models [8]. Joint models simultaneously analyze the longitudinal and event time processes by assuming a structure of association built on summary variables of the marker dynamics [9]. This model which uses all the information on the longitudinal and time-to-event processes to derive the prediction tool is widely used in the case of a single repeated marker but becomes intractable in the presence of more than a few repeated markers due to high computational complexity [7].

An alternative is to use partly conditional survival model [10] or landmark models [8] which consist in directly focusing on the individuals still at risk at the landmark time and consider their history up to the landmark time (see Figure 1). When individual history includes repeated measures of an endogenous marker, summaries of the marker derived from preliminary mixed models can be included in the survival model, instead of only the last observed value [1, 5]. Although the landmark models do not use as much information as the joint model (only information from the at-risk individuals at the landmark time is exploited) and thus may lack of efficiency, they have shown competitive predictive performances, easier implementation (much less numerical problems) and better robustness to misspecification than joint models [7]. However, as joint models, they necessitate to consider the actual nature of the relationship between the marker and the event.

Although the landmark approach is *per se* very general, in practice its definition is based on standard survival models, namely the Cox model, which prevents the methodology to be applied in large dimensional contexts usually encountered in applications. Indeed the Cox model becomes rapidly limited in the presence of: 1) a large number of predictors, 2) highly correlated predictors, and 3) complex relationships between the predictors and the event [11]. Yet, in the context of dynamic prediction from multiple repeated markers, these three limits are rapidly reached. Indeed, the large dimension of the predictors does not only come from the number of markers but also from the number of (potentially correlated with each other) marker-specific summaries that are necessary to approximate the actual nature of the relationship between the marker and the event.

Machine learning methods, including regularized regressions or decision trees and random forests, have been specifically developed to predict outcomes while tackling the aforementioned issues [12]. Their good predictive performances have been largely demonstrated in the literature [13]. Initially proposed for continuous or binary outcomes, they have been recently extended to handle right censored time-to-event data. For instance, Simon *et al.* [14] developed penalized Cox models either using Ridge, Lasso or Elastic-Net penalty, Bastien *et al.* [15] developed a Cox model based on deviance residuals-based sparse-Partial Least Square, as an extension of sparse-Partial Least Square [16] for survival data, and Ishwaran *et al.* [17] extended random forests to survival

data. However, they were mostly applied to predict time-to-event from time-independent marker information. Our purpose is thus to show how these machine learning methods can also be leveraged to provide dynamic individual predictions from large dimensional longitudinal biomarker data.

Computing dynamic predictions in the context of a large number of repeated markers is a very new topic in statistics, and only a few proposals have been made very recently. Zhao *et al.* [18] and Jiang *et al.* [19] focused on random forests. Using a landmark approach, Zhao *et al.* transformed the survival data into pseudo-observations and incorporated in each tree the marker information at a randomly selected time. Although handling repeated markers, this method neither accounts for measurement errors of the biomarkers nor their trajectory shapes. By considering a functional ensemble survival tree, Jiang *et al.* overcame this issue. They characterized the changing patterns of continuous time-varying biomarkers using functional data analysis, and incorporated those characteristics directly into random survival forests. By concomitantly analyzing the markers and the event, this approach belongs to the two-stage calibration approaches [20] and may suffer from the same biases [21]. Finally Tanner *et al.* [22] proposed to extend the landmark approach to incorporate multiple repeated markers with measurements errors. For the survival prediction method, they chose to discretize the time and use an ensemble of classical binary classifiers to predict the event.

In comparison with this emerging literature, our proposal goes one step forward. As in Tanner *et al.*, we chose to rely on a landmark approach and consider various prediction methods rather than only random forests. However, we also chose to directly exploit the survival data in continuous time. In addition, our methodology handles markers of different nature, accounts for their measurement error and intermittent missing data, and for a possibly large number of summary characteristics of each marker.

In the following sections, we first describe the proposed method. We then demonstrate in a simulation study the performances of the methodology and the benefit of using machine learning methods to handle the large dimensional aspect. We then illustrate the methodology in two very different contexts: a clinical context with the prediction of death in primary biliary cholangitis, and a public health context with the prediction of 5-year death at different ages in the general elderly population. The paper ends with the discussion of the strengths and weaknesses of the proposed method.

2 Methods

2.1 Framework, notations and general principle

Let us consider a landmark time t_{LM} of interest and a population of $N_{t_{LM}}$ individuals that are still at risk of the event at t_{LM}. For an individual $i \in \{1, \ldots, N_{t_{LM}}\}$, we denote T_i the true event time, C_i the independent censoring time. We define $T_i^\star = \min(T_i, C_i)$ the observed time event and $\delta_i = \mathbb{1}(T_i < \min(C_i, t_{LM} + t_{Hor}))$ the event indicator with t_{Hor} the horizon time. We consider a single event for simplicity.

At the landmark time, P time-independent covariates X_i are available, and the history of K time-dependent markers Y_{ijk} ($k \in \{1, \ldots, K\}$) measured at time t_{ij} ($j \in \{1, \ldots, n_i\}$) and $t_{ij} \leq t_{LM}$.

The target individual probability of event from the landmark time t_{LM} to the horizon time t_{Hor} of a subject $\star$ is defined as:

$$\pi_\star(t_{LM}, t_{Hor}) = P(T_\star \leq t_{LM} + t_{Hor} \mid T_\star > t_{LM}, \{Y_{\star jk}; k = 1, ..., K, t_{\star jk} \leq t_{LM}\}, X_\star) \qquad (1)$$

By assuming that the history of the K marker trajectories up to t_{LM} can be summarized into a

vector $\Gamma_\star$, we define the following probability:

$$\widetilde{\pi}_\star(t_{LM}, t_{Hor}) = P(T_\star \leq t_{LM} + t_{Hor} \mid T_\star > t_{LM}, \Gamma_\star(t_{LM}), X_\star) \tag{2}$$

This probability is estimated by $\widehat{\pi}_\star^{(m)}(t_{LM}, t_{Hor})$ in 4 steps on a learning sample for each survival prediction method m:

1. Each marker trajectory is modeled using the information collected up to t_{LM}
2. The vector of summary variables $\Gamma_i(t_{LM})$ is computed for each individual i
3. $\Gamma_i(t_{LM})$ and additional baseline covariates X_i are entered into survival prediction method m
4. The predicted probability of event $\widehat{\pi}_\star^{(m)}(t_{LM}, t_{Hor})$ is computed from survival method m

Once the estimator defined (i.e., the survival prediction method trained) on the learning sample, the summary variables $\Gamma_\star(t_{LM})$ can be computed for any new external individual $\star$ at risk of event at t_{LM}, and the corresponding individual predicted probability of event can be deduced.

2.2 Step 1. Longitudinal model for markers history

Longitudinal markers are usually measured at intermittent times with error. The first step consists to estimate the error-free trajectory of the marker of each individual over the history period. We propose to use generalized mixed models [23] defined as:

$$g(E(Y_{ijk}|b_{ik})) = Y_{ik}^*(t_{ijk}) = X_{ik}^\top(t_{ijk})\beta_k + Z_{ik}^\top(t_{ijk})b_{ik} \tag{3}$$

where $X_{ik}^\top(t_{ijk})$ and $Z_{ik}^\top(t_{ijk})$ are the p_k- and q_k-vectors associated with the fixed effects β_k and random effects b_{ik} (with $b_{ik} \sim \mathcal{N}(0, B_k)$), respectively. The link function $g(.)$ is chosen according to the nature of Y_{ijk} (e.g. identity function for Gaussian continuous markers or logit function for binary markers).

2.3 Step 2. Summary characteristics of the marker trajectories

Once the parameters of the model have been estimated (indicated by $\widehat{\ }$ below), any summary that captures the marker behavior up to the time t_{LM} can be computed. We give here a non-exhaustive list for individual i:

- Predicted individual deviations to the mean trajectory: $\widehat{b}_{ik} = \widehat{B}_k Z_{ik}^\top \widehat{V}_{ik}^{-1}(Y_{ik} - X_{ik}\widehat{\beta}_k)$ where $\widehat{V}_{ik} = Z_{ik}\widehat{B}_k Z_{ik}^\top + \widehat{\sigma}_{\epsilon k} I_{n_i}$, if the marker is continuous. Otherwise, $\widehat{b}_{ik} = \underset{b_{ik}}{\text{argmax}}\ f(b_{ik}|Y_{ik}^*) = \underset{b_{ik}}{\text{argmax}}\ f(Y_{ik}^*|b_{ik})f(b_{ik})$ with $f(.)$ the density function;
- Error-free level at time $u \leq t_{LM}$: $\widehat{Y}_{ik}^*(u) = X_{ik}^\top(u)\widehat{\beta}_k + Z_{ik}^\top(\tau)\widehat{b}_{ik}$;
- Error-free slope at time $u \leq t_{LM}$: $\widehat{Y}_{ik}^{*\prime}(u) = \frac{\partial \widehat{Y}_{ik}^*(t)}{\partial t}|_{t=u}$;
- Cumulative error-free level during period $\mathcal{T}$: $\widehat{h}_{ik}(t_{LM}) = \int_{t_{LM}-\mathcal{T}}^{t_{LM}} \widehat{Y}_{ik}^*(u)du$.

Any additional summary that is relevant for a specific disease can be considered as soon as it is a function of the error-free marker trajectory (e.g., time spent above/below a given threshold). All the individual summary characteristics across the K markers are stored into a vector

Γ_i. Using the list above and $u = t_{LM}$, $\Gamma_i(t_{LM}) = \{\Gamma_{ik}(t_{LM}), k = 1, ..., K\}$ with $\Gamma_{ik}(t_{LM}) = (\widehat{b}_{ik}, \widehat{Y}^*_{ik}(t_{LM}), \widehat{Y}^{*\prime}_{ik}(t_{LM}), \widehat{h}_{ik}(t_{LM}))^\top$ is of length $\sum_{k=1}^{K}(q_k+3)$. This vector may have a large amount of summaries which can also be highly correlated with each other. These particularities have to be taken into account in survival prediction methods.

2.4 Step 3. Prediction methods for survival data in a large dimensional context

To predict the risk of event from t_{LM} to a horizon time t_{Hor} using the vector $\mathcal{X}_i = (\Gamma_i, X_i)$ of summaries Γ_i and time-independent variables X_i of length P, we can use any technique that handles 1) right-censored time-to-event data, 2) the possibly high dimension, 3) and the correlation between the predictors. We focused in this work on Cox model, Penalized-Cox model, Deviance residuals-based sparse-Partial Least Square and Random Survival Forests, although other techniques could also be applied. For each technique, several sub-methods were considered that differ according to the type of variable selection and/or the hyperparameters choices. We briefly describe the different techniques and sub-methods below, and refer to Section 1 in supplementary material for further details.

2.4.1 Cox models

The Cox model is a semi-parametric regression which models the instantaneous risk of event according to a log-linear combination of the independent covariates:

$$\lambda_i(t|\Gamma_i, X_i) = \lambda_0(t) \exp\left(X_i\gamma + \Gamma_i\eta\right) \tag{4}$$

with λ_0 the baseline hazard function, γ and η the coefficients estimated by partial likelihood. We defined two sub-models whether variable selection was performed according to backward selection procedure using `step`() R function (called *Cox-SelectVar*) or not (*Cox-AllVar*).

2.4.2 Penalized-Cox models

Penalized-Cox models extend the Cox model defined in (4) to handle a high number of possibly correlated predictors. The partial log-likelihood is penalized with norm ℓ_2 (Ridge penalty), norm ℓ_1 (Lasso penalty [24]) which enables covariate selection, or a mixture of both (Elastic-Net [14]). These methods require the tuning of the norms mixing parameter (0 for Lasso, 1 for Ridge, $]0;1[$ for Elastic-Net) and the penalty parameter. We used `cv.glmnet`() function (from the `glmnet` R package) with internal cross-validation to tune the penalty parameter, and we defined three sub-models according to the norms mixing parameter (i.e. Lasso, Ridge or Elastic-Net). There are called *Penal-Cox-Lasso*, *Penal-Cox-Ridge* and *Penal-Cox-Elastic*, respectively.

2.4.3 Deviance residuals-based sparse-Partial Least Square (sPLS-DR)

Partial Least Square (PLS) is a method of dimension reduction where components (or latent variables) are built to maximize the covariance with the outcome. Sparse-PLS (sPLS) [16] adds a variable selection within each component using Lasso penalty. First developed in the framework of linear regression, this method was extended to survival data [15] (sPLS-DR). The principle is to apply a sPLS regression on the deviance residuals which are a normalized transformation of the martingale residuals $\widehat{\mathcal{M}}_i = \delta_i - \widehat{\Lambda}_i(t)$, with $\widehat{\Lambda}_i(t)$ the Nelson-Aalen cumulative hazard function estimate. Then, a Cox model is applied using the C identified components $f_c(\Gamma_i, X_i)$ as covariates. In sPLS, the number of components C and the Lasso penalty parameter on each component (which controls

the sparsity on each component) have to be properly tuned. We used `cv.coxsplsDR()` function (from `plsRcox` R package) with internal cross-validation to tune the number of components, and considered three variants for the penalty: no penalty (called *sPLS-NoSparse*), maximum penalty (called *sPLS-MaxSparse*), or an optimized penalty from a grid of values (called *sPLS-Optimize*).

2.4.4 Random Survival Forests

Random forests [12] are a non-parametric machine learning tool that can handle high-dimensional data with possibly complex input-output relationships. Random forests, originally developed in a context of regression or classification, were later adapted to right-censored survival data [17] and called random survival forests (RSF). A RSF aggregates B survival trees, each one built on a different bootstrap sample from the original data (subjects not included in one bootstrap sample are called out-of-bag (OOB)). As any tree-based predictor, a survival tree recursively splits the sample into subgroups until the subgroups reach a certain minimal size S. To deal with time-to-event data, the splitting rule is usually based on the log-rank statistics although other splitting rules have also been proposed (e.g. gradient-based brier score splitting [17]). In RSF, at each node of each tree, a subset of M predictors is randomly drawn and the split is optimized among splits candidates only involving those predictors. The size of the predictors subset M and the minimal size S have to be tuned.

The interpretation of the link between the predictors and the event is not as easy in RSF as in (penalized) regression methods. To address this issue, RSF provide a quantification of this association, also known as variable importance (VIMP). For a given predictor p, $VIMP^{(p)}$ measures the mean (over all trees in the forest) increase of a tree error on its associated OOB sample, after randomly permuting the values of the p^{th} predictor in the OOB sample. Large VIMP values indicate variables with prediction ability while null (or even negative) VIMP values indicate variables that could be removed from the prediction tool.

Using `rfsrc()` function (from `randomForestSRC` R package), three RSF sub-methods were considered that differed according to M and S parameter tuning: default software parameters $M =$ square root of the number of predictors, $S = 15$ (called *RSF-Default*), M and S that minimize the OOB error (called *RSF-Optimize*) or M and S optimized plus a variable selection using the VIMP statistic (called *RSF-SelectVar*).

2.5 Step 4. Predicted individual probability of event

The estimator of individual probability of event $\widehat{\pi}_{\star}^{(m)}(t_{LM}, t_{Hor})$ for a new patient $\star$ becomes:

- For Cox, penalized-Cox and sPLS-DR models:

$$\widehat{\pi}_{\star}^{(m)}(t_{LM}, t_{Hor}) = 1 - \exp\left(-\widehat{\Lambda}_0(t_{Hor}) \exp\left(\widehat{\mathcal{P}}_{\star}\right)\right) \tag{5}$$

  with $\widehat{\Lambda}_0(.)$ the Nelson-Aalen estimator, and $\widehat{\mathcal{P}}_{\star}$ the predicted linear predictor directly obtained from $\Gamma_{\star}$ and $X_{\star}$ for Cox and Penalized-Cox models, or from the C components $f_c(\Gamma_{\star}, X_{\star})$ $(c = 1, ..., C)$ for sPLS-DR.

- For RSF:

$$\widehat{\pi}_{\star}^{(m)}(t_{LM}, t_{Hor}) = 1 - \exp\left(-\frac{1}{B}\sum_{b=1}^{B} \widehat{\Lambda}_{\star}^{b}(t_{Hor})\right) \tag{6}$$

with $\widehat{\Lambda}^b_\star(t_{Hor})$ the Nelson-Aalen estimator in the leaf of tree b containing individual $\star$.

2.6 Predictive accuracy assessment

We assessed the predictive performances of the models using the time-dependent Area Under the ROC Curve (AUC) [25] defined as:

$$AUC(t_{LM}, t_{Hor}) = P\Big(\pi_i(t_{LM}, t_{Hor}) > \pi_j(t_{LM}, t_{Hor}) \Big| D_i(t_{LM}, t_{Hor}) = 1, \\ D_j(t_{LM}, t_{Hor}) = 0, T_i > t_{LM}, T_j > t_{LM}\Big) \quad (7)$$

and time-dependent Brier score [26] defined as:

$$BS(t_{LM}, t_{Hor}) = E\Big[\left(D_i(t_{LM}, t_{Hor}) - \pi(t_{LM}, t_{Hor})\right)^2 \Big| T > t_{LM}\Big] \quad (8)$$

where $D_i(t_{LM}, t_{Hor})$ is the survival status at time $t_{LM} + t_{Hor}$. We used estimators of these quantities that specifically handle the censored nature of $D_i(t_{LM}, t_{Hor})$ using inverse censoring probability weighting (see [26, 27] for details).

In the applications, predictive accuracy assessment was done using a cross-validation approach to ensure independence between the samples on which each predictive tool was learnt and the samples on which their predictive accuracy was assessed (Figure 2A). This induced a two-layer cross-validation since a cross-validation (or a bootstrap) was also performed within each training set to determine the method-specific hyperparameters.

2.7 Combining the predictions into a single Super Learner

Each survival prediction method m ($m = 1, ..., M$) provides a different individual predicted probability $\widehat{\pi}^{(m)}_\star$ (equation 2). In some cases, one will prefer to select the best predictive tool and rely on it. In other cases, one can also choose to combine the predictive tools into a Super-Learner predictive tool [28, 29]. It consists in defining the final predicted probability as a weighted mean over the survival method-specific predictions:

$$\widehat{\Pi}_\star = \sum_{m=1}^{M} \omega_m \widehat{\pi}^{(m)}_\star \quad (9)$$

where the weights ω_m (defined in $[0, 1]$ with $\sum_{m=1}^{M} \omega_m = 1$) are determined so that the Super-Learner predictive tool $\widehat{\Pi}$ minimizes a loss function. In our work, we chose to minimize the BS function defined in equation (8) by internal cross-validation. This lead to a three-layer cross-validation for the superlearner building and validation (see Figure 2B).

3 Results

3.1 Performances of the methodology through a simulation study

We contrasted the performances of the different survival prediction methods according to a serie of scenarios, based on Ishwaran *et al.* [30], in an extensive simulation study. Prediction tools were trained on $R = 250$ learning datasets and their predictive performances were compared on a unique external validation dataset.

3.1.1 Design

The R learning datasets and the validation dataset were generated according to the same design. They included $N = 500$ individuals at risk of the event at a landmark time t_{LM} of 4 years. Up to landmark time, repeated information on 17 continuous biomarkers was generated according to linear mixed models, as described in equation (3) with identity link.

For each biomarker, measurement times were randomly generated according to a $\mathcal{N}(0, 0.15)$ around 5 theoretical visit times at -4, -3, -2, -1 and 0 years prior to t_{LM}. Different shapes of individual trajectory were considering depending on the biomarker, although all followed an individual polynomial function of time (see figure 1 in additional file). Summary characteristics of each error-free marker trajectory were computed (as defined in "Methods" Section) leading to a total of 92 summaries statistics, stored in a vector Γ_i^0. An additional vector X_i^0 of 10 time-independent covariates was generated at the landmark time: 5 according to a standard normal distribution and 5 according to binomial distribution with success probability of 0.5.

The risk of event after the landmark time was defined according to a proportional hazard model with Γ_i^0 and X_i^0, and a Weibull distribution for the base hazard function, in order to not disadvantage the methods based on the Cox model. Five different scenarios were built according to the number of summaries actually associated to the event (18 or 4 summaries) and the form of the dependence function: biomarkers summaries were entered into the linear predictor either linearly, linearly with interactions across biomarkers, or non-linearly with polynomial functions and binarization of summaries. Details on the generation model are given in section 1 of additional file.

The target prediction was the probability of event up to a horizon of 3 years. The predictive performances of all the survival methods were compared on the external dataset using the BS and AUC introduced in section 2.6, as well as the Mean Square Error of Prediction (MSEP), $MSEP = \frac{1}{N}\sum_{i=1}^{N}(\widehat{\pi}_i - \pi_i^0)^2$, which measures the average squared difference between the estimated probability $\widehat{\pi}_i$ and the true generated probability π_i^0 over all individuals.

3.1.2 Results

Predictive performances for scenarios with 18 summaries are summarized in Figure 3. The same figure for scenarios with 4 summaries is given in Figure S2 of supplementary material.

When considering summaries entered linearly, the penalized-Cox provided the smallest BS and MSEP, and the highest AUC in both scenarios with 4 or 18 summaries associated with the event. When the relationships became increasingly complex (linear with interactions and non-linear), RSF provided better predictive performance than the other methods for both AUC, BS and MSEP regardless of the number of summaries considered.

This simulation study highlights that the penalized-Cox model provides more accurate predictions in the case of simple relationships between the predictors and the event while RSF outperforms the others in the case of complex relationships (no matter how many summaries are considered). In contrast, classical Cox model was systematically outperformed by the other methods which points out the potential benefit of using advanced methods to predict the event in landmark approaches.

3.2 Individual prediction of death in primary biliary cholangitis

We first illustrated our method for predicting death among patients with primary biliary cholangitis (PBC). PBC is a chronic liver disease possibly leading to liver failure. For these patients, the only useful treatment is a liver transplantation [31], and prediction of the risk of death can be useful in that context for patient stratification. We focused on the widely known PBC data from a clinical

trial [32] including repeated measures of 11 biomarkers (7 continuous and 4 binary), such as bilirubin value, albumin value or presence of hepatomegaly, and 3 additional demographic variables collected at the enrollment in the study (see Table S2 in supplementary material for the complete list). We aimed to predict the occurrence of death at horizon time $t_{Hor} = 3$ using information collected up to landmark time $t_{LM} = 4$ years on the $N = 225$ patients still at risk at t_{LM} (see the flow chart Figure S3 in supplementary material).

After a normalization for continuous markers which did not follow a gaussian distribution using splines [33], we modeled independently the markers according to generalized mixed models (see equation 3) with natural splines on time measurements to capture potentially complex behavior over time [34] (See Section 3.1 in supplementary material for details on the models).

We used a 10-fold cross-validation to compute the predictive performances of the individual predicted probabilities. The distribution of the event times did not differ across folds (Figure S4 in supplementary material). For the superlearner, the optimal weights were determined in a second-layer 9-fold cross-validation. We repeated this process $R = 50$ times for all methods to assess the variability of the results across different cross-validation partitions.

Predictive performances are displayed in Figure 4A. All the prediction tools provided satisfying predictive performance for both BS (from 0.076 to 0.089 in mean) and AUC (from 0.73 to 0.87 in mean). Nevertheless, we found that Cox models gave much worst indicators, especially for AUC (the only ones below 0.80 in mean), illustrating the limits of classical methods compared to machine learning methods that handle high dimension and correlation. In this application, the most discriminating and accurate predictions were given by the Cox model with Lasso penalty according to BS (0.076 in mean) and AUC (0.87 in mean). Results from the superlearner did not show substantial improvement in predictive performance. The weights of the superlearner indicated that it was mostly driven by penalized Cox models and RSF (Figure 4B).

For comparison, we also developed predictive tools based on (1) only baseline information for the 11 biomarkers and 3 covariates, (2) information on the 3 covariates and the trajectory of one biomarker over time (either serum bilirubin, albumin or platelets). The predictive tools based only on baseline information provided poorer cross-validation BS (32% higher in mean over the methods) and AUC (8% lower in mean over the methods) nicely illustrating the gain in updating the biomarker information over follow-up (Figure 5). The predictive performances were also worse when considering only repeated albumin or platelets with in mean 22% and 37% higher BS (1% and 11% lower AUC), respectively. In contrast, the predictive tools based on serum bilirubin (the main biomarker in PBC) provided similar performances as the multivariate predictive tool.

3.3 Individual prediction of 5-years death at 80 and 85 years old in the general population

In this second application, we aimed to predict the 5-year risk of death from any cause in the general older population at two different ages: 80 and 85 years old. We relied on the French prospective population-based aging cohort Paquid [35] which included 3777 individuals aged 65 years and older, and followed them up to more than 30 years with general health assessment every two to three years and continuous reporting of death. Beyond the individual quantification of the risk of death, our aim was to identify the main predictors of death and assess whether they differed according to age. The use of landmark models was perfectly adapted to this context with the definition of an age-specific prediction model. We chose to predict the 5-year risk of death from information on 9 markers of aging: depressive symptoms, 3 cognitive functions (general cognition, verbal fluency and executive function), functional dependency, incontinence, dyspnea, the live alone status, and polymedication as a global and easily collected marker of multimorbidity [36]. For each one, we focused on the trajectory over the last 5 years prior to the landmark age. In addition,

we considered 18 other predictors including socio-demographic information (such as generation or gender) and medical history at the last visit prior to the landmark age (such as cardiovascular disease). Complete information on the markers and covariate definitions is given in Section 3.2 and Table S3 of supplementary material. The analysis was done on the samples of individuals still alive at $t_{LM} = 80$ and $t_{LM} = 85$, and with at least one measure for each of the predictors resulting in $N = 1561$ and $N = 1240$ subjects for $t_{LM} = 80$ and $t_{LM} = 85$, respectively (see flowchart Figure S6 in supplementary material).

We used the exact same strategy as explained in the previous application for (i) modeling the trajectories of each marker except that time was the backward time (from -5 to 0 years) from landmark; (ii) computing the external probabilities with a 10-fold cross-validation and computing the superlearner with an internal 5-fold cross-validation. The event time distribution did not differ across folds (see Figures S7 and S8 in supplementary material). Note that due to the impossibility of using predictors with zero or near zero variance in sPLS-DR models, we removed from these models the following predictors: level of education, hearing, dementia, housing and dependency (ADL). RSF hyperparameters tuning (according to OOB error) is reported in supplementary material Figures S9 and S10.

Overall, the predictive performances of all the prediction models were very low with AUC ranging from 0.55 to 0.64 in mean and BS ranging from 0.123 to 0.135 in mean (see Figure S11A in supplementary material) showing the difficulty to accurately predict the age-specific risk of all-cause death in the general population. For both $t_{LM} = 80$ and $t_{LM} = 85$, RSF and the superlearner (which was mostly driven by the RSF (see Figures 11B and 12B in supplementary material) provided the lowest BS, whereas Cox with variable selection and penalized Cox models gave the highest AUC (0.66 in mean).

This application mainly aimed at identifying and contrasting the main age-specific predictors of death at 80 and 85 years old. Figure 6 reports the VIMP from the optimized RSF (variables selected by the Lasso are shown in supplementary material Figure S14). The main predictors of 5-year death were mainly the trajectory of moderate functional dependency and polymedication both at 80 and 85 years old, dyspnea, gender and dementia at 80 years old as well as general self-assessment of health and severe dependency status at 85 years old. The predictors of 5-year death did not substantially differ between the two landmark times for RSF, except for dyspnea, general self-assessment of health and gender.

4 Discussion

We introduced in this paper an original methodology to compute individual dynamic predictions from a large number of time-dependent markers. We proposed to compute this prediction using a landmark approach combined with machine learning methods adapted to survival data. The idea was to incorporate a set of individual summaries of each marker trajectory (obtained in a preliminary longitudinal analysis) as well as other covariates in various prediction methods that could handle a large number of possibly correlated predictors, and complex associations. In addition to each prediction tool, we also proposed a superlearner, as a weighted mean of tool-specific predictions where weights were determined in an internal cross-validation to provide a minimal Brier Score.

Through an extensive simulation study, we showed that regularized Cox models and RSF provided better cross-validated predictive performance over standard Cox model in different scenarios where there was a large number of markers and/or complex associations with the event. This was also observed in two real case applications: a clinical setting where death was predicted from mon-

itored markers in primary biliary cholangitis, and in a setting where all-cause age-specific death was predicted in the general population from main markers of aging.

Providing accurate predictions of health events that can exploit all the available individual information, even measured repeatedly over time, has become a major issue with the expansion of precise medicine. After the first proposals of dynamic predictions from repeated marker information [1, 6], some authors have recently begun to tackle the problem of large dimension of longitudinal markers [22, 18, 19]. In comparison with this recent literature, our method has the advantage of (i) considering any nature of markers with measurement error while other considered only continuous outcomes [19], (ii) proposing the use of many summaries from the biomarkers as individual posterior computation from the longitudinal model (compared for instance to [22] who only include one or two summaries), (iii) exploiting the time-continuous information from survival data rather than discretized scale as in [22], and (iv) considering a vast variety of machine learning techniques as well as a superlearner rather than focusing only on one specific technique [18]. Our methodology does not limit to the specific model and techniques described in the paper, it allows the use of any relevant method at each step. For example, we suggested to capture individual trajectories using generalized mixed models, but we also used functional principal component analysis [37] to characterize the individual variation of the trajectories using eigenfunctions leading to similar results (not shown here). We could also estimate the individual probability using other techniques such as deep learning [38] or random forests based on pseudo-observations [18]. Finally, although we considered for simplicity a single cause of event in this paper, our methodology could be extended to take into account several events through competing risks. For example, we could easily replace random survival forests by their extension that takes into account competing risks [30].

In our simulations and applications, we considered only a few dozens of markers repeatedly measured over time since this is already a challenging situation in individual dynamic prediction context where classical techniques are limited to a few markers. Yet, the method would also apply in a much higher dimensional context (e.g., with omics repeated data) or with a much larger amount of subjects. Indeed, thanks to the landmark approach, the prediction of the summary features for each biomarker and the prediction of the event are done in different steps, and each step-specific technique scales well in higher contexts (i.e., mixed model in large samples, and machine learning techniques in high dimensional context).

Our methodology is relevant for the prediction of an event from a landmark time that is common over subjects or for a small number of common landmark times as done in the application. In other settings where any landmark time should be considered, our methodology would need to be adapted as it currently involves as many prediction tools and the number of landmark times which would result in a considerable increase of computational burden. A possible solution might be to define the prediction tools as a continuous function of the landmarks, following the super landmark models idea [39] but we leave such development for future research.

5 Conclusions

By extending the landmark approach to the large dimensional and repeated setting, our methodology addresses a current major issue in biomedical studies with a complete methodology that has the assets of being (i) easy to implement in standard software (R code is provided at https://github.com/anthonydevaux and (ii) generic as it can be used with any new machine learning technique adapted to survival data, any methodology to model repeated markers over time, any type of possible summary characteristics for the markers, and any number of markers.

References

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

[2] Sène, M., Bellera, C.A., Proust-Lima, C.: Shared random-effect models for the joint analysis of longitudinal and time-to-event data: application to the prediction of prostate cancer recurrence. Journal de la Société Française de Statistique **155**(1), 134–155 (2014). Accessed 2014-05-07

[3] Taylor, J.M.G., Park, Y., Ankerst, D.P., Proust-Lima, C., Williams, S., Kestin, L., Bae, K., Pickles, T., Sandler, H.: Real-Time Individual Predictions of Prostate Cancer Recurrence Using Joint Models: Real-Time Individual Predictions of Prostate Cancer Recurrence Using Joint Models. Biometrics **69**(1), 206–213 (2013). doi:10.1111/j.1541-0420.2012.01823.x

[4] Paige, E., Barrett, J., Stevens, D., Keogh, R.H., Sweeting, M.J., Nazareth, I., Petersen, I., Wood, A.M.: Landmark Models for Optimizing the Use of Repeated Measurements of Risk Factors in Electronic Health Records to Predict Future Disease Risk. American Journal of Epidemiology **187**(7), 1530–1538 (2018). doi:10.1093/aje/kwy018

[5] Sweeting, M.J., Barrett, J.K., Thompson, S.G., Wood, A.M.: The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the ARIC study. Statistics in Medicine **36**(28), 4514–4528 (2017). doi:10.1002/sim.7144

[6] Rizopoulos, D.: Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data. Biometrics **67**(3), 819–829 (2011). doi:10.1111/j.1541-0420.2010.01546.x

[7] Ferrer, L., Putter, H., Proust-Lima, C.: Individual dynamic predictions using landmarking and joint modelling: Validation of estimators and robustness assessment. Statistical Methods in Medical Research **28**(12), 3649–3666 (2019). doi:10.1177/0962280218811837

[8] Van Houwelingen, H.C.: Dynamic Prediction by Landmarking in Event History Analysis. Scandinavian Journal of Statistics **34**(1), 70–85 (2007). doi:10.1111/j.1467-9469.2006.00529.x

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

[10] Maziarz, M., Heagerty, P., Cai, T., Zheng, Y.: On longitudinal prediction with time-to-event outcome: Comparison of modeling options: Prediction Based on Longitudinal and Time-to-Event Data. Biometrics **73**(1), 83–93 (2017). doi:10.1111/biom.12562

[11] Goldstein, B.A., Navar, A.M., Carter, R.E.: Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. European Heart Journal **38**(23), 1805–1814 (2016). doi:10.1093/eurheartj/ehw302

[12] Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (2001). doi:10.1023/A:1010933404324

[13] Lebedev, A.V., Westman, E., Van Westen, G.J.P., Kramberger, M.G., Lundervold, A., Aarsland, D., Soininen, H., Kłoszewska, I., Mecocci, P., Tsolaki, M., Vellas, B., Lovestone, S., Simmons, A.: Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. NeuroImage: Clinical **6**, 115–125 (2014). doi:10.1016/j.nicl.2014.08.023

[14] Simon, N., Friedman, J., Hastie, T., Tibshirani, R.: Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. Journal of Statistical Software **39**(5) (2011). doi:10.18637/jss.v039.i05

[15] Bastien, P., Bertrand, F., Meyer, N., Maumy-Bertrand, M.: Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data. Bioinformatics **31**(3), 397–404 (2015). doi:10.1093/bioinformatics/btu660

[16] Chun, H., Keles, S.: Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72**(1), 3–25 (2010). doi:10.1111/j.1467-9868.2009.00723.x

[17] Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S.: Random survival forests. The Annals of Applied Statistics **2**(3), 841–860 (2008). doi:10.1214/08-AOAS169

[18] Zhao, L., Murray, S., Mariani, L.H., Ju, W.: Incorporating longitudinal biomarkers for dynamic risk prediction in the era of big data: A pseudo-observation approach. Statistics in Medicine **39**(26), 3685–3699 (2020). doi:10.1002/sim.8687

[19] Jiang, S., Xie, Y., Colditz, G.A.: Functional ensemble survival tree: Dynamic prediction of Alzheimer's disease progression accommodating multiple time-varying covariates. Journal of the Royal Statistical Society: Series C (Applied Statistics), 12449 (2020). doi:10.1111/rssc.12449

[20] Ye, W., Lin, X., Taylor, J.M.G.: Semiparametric Modeling of Longitudinal Measurements and Time-to-Event Data-A Two-Stage Regression Calibration Approach. Biometrics **64**(4), 1238–1246 (2008). doi:10.1111/j.1541-0420.2007.00983.x

[21] Albert, P.S., Shih, J.H.: On Estimating the Relationship between Longitudinal Measurements and Time-to-Event Data Using a Simple Two-Stage Procedure. Biometrics **66**(3), 983–987 (2010). doi:10.1111/j.1541-0420.2009.01324_1.x

[22] Tanner, K.T., Sharples, L.D., Daniel, R.M., Keogh, R.H.: Dynamic survival prediction combining landmarking with a machine learning ensemble: Methodology and empirical comparison. Journal of the Royal Statistical Society: Series A (Statistics in Society) (2020). doi:10.1111/rssa.12611

[23] Laird, N.M., Ware, J.H.: Random-Effects Models for Longitudinal Data. Biometrics **38**(4), 963–974 (1982). doi:10.2307/2529876

[24] Goeman, J.J.: L1 Penalized Estimation in the Cox Proportional Hazards Model. Biometrical Journal **52**(1), 70–84 (2009). doi:10.1002/bimj.200900028

[25] Blanche, P., Dartigues, J.-F., Jacqmin-Gadda, H.: Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Statistics in Medicine **32**(30), 5381–5397 (2013). doi:10.1002/sim.5958

[26] Mogensen, U.B., Ishwaran, H., Gerds, T.A.: Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. Journal of Statistical Software **50**(11) (2012). doi:10.18637/jss.v050.i11

[27] Blanche, P., Proust-Lima, C., Loubère, L., Berr, C., Dartigues, J.-F., Jacqmin-Gadda, H.: Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks: Comparing Dynamic Predictive Accuracy of Joint Models. Biometrics **71**(1), 102–113 (2015). doi:10.1111/biom.12232

[28] van der Laan, M.J., Polley, E.C., Hubbard, A.E.: Super Learner. Statistical Applications in Genetics and Molecular Biology **6**(1) (2007). doi:10.2202/1544-6115.1309

[29] Golmakani, M.K., Polley, E.C.: Super Learner for Survival Data Prediction. The International Journal of Biostatistics **16**(2), 20190065 (2020). doi:10.1515/ijb-2019-0065. Place: Berlin, Boston Publisher: De Gruyter

[30] Ishwaran, H., Gerds, T.A., Kogalur, U.B., Moore, R.D., Gange, S.J., Lau, B.M.: Random survival forests for competing risks. Biostatistics **15**(4), 757–773 (2014). doi:10.1093/biostatistics/kxu010

[31] Kaplan, M.M.: Primary Biliary Cirrhosis. New England Journal of Medicine **335**(21), 1570–1580 (1996). doi:10.1056/NEJM199611213352107

[32] Murtaugh, P.A., Dickson, E.R., Van Dam, G.M., Malinchoc, M., Grambsch, P.M., Langworthy, A.L., Gips, C.H.: Primary biliary cirrhosis: Prediction of short-term survival based on repeated patient visits. Hepatology **20**(1), 126–134 (1994). doi:10.1002/hep.1840200120

[33] Proust-Lima, C., Philipps, V., Liquet, B.: Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm. Journal of Statistical Software **78**(2), 1–56 (2017). doi:10.18637/jss.v078.i02

[34] Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., Schmid, M.: A review of spline function procedures in R. BMC Medical Research Methodology **19**(1), 46 (2019). doi:10.1186/s12874-019-0666-3

[35] Helmer, C., Joly, P., Letenneur, L., Commenges, D., Dartigues, J.-F.: Mortality with Dementia: Results from a French Prospective Community-based Cohort. American Journal of Epidemiology **154**(7), 642–648 (2001). doi:10.1093/aje/154.7.642

[36] Schneeweiss, S., Seeger, J.D., Maclure, M., Wang, P.S., Avorn, J., Glynn, R.J.: Performance of comorbidity scores to control for confounding in epidemiologic studies using claims data. American Journal of Epidemiology **154**(9), 854–864 (2001). doi:10.1093/aje/154.9.854

[37] Yao, F., Müller, H.-G., Wang, J.-L.: Functional Data Analysis for Sparse Longitudinal Data. Journal of the American Statistical Association **100**(470), 577–590 (2005). doi:10.1198/016214504000001745

[38] Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology **18**(1), 24 (2018). doi:10.1186/s12874-018-0482-1

[39] Houwelingen, J.C.v., Putter, H.: Dynamic Prediction in Clinical Survival Analysis. Monographs on statistics and applied probability, vol. 123. CRC Press, Boca Raton (2012)

Figures

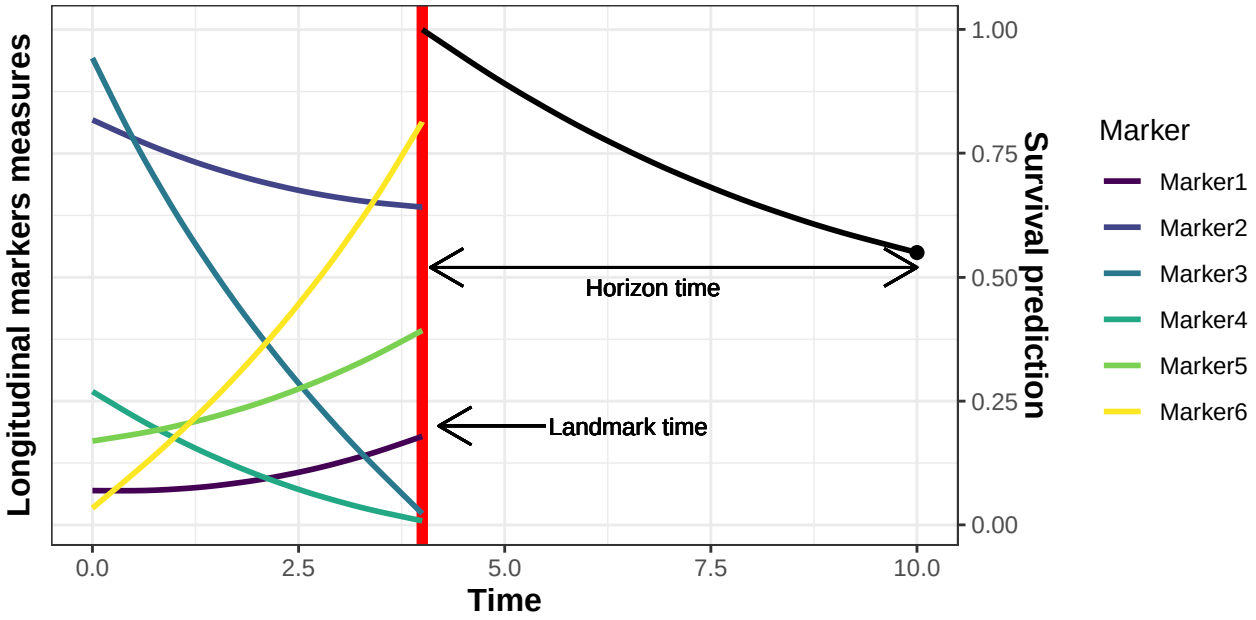

Figure 1: Illustration of individual dynamic prediction of an event computed using history of multiple repeated markers (here 6). The individual probability of event is computed from a landmark time to a horizon time by using the information on the markers trajectories collected up to the landmark time.

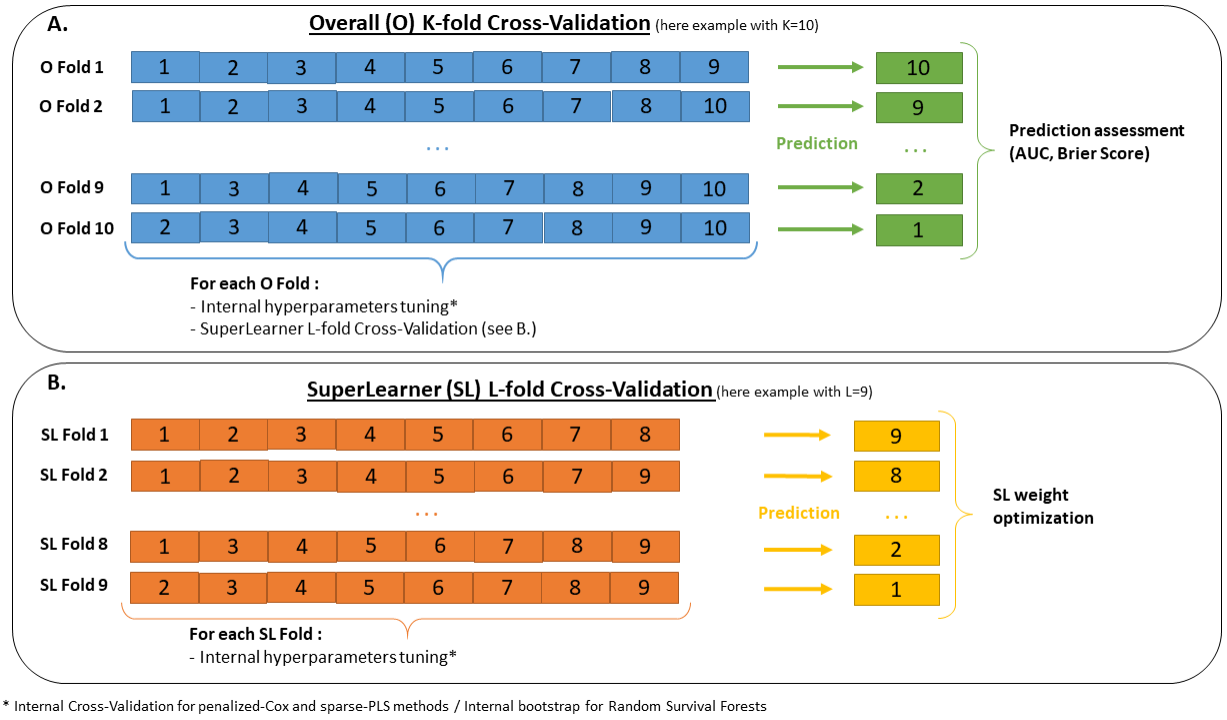

Figure 2: Multi-layer cross-validation framework: (A) Overall cross-validation to assess the predictive performances on independent samples, (B) Intermediate-layer cross-validation for the super-learner only performed on the learning sample to determine the weights. A final internal cross-validation (or Bootstrap for RSF) is done to tune each method.

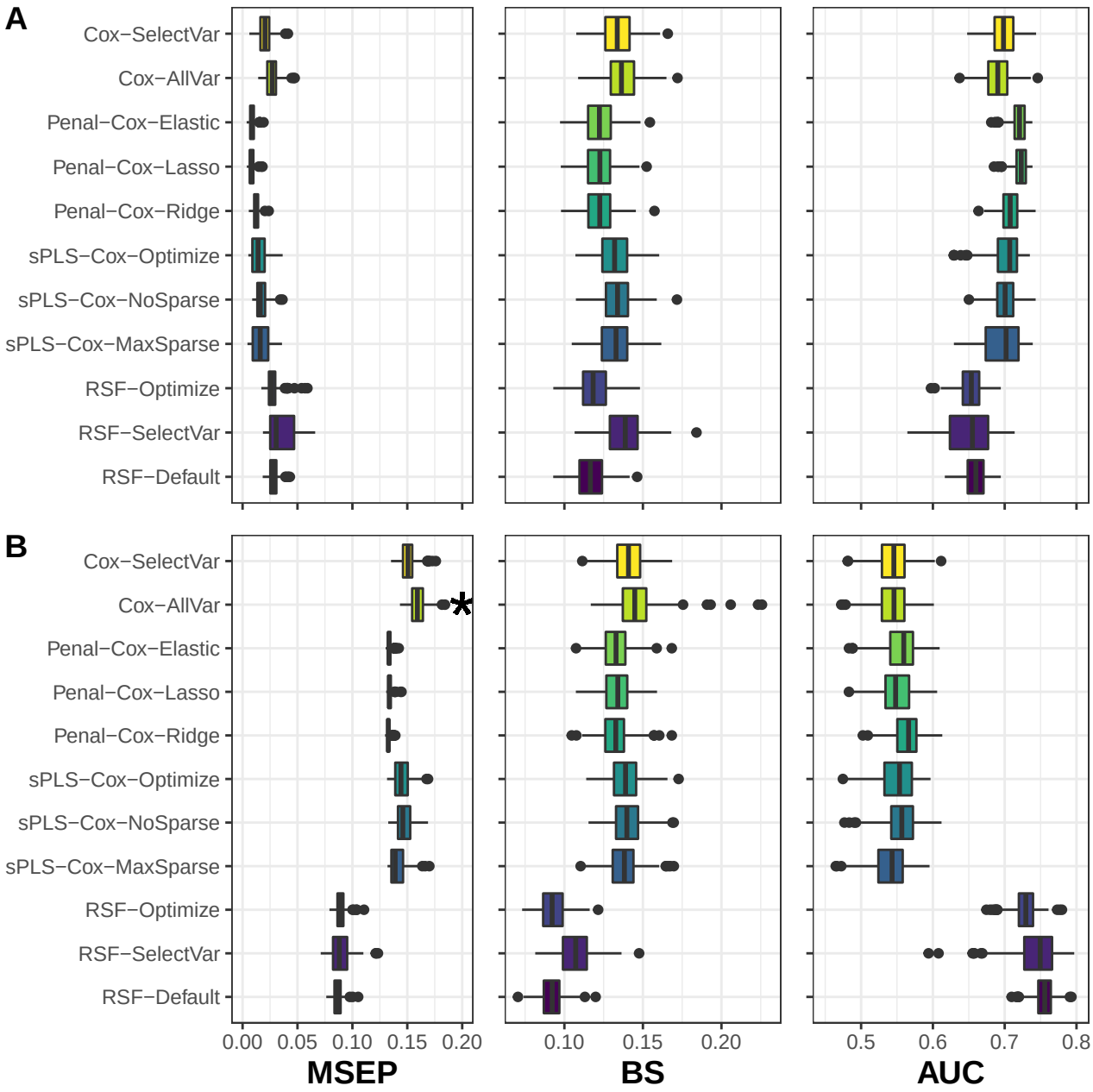

Figure 3: Simulation results over 250 replicates when considering 18 summaries associated to the event either assuming a linear form (figure A) or non-linear form (figure B). Methods are assessed using Mean Square Error of Prediction (MSEP), Brier Score (BS) and Area Under the ROC Curve (AUC). ($*$) symbol indicates the presence of MSEP values above 0.2, but not displayed.

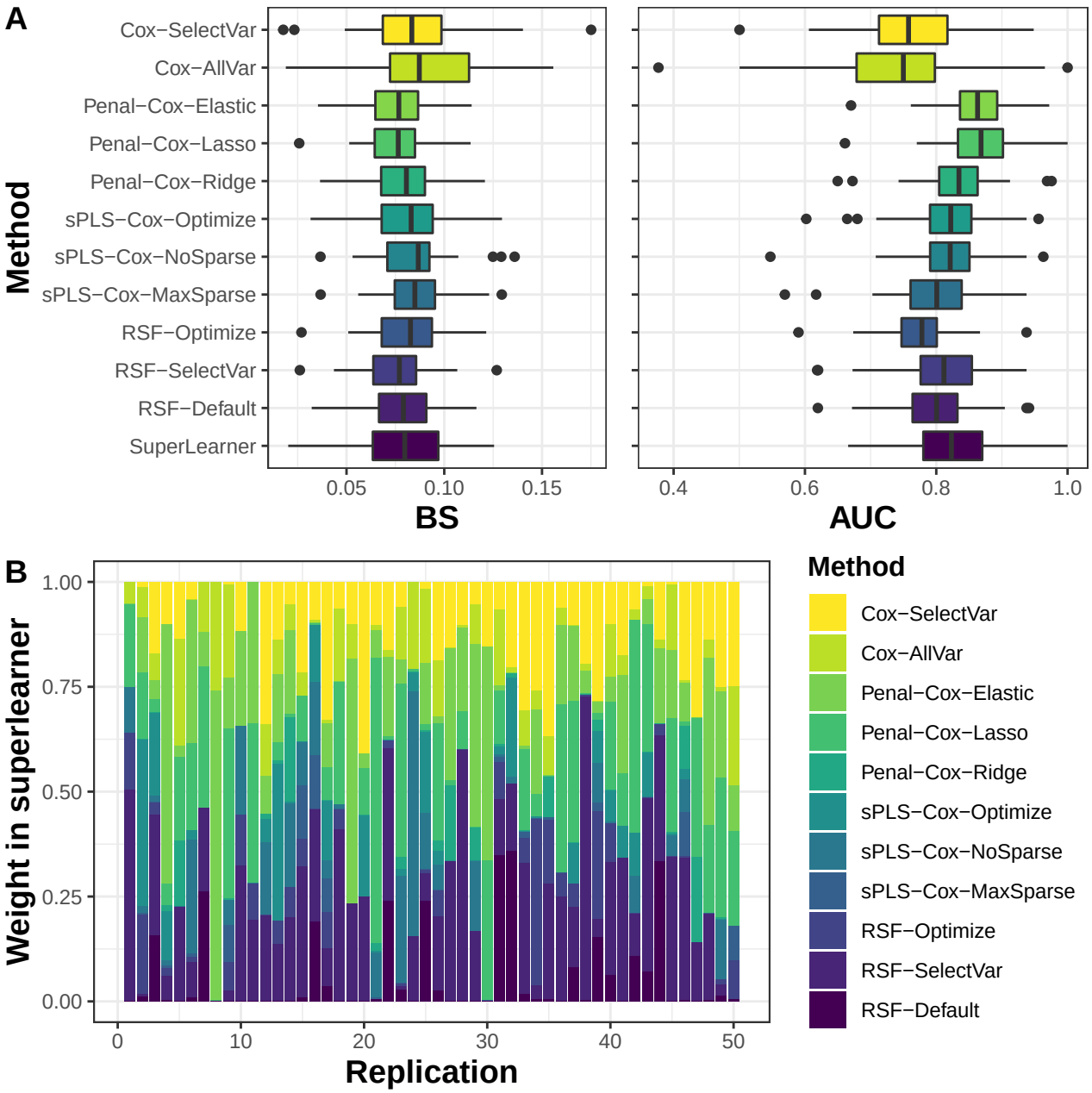

Figure 4: Assessment (figure A) and weights in superlearner (figure B) of 3-years death survival probability in primary biliary cholangitis patients using information collected up to 4 years over 50 replicates. Methods are assessed using Brier Score (BS) and Area Under the ROC Curve (AUC).

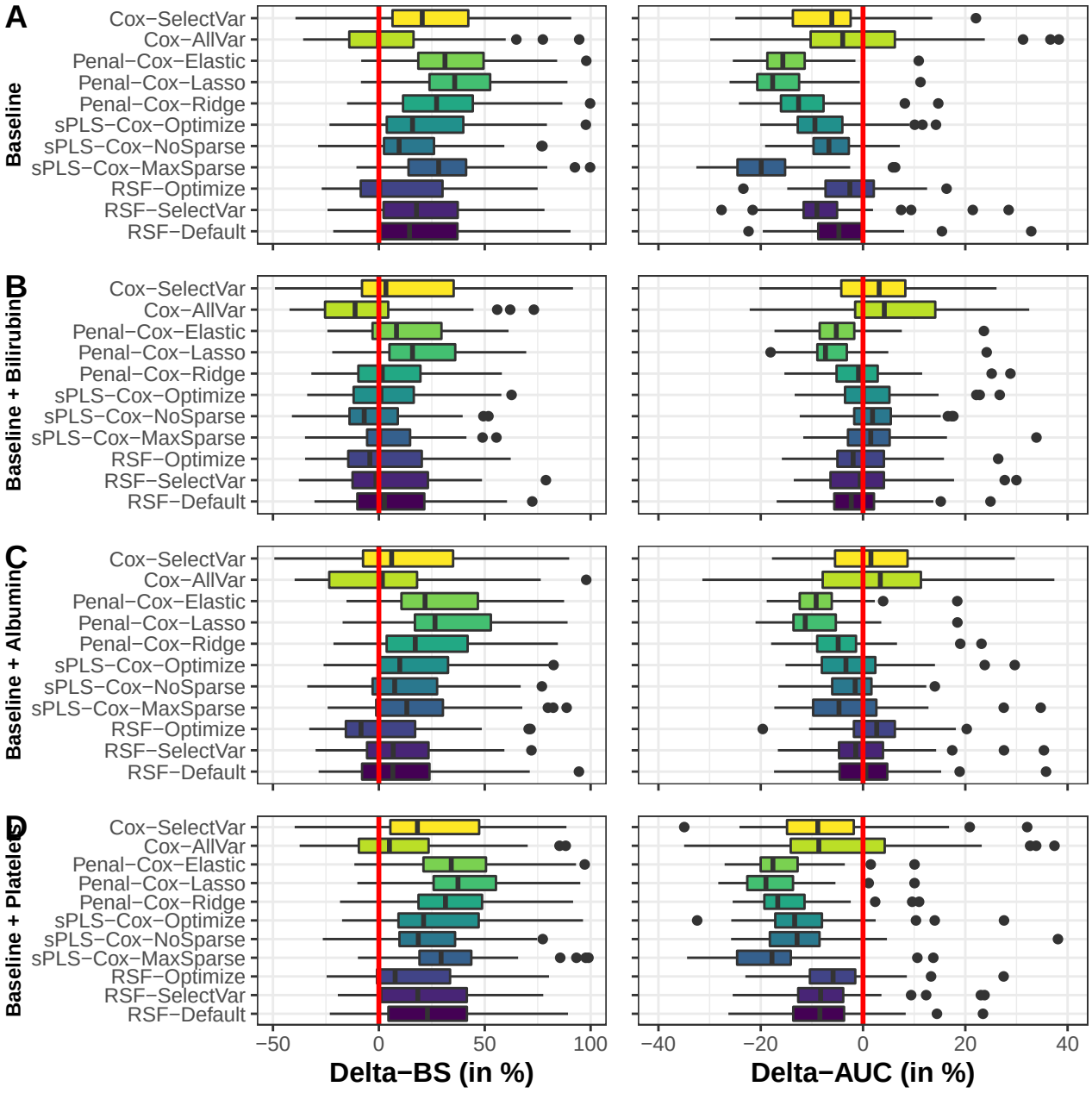

Figure 5: Assessment of 3-year survival probability in primary biliary cholangitis patients using baseline information on the 11 biomarkers and 3 covariates (figure A), baseline information and repeated measures collected up to 4 years of either serum bilirubin (figure B), albumin (figure C) or presence of platelets (figure D). The 10-fold cross-validation was replicated 50 times. The figure displays the difference (in percentage) of Brier Score (BS) and Area Under the ROC Curve (AUC) compared to the method using all the information with positive values for BS and negative values for AUC indicating a lower predictive accuracy.

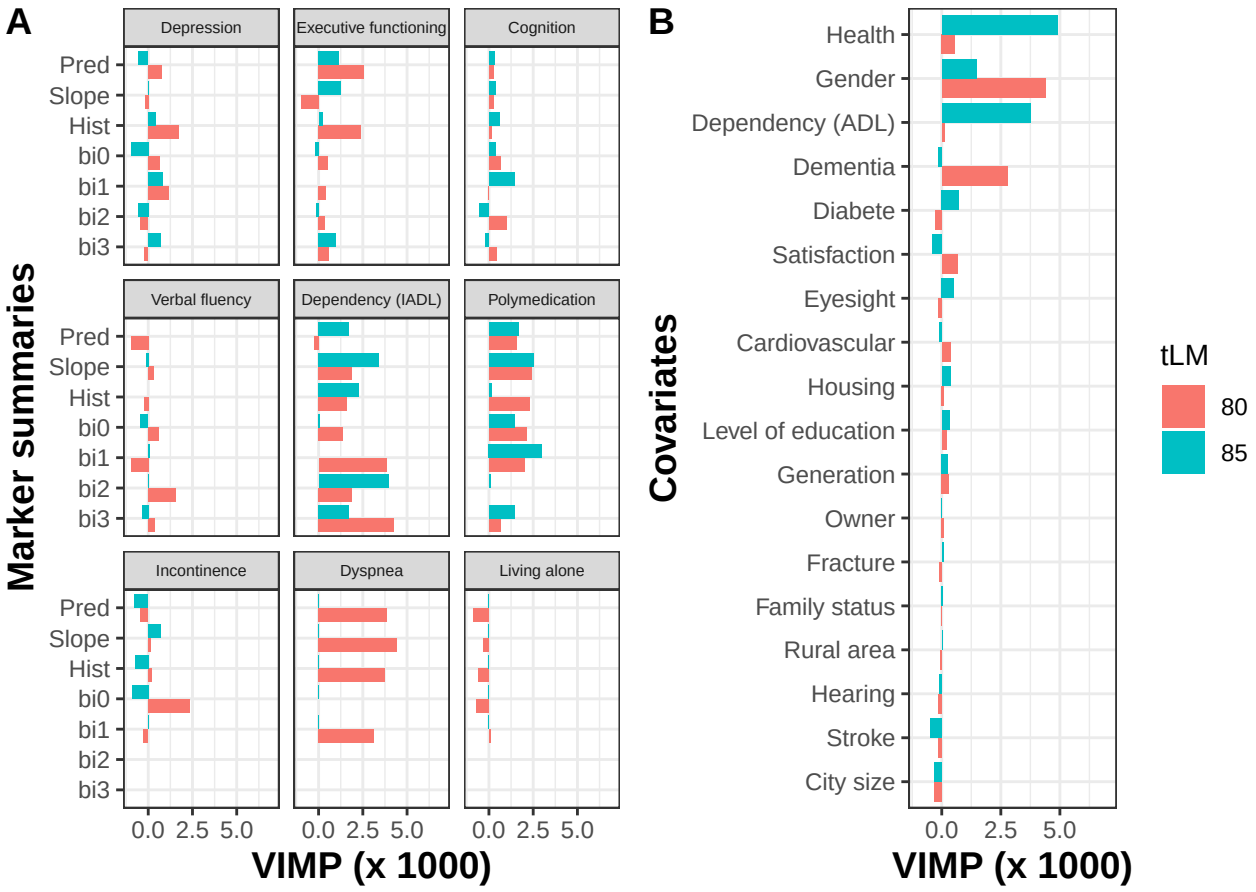

Figure 6: Variables associated with all-cause death in the *RSF-Optimize* model at 80-year and 85-year landmark age. Are displayed the VIMP value for each marker summaries (figure A) and covariate (figure B). A large VIMP value indicates that the variable is predictive of the event.