# Modélisation et prédiction dynamique individuelle d'événements de santé à partir de données longitudinales multivariées

Anthony Devaux

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

# DOCTEUR
# DE L'UNIVERSITÉ DE BORDEAUX

Ecole Doctorale Sociétés, Politique, Santé Publique

Spécialité Santé Publique, option Biostatistique

Par **Anthony DEVAUX**

## Modélisation et prédiction dynamique individuelle d'événements de santé à partir de données longitudinales multivariées

Sous la direction de : **Cécile PROUST-LIMA**
Co-directeur : **Robin GENUER**

Soutenue le 29 novembre 2022

Membres du jury :

| | | | |
|---|---|---|---|
| M. Yohann FOUCHER | Professeur des universités | Poitiers | Rapporteur |
| M. Robin GENUER | Maître de Conférence | Bordeaux | Co-directeur de thèse |
| Mme. Agathe GUILLOUX | Directrice de Recherche | Paris | Rapporteur |
| Mme. Cécile PROUST-LIMA | Directrice de Recherche | Bordeaux | Directrice de thèse |
| Mme. Virginie RONDEAU | Directrice de Recherche | Bordeaux | Présidente |
| M. Philippe SAINT-PIERRE | Professeur assistant | Toulouse | Examinateur |

# Remerciements

## A mes directeurs de thèse

**A Cécile.**

Je te remercierai jamais assez de m'avoir fait confiance et de m'avoir proposé de travailler à tes côtés pendant ces 3 ans. J'ai énormément appris grâce à toi, sûrement parce que tu connais tellement de choses, c'est assez incroyable d'ailleurs ! Tu as toujours été présente pour me relire, me corriger, me rerelire, me recorriger, à n'importe quelle heure du jour et de la nuit. Malgré ton emploi du temps plus que chargé, tu as toujours su trouver une place quand j'en avais besoin, et me remotiver dans les moments difficiles. Merci encore pour tout ce que tu as fais pour moi.

**A Robin.**

Tu as été mon guide durant cette thèse pour appréhender les différentes méthodes en grande dimension. Grâce à ta qualité d'enseignement, tu as su rendre simple ce qui au préalable était obscur. J'ai pris un grand plaisir à travailler avec toi et toujours dans la bonne humeur.

# Aux membres du jury

**A Virginie.**

Je te remercie de me faire l'honneur de présider ce jury de thèse.

**A Agathe Guilloux.**

Je vous remercie d'avoir accepté d'évaluer mes travaux de thèse. Votre expertise en grande dimension me permettra de progresser dans mon travail.

**A Yohann Foucher.**

Merci de me faire l'honneur de juger cette thèse. Votre grande connaissance en survie m'apportera sans aucun doute une vision nouvelle.

**A Philippe Saint-Pierre.**

Je vous remercie d'avoir accepté de faire partie mon jury de thèse.

# A mes collègues et amis

**Au bureau S157.**

**Les filles**, on a formé une bonne team dans ce bureau, dommage qu'on en est pas assez profité... En tout cas, j'espère que je n'ai pas été un collègue trop chiant à râler tout le temps !

**Tiphaine**, j'ai squatté tous les serveurs pendant ces 3 ans, tu as souvent du me détester ! Je t'ai ramené du thé au ginseng pour m'excuser mais visiblement ça n'a pas eu l'effet escompté ! Je te l'accorde c'est dégueulasse ! En tout cas, tu as remarqué que j'ai évolué pendant ces 3 ans où désormais j'accepte ton eau chaude !

**Kateline**, je suis triste que tu ne sois pas présente pour ma soutenance, mais tu es à New-York donc finalement c'est pas trop mal ! Merci de ton soutien durant les moments difficiles, ça fait toujours du bien de savoir qu'on n'est pas seul. La thèse est un long chemin et tu es bientôt au bout, courage ! Et dans le pire des cas tu sais que tu as également un avenir dans l'événementiel ;)

**Ariane**, quand je serai parti, tu pourras prendre ma chaise de bureau parce que c'est pas possible que tu puisses bosser avec ta chaise pourrie ! En tout cas, merci pour tous les petits gâteaux que tu as ramené, on s'est bien régalé !

**Lisa**, je ne sais pas si tu pourras lire ces remerciements, j'imagine que tu es en réunion... Tu m'as toujours impressionné à t'intéresser à tous les sujets possibles (même les plus pourries). Je suis persuadé que tu seras une grande chercheuse ! En tout cas, tu as tout fait pour me réconcilier avec les étudiants de médecine, mais je crois que le fossé est trop grand !

**Au bureau S158.**

**Manel**, **Léonie**, **Quentin**, **Maris**, okay j'ai eu une illumination pendant que j'écrivais les remerciements. Je me disais qu'on n'a pas passé assez de temps ensemble et que c'était bien dommage... Mais en fait, quelqu'un peut m'expliquer pourquoi personne a eu l'idée d'inverser la salle de réunion et votre bureau ? ? ? On aurait pu pété le mur et faire un bureau commun nan ? A noter pour la prochaine RLP !

**A Gaëlle.**

C'était sympa les pauses cafés en début de thèse ! Dommage que le Covid est passé par la... bon après ça nous a sûrement permis de bosser plus ! Bon courage pour la fin de thèse !

**A l'équipe Biostatistique.**

Un grand merci à tous mes collègues de l'équipe Biostatistique pour votre sympathie et la qualité de vos remarques durant nos réunions d'équipe. Je remercie plus particulièrement *Sandrine* pour toute ton aide pendant les 5 dernières années.

**A Boris.**

Tu m'as pris sous ton aile pendant mes années de master et tu m'as convaincu de me lancer dans cette grande aventure qu'est la thèse. Dans les moments difficiles, je dois t'avouer que je t'ai détesté, mais finalement, je te suis reconnaissant de m'avoir pousser dans cette voie.

**A Rodolphe.**

A travers ta présidence de l'École Universitaire de Recherche Digital Public Health, je te remercie d'avoir rendu possible la réalisation de ma thèse. Un grand merci également pour avoir évaluer mes travaux au cours des comités de suivi de thèse et pour tes précieux conseils.

## Aux serveurs de l'ISPED

A **akouda**, **bizerte**, **kasserine**, **madhia** et **septimi**. Vous avez été mes plus fidèles compagnons pendant toute cette thèse. Je vous ai exploité pendant des heures, des jours, des semaines, des mois... et pourtant vous n'avez jamais failli. Les rumeurs dises que vous êtes obsolètes, pourtant cette thèse n'aurait jamais été possible sans vous. Vous allez être très probablement être débranché dans les semaines qui viennent, mais votre mémoire restera à jamais présente à travers ces remerciements.

## A mes parents

Je vous remercie d'avoir été présents à un tournant de ma vie. Vous n'avez pas hésité à me faire confiance dans mon projet de reprise d'études, malgré les incertitudes. Si j'en suis là aujourd'hui, c'est en grande partie grâce à vous !

## A ma femme

Mon baby, mon amour, depuis le début, tu as toujours cru en moi. Tu as été présente à chaque moment pour me remotiver. Tu m'as rendu la vie plus simple et tu m'as supporté dans les périodes de stress. Je sais que je n'ai pas toujours été facile, et je m'en excuse... J'ai maintenant le reste de notre vie pour me faire pardonner ! Maintenant on va pouvoir reprendre nos aventures et avoir *La vie qu'on a décidé de mener !* Je t'aime tellement <3

# Valorisations scientifiques

## Publications

- Devaux A., Genuer R., Peres K. and Proust-Lima C. (2022). Individual dynamic prediction of clinical endpoint from large dimensional longitudinal biomarker history : a landmark approach. *BMC Medical Research Methodology*, 22(1) :1–14
- Devaux A., Helmer C., Genuer R., and Proust-Lima C. (2022). Random survival forests for competing risks with multivariate longitudinal endogenous covariates. *Submitted in Biostatistics*
- Devaux A., Genuer R. and Proust-Lima C. (2022). Random Forest with longitudinal irregularly measured predictors : The `DynForest` R package. *To be submitted to Journal of Statistical Software*
- Bercu A., Devaux A., Helmer C., Dufouil C., Proust-Lima C. and Jacqmin-Gadda H. (2022). Identify risk factors of dementia with competing landmark approach. *To be submitted*
- De Courson H., Devaux A., Derot S., d'Auzac A., Marnat G., Verchère E., Proust-Lima C. and Biais M. (2022). Dynamic prediction of cerebral vasospasm after subarachnoid hemorrhage in neuro-intensive care unit. *In preparation*

## Packages R

- Devaux A. (2022). `hdlandmark` : *Dynamic Predictions with Multivariate Longitudinal Predictors by Landmark Approach*. GitHub development version. https://github.com/anthonydevaux/hdlandmark
- Devaux A. (2022). `DynForest` : *Random Forest with Multivariate Longitudinal Predictors*. R package version : 1.0.0. https://CRAN.R-project.org/package=DynForest

## Présentations orales en conférence

- Devaux A., Genuer R. and Proust-Lima C., *Dynamic modelling and prediction of health events from multivariate longitudinal data*, GDR Statistique et Santé, Paris, France, 2020 (en ligne).
- Devaux A., Genuer R., Peres K. and Proust-Lima C., *Individual dynamic prediction of clinical endpoint from large dimensional longitudinal biomarker history : a landmark approach*, 8ème Channel Network Conference, Paris, France, 2021 (en ligne).
- Devaux A., Genuer R., Peres K. and Proust-Lima C., *Individual dynamic prediction of clinical endpoint from large dimensional longitudinal biomarker history : a landmark approach*, 42nd Annual Conference of the International Society for Clinical Biostatistics, Lyon, France, 2021 (en ligne).
- Devaux A., Helmer C., Dufouil C., Genuer R., and Proust-Lima C., *Random survival forests for competing causes with multivariate longitudinal endogenous covariates*, Intelligence Artificielle et santé : approches interdisciplinaires, Nantes, France, 2022.
- Devaux A., Helmer C., Dufouil C., Genuer R., and Proust-Lima C., *Random survival forests for competing causes with multivariate longitudinal endogenous covariates*, 43rd Annual Conference of the International Society for Clinical Biostatistics, Newcastle, United Kingdom, 2022.

## Communication invitée en séminaire

- Devaux A., Genuer R., Peres K. and Proust-Lima C., *Individual dynamic prediction of clinical endpoint from large dimensional longitudinal biomarker history : a landmark approach*, Biostatistic seminar, University of Waterloo, Canada, 2022

## Récompense scientifique

- Student Conference Award of the 2022 International Society for Clinical Biostatistics Conference, Newcastle, United Kingdom, 2022

# Table des matières

# Chapitre I

# Introduction

## Sommaire

## I.1 Médecine personnalisée

La médecine personnalisée est, par définition, une médecine basée sur les caractéristiques individuelles du patient, dans le but de prédire au mieux un évènement clinique (récidive du cancer, décès, ...) pour adapter la prise en charge. La philosophie de la médecine individuelle est donc d'aller plus loin que l'utilisation de simples données sociodémographiques du patient (comme l'âge et le sexe), en intégrant en plus d'autres sources de données collectées au cours du suivi. Parmi le type de données à inclure en plus, il existe notamment les données cliniques (température, fréquence cardiaque, pression artérielle, ...) ou les données biologiques (glycémie, natrémie, cholestérol, ...). Des données, plus spécifiques à certaines maladies, peuvent également être considérées comme par exemple, les données d'imagerie (taille d'une tumeur, d'un organe, ...) ou des données issues d'échelles de mesure (tests cognitifs, échelles psychologiques, ...).

La médecine personnalisée permet désormais de fournir une prédiction du risque d'un évènement clinique dans un futur proche. Ainsi, plusieurs scores ont été développés pour divers évènements cliniques. Parmi les nombreux exemples issus de la littérature, nous retrouvons des scores pour calculer le risque de mortalité chez les adultes en unité de soins intensifs [Zimmerman et al., 2006] ou suite à une hémorragie traumatique [Perel et al., 2012], le risque de maladie cardio-vasculaire [Wilson et al., 1998, American Heart Association, 2013, Hippisley-Cox et al., 2017] ou encore le risque de progression de maladie chronique rénale [Tangri et al., 2011]. Certains scores présentés dans la littérature ont ensuite été implémentés dans des applications web [Zimmerman et al., 2006, Hippisley-Cox et al., 2017, Tangri et al., 2011] pour une utilisation plus simple en routine.

Ces scores n'ont pas à vocation d'automatiser une quelconque décision à la place du corps médical, mais de fournir des outils supplémentaires pour une aide à la décision. Dans le cadre des patients suivis en unité de soins intensifs, quantifier le risque de mortalité à court terme permet de stratifier les patients pour, par exemple, renforcer la surveillance de ceux ayant un risque de décès élevé. Concernant la progression d'une maladie sur un plus

long terme, l'estimation du risque peut permettre d'orienter les stratégies thérapeutiques à mettre en place, notamment en administrant un traitement alternatif pour les patients à fort risque.

Pour une même maladie, le risque peut être quantifié à partir de divers modèles de prédiction qui se différencient par la méthode statistique utilisée et par les données incluses. Pour définir quel est le meilleur modèle parmi les divers modèles proposés, la qualité de prédiction peut être évaluée au moyen de plusieurs critères [Steyerberg et al., 2010]. Bien souvent, seule la discrimination du modèle est évaluée, autrement dit si le modèle performe à identifier les malades des non malades. Mais la qualité de prédiction peut également être évaluée en terme de calibration, en d'autres mots, si le risque prédit est proche du risque attendu.

Pour avoir des modèles prédictifs toujours plus performants, il est important d'inclure les variables les plus pertinentes possibles. En particulier, les données collectées au cours du temps peuvent se révéler essentielles pour quantifier le risque de certaines maladies.

## I.2 Définition des prédictions dynamiques

La plupart des modèles prédictifs utilisent uniquement des données dites transversales (récoltées à un seul temps). Cependant, l'incorporation de données longitudinales (mesurées à plusieurs temps) est désormais possible. Ces multiples mesures peuvent notamment être collectées dans le cadre d'un suivi régulier ou par un enregistrement électronique des données [Birkhead et al., 2015] lors du monitorage des patients. L'introduction de données longitudinales dans les modèles de prédiction s'est révélée pertinente dans le cadre de certaines maladies, par exemple avec l'antigène prostatique spécifique pour prédire la récidive du cancer de la prostate [Proust-Lima and Taylor, 2009, Taylor et al., 2013] ou encore du glucose et de l'hémoglobine pour prédire la survenue du diabète [Parast et al., 2019].

La figure I.1 introduit le concept de prédiction dynamique dans le cadre d'une seule

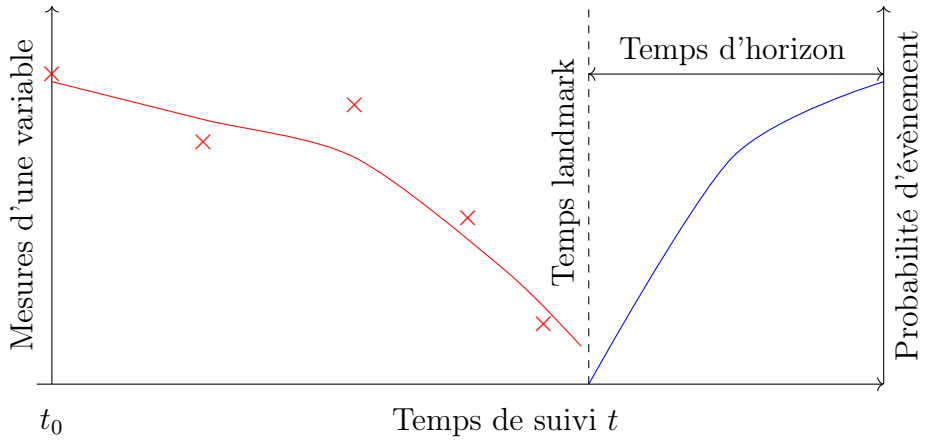



FIGURE I.1 – Illustration d'une prédiction dynamique individuelle. Les données (croix rouges) sont collectées au cours du temps jusqu'à un temps *landmark* (ligne pointillée) puis modélisées (courbe rouge). Ces informations peuvent ainsi être utilisées pour calculer la probabilité de survenue d'un évènement (courbe bleue) au cours d'un temps d'horizon.

variable explicative répétée au cours du temps. La probabilité de survenue d'un évènement est calculée à partir d'un temps appelé *landmark* jusqu'à un certain horizon en utilisant les données collectées jusqu'à ce temps *landmark*. Dans les études, la prise en compte de données longitudinales est souvent très simpliste avec uniquement la dernière mesure utilisée [Keogh et al., 2019]. En effet, l'utilisation de la dernière mesure entraîne deux problèmes en ne prenant pas en compte : (i) l'erreur de mesure ; (ii) le délai entre la dernière mesure et le temps *landmark*. Pour éviter ces problèmes, la dynamique de la variable longitudinale peut être modélisée [Proust-Lima and Taylor, 2009] pour calculer des résumés tels que la valeur à un temps donné, l'accélération à un temps donné ou une différence entre les mesures du patient par rapport à la moyenne.

Plusieurs méthodes ont été proposées pour prédire le risque d'un évènement en fonction d'une variable mesurée à différents temps et avec erreur : les méthodes en deux étapes (par *regression calibration* [Ye et al., 2008] ou *landmark* [Van Houwelingen, 2007]) et les modèles conjoints [Proust-Lima and Taylor, 2009, Tsiatis and Davidian, 2004, Rizopoulos, 2012]. Dans les méthodes en deux étapes, la variable longitudinale est modélisée par des méthodes adaptées aux données répétées, le plus souvent des modèles mixtes [Laird and Ware, 1982], pour obtenir des résumés de la dynamique de la variable (étape 1). Ces

résumés sont ensuite inclus pour modéliser le risque d'évènement [Cox, 1972] (étape 2). Les méthodes en deux étapes se différencient entre elles par la quantité d'information utilisée. Dans la méthode de *regression calibration*, toutes les données sont utilisées, en particulier l'information postérieur à l'évènement, pouvant induire un biais dans l'estimation de la prédiction du risque [Albert and Shih, 2010]. L'approche *landmark* a ainsi été développée pour éliminer ce biais au détriment d'une perte de puissance statistique, en n'utilisant que les données disponibles jusqu'au temps *landmark* pour les individus encore à risque à ce temps. Les modèles conjoints, quant à eux, réalisent ces deux étapes simultanément et permettent ainsi de mieux estimer le risque d'évènement, mais sont plus complexes d'utilisation [Ferrer et al., 2019].

Les développements statistiques décrits précédemment se sont limités le plus souvent à une seule variable longitudinale. Pourtant, dans de nombreuses situations, l'utilisation d'une seule variable longitudinale ne permet pas de prédire au mieux l'évènement. Pour aller plus loin, il peut alors être pertinent d'intégrer de multiples variables longitudinales, dont certaines peuvent être potentiellement prédictives.

## I.3 Dimension des données

Dans le cas où nous souhaitons inclure un grand nombre de variables (et notamment longitudinales), le problème de dimensionnalité émerge. La dimensionnalité se définit par le nombre de variables d'un jeu de données. Dans un contexte de grande dimension, deux facteurs sont à prendre en compte : (i) la possible forte corrélation entre les différentes variables (ii) la parcimonie des observations dans un espace très grand. Ce concept est appelé *fléau de la dimension* et traduit bien le problème qui existe pour exploiter une quantité d'information toujours plus importante. Ce concept est omniprésent dans certains domaines de la santé, en particulier, en génétique où des milliers de variables peuvent être utilisées [Yu et al., 2017].

Pour résoudre le problème de dimensionnalité, trois techniques ont été proposées :

(i) les méthodes de réduction de dimension [Bastien et al., 2015] pour capter le plus d'informations possible sur un espace de plus petite dimension ; (ii) les méthodes de régularisation [Tibshirani, 1997, Zou and Hastie, 2005] où une contrainte sur les paramètres est fixée à l'aide d'une pénalité ; (iii) les méthodes non paramétriques [Ishwaran et al., 2008] ne s'appuyant pas sur des hypothèses *a priori* de distributions. Néanmoins, toutes ces méthodes ont été développées à partir de données transversales. Leur extension aux données longitudinales est actuellement un domaine de recherche car l'intégration de ces données n'est pas directe.

## I.4 Objectif de la thèse

L'objectif de cette thèse est le développement de nouvelles méthodologies pour prédire un évènement de santé à partir de multiples données longitudinales. Pour répondre à cette problématique, nous proposons deux travaux combinant l'association entre modèles mixtes, pour modéliser les dynamiques des variables longitudinales, et méthodes de survie en grande dimension, pour prédire le risque d'un évènement de santé. Ces deux travaux ont été développés pour des données de survie censurées à droite dans un contexte de risques compétitifs.

Le chapitre 2 présente un état de l'art de l'ensemble des notions utilisées dans cette thèse. Nous introduisons les données de survie dans le cas général, dans le contexte des risques compétitifs puis les extensions développées dans le cadre de données en grande dimension. Ensuite, nous présentons les modèles mixtes pour l'analyse de données longitudinales, puis en association avec les données de survie. Enfin, nous introduisons la notion de prédiction dynamique individuelle et son évaluation.

Le chapitre 3 est composé de deux parties. La première partie introduit un développement méthodologique où les variables longitudinales et les données de survie sont modélisées en deux étapes au moyen d'une approche *landmark*. L'approche *landmark* consiste à se focaliser sur un temps de suivi et développer l'outil de prédiction sur les sujets à risque

d'évènement à ce temps. Des caractéristiques provenant des variables longitudinales sont incluses dans plusieurs modèles de survie adaptés à la grande dimension pour prédire l'évènement de santé. Des simulations ont été effectuées pour comparer les différentes méthodes de prédiction dans un cadre de petite et moyenne dimension avec des associations plus ou moins complexes. Puis, la méthode a été illustrée dans deux applications : (i) pour prédire le risque de décès chez les patients atteints de la cholangite biliaire primitive ; (ii) pour prédire le risque de décès chez les personnes âgées en population générale. Dans une deuxième partie, nous introduisons l'extension de notre méthode *landmark* dans un contexte de risques compétitifs.

Le chapitre 4 est ordonné en trois sous-sections. La première section présente une extension des forêts aléatoires en survie dans laquelle les prédicteurs peuvent être des données longitudinales de marqueurs endogènes (expliqués par d'autres variables). Cette méthode a été comparée à des méthodes concurrentes dans une étude de simulation. Puis, cette méthode a été illustrée pour prédire la survenue de la démence en prenant en compte le décès en tant que risque competitif. La deuxième partie introduit le package R `DynForest` permettant d'utiliser cette méthode à travers plusieurs exemples. Enfin, la troisième partie illustre une autre application de cette méthode, pour prédire la survenue du vasospasme chez les patients ayant subi une hémorragie sous-arachnoïdienne.

Le dernier chapitre de la thèse est consacré à une discussion concernant les avantages et les limites des deux méthodes proposées, ainsi qu'aux possibles perspectives à la suite de cette thèse.

# Chapitre II

# État de l'art

**Sommaire**

INTRODUCTION

Ce chapitre introduit les différents concepts utilisés dans le reste de la thèse. Nous abordons le principe des données de survie, en particulier dans un contexte de grande dimension. Ensuite, nous présentons le principe des données longitudinales et comment elles peuvent être couplées aux données de survie. Enfin, nous définissons qu'est-ce qu'un outil de prédiction, de son développement jusqu'à son évaluation.

## II.1 Analyse des données de survie

Dans le domaine de la santé publique, l'étude d'un évènement clinique est nécessaire pour en comprendre les mécanismes, comme par exemple la survenue du cancer, de la démence ou du décès. Le temps d'apparition de l'évènement associé à sa survenue, en cas de censure, définit les données de survie.

### II.1.1 Définition de la survie dans un contexte de données censurées

Pour l'analyse des données de survie, nous allons nous intéresser au temps de suivi $T$ pour un individu jusqu'à la survenue d'un évènement, par exemple le décès. Ce temps $T$ peut être défini soit par : (i) le temps depuis la naissance de l'individu ; (ii) le temps depuis l'entrée dans l'étude de l'individu ; (iii) le temps depuis une date fixe commune à tous les individus. Dans les études de cohorte, où la durée de suivi est limitée dans le temps, l'évènement n'est pas toujours observable. Cela peut notamment se produire lorsqu'un individu sort de l'étude prématurément, de façon volontaire ou non, également appelé perdu de vue. Les données sont alors censurées au temps de dernières nouvelles $C$ (communément appelée censure à droite). Lorsque le temps de censure $C$ est indépendant de l'évènement, autrement dit que la dégradation du patient n'est pas liée au fait que les

données du patient sont observables, les données de survie sont définies par :

$$(\widetilde{T}, \delta) \text{ avec } \begin{cases} \widetilde{T} = \min(T, C) \\ \delta = \mathbb{1}_{(T \leq C)} \end{cases} \tag{II.1}$$

où $\widetilde{T}$ représente le temps observé et $\delta$ l'indicatrice d'évènement.

Le principe des données de survie est illustré dans la figure II.1, où l'évènement n'est pas observé pour les individus 1 et 2 aux dates de dernières nouvelles $C$ contrairement aux individus 3 et 4 où l'évènement est apparu au temps $T$.

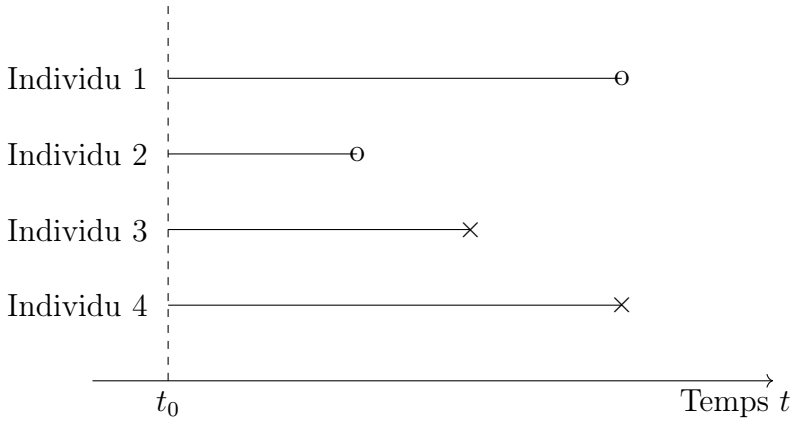



FIGURE II.1 – Illustration des données de survie dans le cas d'une censure à droite. Chaque ligne horizontale représente la durée de suivi pour un individu $i$. Une croix indique que l'évènement est survenu ($\delta_i = 1$) au temps $T_i$. Un rond blanc signifie une absence d'évènement ($\delta_i = 0$) avant le dernier temps $C_i$ observé, les données sont dites censurées.

Pour étudier le temps de survie $T$, la fonction de survie $S(t)$ est couramment utilisée. Il s'agit de la probabilité de ne pas subir l'évènement avant un temps $t$, d'où la probabilité de survie lorsque l'évènement est le décès. La fonction de survie se définie par :

$$\begin{aligned} S(t) &= P(T > t) \\ &= \exp(-\Lambda(t)) \\ &= \exp\left(-\int_0^t \lambda(u)du\right) \end{aligned} \tag{II.2}$$

où $\Lambda(t)$ représente la fonction de risque cumulé et $\lambda(u)$ la fonction de risque instantané de l'évènement. La fonction de survie est une probabilité, donc à valeurs entre 0 et 1, avec

$S(0) = 1$. La fonction de répartition $F(t)$ peut également être définie à partir de la fonction de survie par :

$$\begin{aligned} F(t) &= 1 - S(t) \\ &= P(T \leq t) \\ &= \int_0^t f(u)du \end{aligned} \tag{II.3}$$

avec $f(t)$ est la fonction de densité égal à :

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \tag{II.4}$$

où $\Delta t$ un petit interval de temps.

### II.1.2 Estimation de la fonction de survie par Kaplan-Meier

L'analyse de la fonction de survie $S(t)$ est très utile, car en plus d'être simple à comprendre (i.e. la probabilité d'être vivant en $t$), cette fonction peut être facilement représentée graphiquement. La fonction de survie peut être estimée de plusieurs manières en utilisant soit une méthode paramétrique (suivant une loi exponentielle ou de Weibull par exemple) ou non-paramétrique. Parmi les estimateurs non-paramétriques, l'estimateur de Kaplan-Meier [Kaplan and Meier, 1958] est couramment utilisé et est adapté aux données censurées à droite. Son estimateur $\widehat{S}(t)$ et sa variance $\widehat{\sigma}^2(t)$ sont définis par :

$$\widehat{S}(t) = \prod_{j:\widetilde{T}_j \leq t} \left(1 - \frac{d_j}{n_j}\right) \text{ et } \widehat{\sigma}^2(t) = [\widehat{S}(t)]^2 \sum_{j:\widetilde{T}_j \leq t} \frac{d_j}{n_j(n_j - d_j)} \tag{II.5}$$

avec $\widetilde{T}_j$ les temps d'évènement observés (i.e. non censurés), $n_j$ et $d_j$ le nombre d'individus à risque et subissant l'évènement en $\widetilde{T}_j$, respectivement.

L'estimateur de Kaplan-Meier $\widehat{S}(t)$ est une fonction en escalier décroissante, avec un *saut* à chaque temps d'évènement $\widetilde{T}_j$, et donc constante entre deux temps d'évènement $\widetilde{T}_j$. Cet estimateur est notamment utile pour trouver la médiane de survie estimée $S(t_M) = 0.5$, qui correspond au temps $t_M$ où 50% des individus sont encore à risque.

### II.1.3 Estimation de la fonction de risque cumulé par Nelson-Aalen

Une estimation de la fonction de risque cumulé $\Lambda(t)$ a été proposée par Nelson-Aalen [Nelson, 1969, Aalen, 1976]. Cet estimateur, non-paramétrique, permet de prendre en compte la censure des données par la relation suivante :

$$\widehat{\Lambda}(t) = \sum_{j:\widetilde{T}_j \leq t} \frac{d_j}{n_j} \tag{II.6}$$

Sachant que la fonction de risque cumulé $\Lambda(t)$ est liée à la fonction de survie $S(t)$ par la relation $S(t) = \exp(-\Lambda(t))$, une nouvelle estimation de $S(t)$ est possible.

### II.1.4 Test du *log-rank*

La fonction de survie $S(t)$ peut être définie dans plusieurs groupes, par exemple en fonction du sexe des sujets. Dans le cas d'une variable continue, un seuil peut être déterminé pour définir les groupes, par exemple en classes d'âge. Dans le cas de deux groupes, $S_1(t)$ et $S_2(t)$ désignent les fonctions de survie pour le groupe de sujet 1 et 2, respectivement. La différence de survie peut être testée entre ces deux groupes à l'aide du test du *log-rank* [Harrigton and Fleming, 1982] où les hypothèses sont :

$$\begin{cases} H_0 : S_1(t) = S_2(t) \forall t \\ H_1 : S_1(t) \neq S_2(t) \text{ pour au moins un } t \end{cases} \tag{II.7}$$

Le test du *log-rank* compare le nombre d'évènements observés dans chaque groupe au nombre d'évènements attendus sous l'hypothèse $H_0$. La statistique de test suit une loi du $\mathcal{X}^2$ (chi-deux) à un degré de liberté. En plus de la statistique de test, il est important d'observer l'évolution des fonctions de survie dans chacun des groupes, en particulier si elles sont amenées à se croiser. En effet, cela peut entraîner une augmentation du risque de second espèce (i.e. de ne pas rejeter l'hypothèse $H_0$ alors que $H_1$ est vraie).

### II.1.5 Modélisation des données de survie par un modèle de Cox

Pour étudier les facteurs de risque associés à la survenue d'un évènement clinique, il est possible de modéliser la fonction de risque $\lambda(t)$ à l'aide des modèles à risques proportionnels. Parmi les modèles à risques proportionnels, le modèle de Cox [Cox, 1972] est le plus couramment utilisé et est défini par :

$$\lambda(t|X) = \lambda_0(t) \exp\left(X^\top \beta\right) \tag{II.8}$$

avec $\lambda_0(t)$ la fonction de risque de base commune à l'ensemble des individus et laissée non-spécifiée. $X$ est un ensemble de prédicteurs et $\beta$ les coefficients associés à $X$. Les coefficients $\beta$ sont estimés en maximisant la vraisemblance partielle $\mathcal{L}(\beta, X)$ calculée en chaque temps d'évènement à partir des $N$ individus par :

$$\mathcal{L}(\beta, X) = \prod_{i=1}^{N} \frac{\exp\left(X_i^\top \beta\right)}{\sum_{l:\widetilde{T}_l \geq t_i} \exp\left(X_l^\top \beta\right)} \tag{II.9}$$

Lorsque qu'il existe plusieurs temps d'évènements identiques, la vraisemblance peut être approchée par l'approximation de Breslow [Breslow, 1974].

Le modèle de Cox est très populaire, notamment en épidémiologie, par sa facilité d'interprétation. En effet, le changement du risque pour l'augmentation d'une unité pour une variable $p$ donnée est appelée risque relatif, et peut être facilement calculé par :

$$\frac{\lambda(t|X_p = x+1)}{\lambda(t|X_p = x)} = \exp\left(\beta\right). \tag{II.10}$$

En revanche, ce modèle fait l'hypothèse de proportionnalité des risques au cours du temps, qu'il est nécessaire de vérifier pour valider le modèle. En effet, $\exp\left(\beta\right)$ n'est pas dépendant du temps $t$, ou en d'autres termes, que le risque est donc constant en tout temps. Cette hypothèse est vérifiable notamment par l'analyse des résidus [Kalbfleisch and Prentice, 2011].

### II.1.6 Extension aux risques compétitifs

En pratique, l'analyse de survie ne se limite pas à un unique évènement. En effet, les sujets sont susceptibles d'être à risque de plusieurs évènements à la fois. La compétition des risques signifie qu'il existe un autre évènement pouvant empêcher la survenue de l'évènement d'intérêt (voir figure II.2), au contraire des modèles multi-états [Andersen and Keiding, 2002] où la transition entre les différents évènements est possible. Lors de l'étude d'un évènement clinique d'intérêt comme le décès par exemple, les différentes causes se trouvent en compétition (ou concurrence), puisque si le sujet décède pour une cause comme le cancer, il ne pourra pas décéder d'une autre cause. Plus largement, les risques compétitifs peuvent se définir par la survenue du premier évènement. Par exemple dans le cas de la démence, qui survient avant le décès sans démence. Dans le cas de risques compétitifs, le couple $(\widetilde{T}, \delta)$ est défini par :

$$(\widetilde{T}, \delta) \text{ avec } \begin{cases} \widetilde{T} = \min(T, C) \\ \delta = k * \mathbb{1}_{(T \leq C)} \end{cases} \tag{II.11}$$

avec $\delta = k$ si l'individu subi l'évènement $k \in \{1, ..., K\}$ avant le temps de censure $C$, sinon $\delta = 0$.

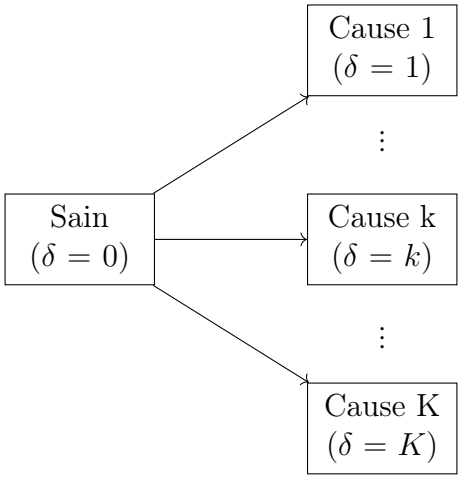



FIGURE II.2 – Illustration de la compétition des risques en survie. Tous les sujets sont considérés comme sain ($\delta = 0$) au début de l'étude, jusqu'à la survenue d'une cause $k \in \{1, ..., K\}$ ($\delta = k$) au temps $\widetilde{T}$.

Dans le contexte d'évènements compétitifs, la fonction de risque instantané $\lambda(t)$ est modélisée par :

$$\lambda(t) = \sum_{k=1}^{K} \lambda_k(t) \tag{II.12}$$

où $\lambda_k(t)$ est la fonction de risque instantané pour chaque cause $k$.
De plus, la fonction d'incidence cumulée $I_k(t)$, désignant la probabilité de survenue de l'évènement $k$ avant $t$, peut également être défini par :

$$\begin{array}{rcl} I_k(t) & = & P(T \leq t, \delta = k) \\ & = & \int_0^t \lambda_k(u) S(u) du \end{array} \tag{II.13}$$

où $S(u)$ est la survie sans évènement caractérisée par :

$$\begin{array}{rcl} S(t) & = & \exp\Big( - \sum_{k=1}^{K} \Lambda_k(t) \Big) \\ & = & \exp\Big( - \sum_{k=1}^{K} \int_0^t \lambda_k(u) du \Big) \end{array} \tag{II.14}$$

avec $\Lambda_k(t)$ la fonction de risque cumulé, pouvant être approchée par l'estimateur de Aalen-Johansen [Aalen and Johansen, 1978].

Pour étudier un effet de variables explicatives sur le risque d'évènement, plusieurs modèles semi-paramétriques ont été proposés, dont le modèle cause-spécifique [Prentice et al., 1978] et le modèle de Fine & Gray [Fine and Gray, 1999] décrits dans la suite.

#### II.1.6.1 Modèle cause-spécifique

Le modèle cause-spécifique décline le modèle à risque proportionnel pour chaque cause $k$ par :

$$\lambda_k(t|X) = \lambda_{k,0}(t) \exp\left(X^\top \beta_k\right) \tag{II.15}$$

avec $\lambda_{k,0}(t)$ la fonction de risque instantané de base de la cause $k$ et $\beta_k$ les coefficients spécifiques à la cause $k$, associés aux prédicteurs $X$. En pratique, l'estimation du modèle est réalisée en censurant les évènements autres que celui d'intérêt $k$.

#### II.1.6.2 Modèle de Fine & Gray

Contrairement au modèle cause-spécifique qui modélise chaque cause $k$ indépendemment, le modèle de Fine & Gray propose une nouvelle formulation du modèle via les fonctions de risque de sous-répartition [Gray, 1988] notées $\alpha_k(t)$ et définies par :

$$\alpha_k(t) = -\frac{\mathrm{d}\log\left(1 - I_k(t)\right)}{\mathrm{d}t} \tag{II.16}$$

où $I_k(t)$ désigne la fonction d'incidence cumulée pour la cause $k$ définie dans l'équation (II.13).

Le modèle de Fine & Gray permet de modéliser le risque instantané de sous-répartition $\alpha_k(t)$ pour chaque cause $k$, en fonction d'un ensemble de covariables $X$ par :

$$\alpha_k(t|X) = \alpha_{k,0}(t)\exp\left(X^\top \beta_k\right) \tag{II.17}$$

avec $\alpha_{k,0}(t)$ le risque de base instantané de sous-répartition.

A partir du modèle de Fine & Gray, il est possible de tester s'il existe une différence significative entre les courbes de fonction d'incidence cumulée $I_k(t)$ de plusieurs groupes. Cette différence peut être calculée à l'aide du test de Gray [Gray, 1988].

## II.2 Modèles de survie adaptés à la grande dimension

Dans certaines applications, notamment en génétique, le nombre de prédicteurs peut être plus élevé que le nombre d'observations. Dans ce cas de figure, les données deviennent éparses dans l'espace de représentation. Ainsi, les méthodes traditionnelles, comme la régression linéaire, conduisent à des résultats biaisés [Hastie et al., 2009]. De plus, l'augmentation du nombre de prédicteurs engendre également une corrélation possiblement importante entre elles, nécessitant d'être prise en compte. Pour résoudre ces problèmes, plusieurs méthodologies ont été proposées dont les régressions dites pénalisées [Goeman, 2009, Simon et al., 2011] et les forêts aléatoires en survie [Ishwaran et al., 2008, Ishwaran

et al., 2014, Wright and Ziegler, 2017].

### II.2.1 Régressions pénalisées

Parmi les méthodes utilisées pour gérer un très grand nombre de prédicteurs, les régressions pénalisées sont des extensions des modèles de régression en survie comme le modèle de Cox. L'écriture du modèle reste la même, mais une pénalisation est ajoutée dans la vraisemblance pour contraindre l'estimation des paramètres. Plusieurs types de pénalisation ont été proposés.

Soit $L(\beta) = \log \mathcal{L}(\beta)$, le logarithme de la vraisemblance partielle du modèle de Cox, défini dans l'équation II.9, les paramètres $\beta$ sont alors estimés à l'aide d'une log-vraisemblance pénalisée avec une pénalisation de type $\ell_1$ et/ou $\ell_2$. Un exemple classique est la pénalisation *Elastic-Net* dans laquelle :

$$\widehat{\beta} = \arg\max_{\beta} \left( \log \mathcal{L}(\beta) - \alpha \Big( r \sum_{p=1}^{P} |\beta_p| + \frac{1-r}{2} \sum_{p=1}^{P} \beta_p^2 \Big) \right) \tag{II.18}$$

avec $\beta_p$ le coefficient associé au prédicteur $p \in \{1, ..., P\}$, $\alpha \geq 0$ un paramètre utilisateur contrôlant la force de pénalité et $r \in [0; 1]$ un paramètre utilisateur contrôlant le rapport entre les pénalités de type $\ell_1$ et $\ell_2$. Dans le cas où $\alpha = 0$, aucune pénalité n'est appliquée et l'estimation des paramètres se fait comme dans le modèle de Cox. Lorsque $r = 0$, la pénalisation est uniquement de type $\ell_1$ (appelée *Lasso*) alors qu'elle sera uniquement de type $\ell_2$ (appelée *Ridge*) quand $r = 1$. Lorsque $r \in ]0; 1[$, la pénalisation choisie est appelée *Elastic-Net* [Zou and Hastie, 2005] et correspond alors à un mixte entre pénalisation $\ell_1$ et $\ell_2$.

Ces types de pénalisation vont entraîner un effet différent sur l'estimation des paramètres $\beta$. En effet, plus le paramètre de pénalisation $\alpha$ sera élevé, plus l'estimation des paramètres va tendre vers 0. Avec la pénalisation *Lasso*, les paramètres peuvent être exactement à 0, contrairement à la pénalisation *Ridge*, où les paramètres pourront être très proche de 0, sans toutefois y être égaux. Par conséquent, la pénalisation *Lasso* est efficace

pour sélectionner un sous-ensemble de variables très associées à l'évènement contrairement à la pénalisation *Ridge* qui est plutôt efficace pour gérer les variables corrélées [Hastie et al., 2009]. La pénalisation *Elastic-Net* est un mixte entre une sélection de variables (par *Lasso*) et gestion des variables corrélées (par *Ridge*).

D'autres pénalisations ont également été développées telles que l'*adaptive lasso* [Zou, 2006], le *smoothly clipped absolute deviations* (SCAD) [Fan and Li, 2001] ou le *minimax concave penalty* (MCP) [Zhang, 2010]. Les pénalisations SCAD et MCP ont notamment l'avantage de pouvoir être utilisées pour faire de l'inférence statistique [Fu et al., 2017], et *in fine*, de tester les possibles associations entre les prédicteurs et l'évènement d'intérêt.

### II.2.2 Forêts aléatoires en survie

Introduite par Breiman, la méthode des forêts aléatoires [Breiman, 2001] est une approche d'ensemble pour prédire une variable à partir d'un très grand nombre de prédicteurs. Initialement développées pour prédire une variable réponse de type continue ou catégorielle, les forêts aléatoires ont ensuite été étendues aux données de survie [Ishwaran et al., 2008]. Malgré la complexité de la méthodologie par rapport à un modèle de Cox par exemple, les forêts aléatoires possèdent de nombreux avantages en prenant en compte : (i) un possible très grand nombre de prédicteurs ; (ii) une possible non-linéarité entre les prédicteurs et la variable réponse ; (iii) une corrélation entre les différents prédicteurs.

Le principe général de la forêt aléatoire est représenté en figure II.3, et se décompose en trois grandes étapes :

1. Créer $B$ nouveaux jeux de données $\mathcal{D}_N^1, ..., \mathcal{D}_N^b, ..., \mathcal{D}_N^B$ à partir des données originelles $\mathcal{D}_N$ composées de $N$ sujets et $P$ variables. Ces nouveaux jeux de données sont construits à l'aide d'un tirage aléatoire avec remise, ou *bootstrap* non paramétrique. Cette étape entraîne en moyenne une exclusion de 37% des individus qui constituent l'ensemble *Out-Of-Bag*, noté $OOB^b$ pour le jeu de données $\mathcal{D}_N^b$ $(b = 1, ..., B)$ ;
2. Construire $B$ arbres de décision $T_1, ..., T_b, ..., T_B$ à partir des $B$ nouveaux jeux de données précédemment créés. Ces arbres ont pour objectif de prédire la variable

réponse pour l'ensemble des individus ayant été utilisés pour les construire ;

3. Agréger l'ensemble des prédictions issues des $B$ arbres pour obtenir une unique prédiction de la forêt aléatoire pour les $N$ individus.

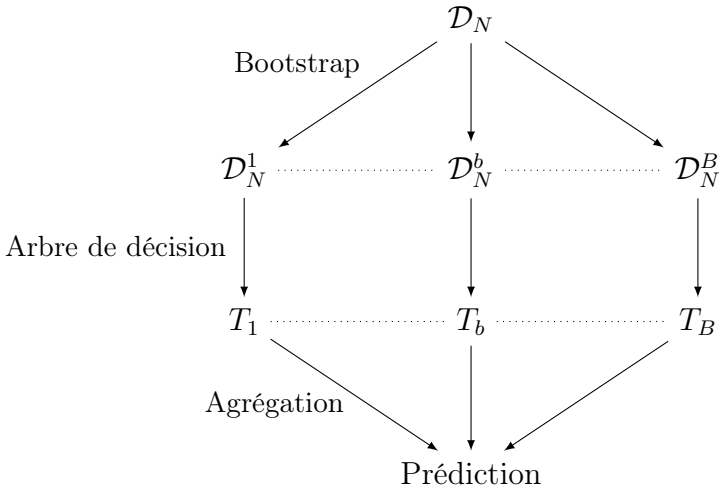


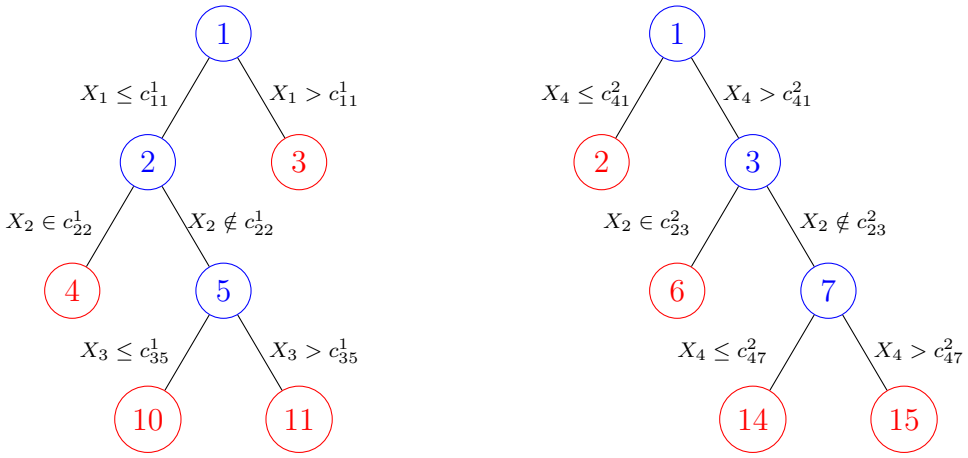

FIGURE II.3 – Structure générale d'une forêt aléatoire

Les étapes de construction de l'arbre et d'agrégation des prédictions dépendent de la nature de la variable réponse. Ces étapes sont détaillées dans la suite de cette section pour l'analyse des données de survie.

#### II.2.2.1 Construction de l'arbre

L'objectif d'un arbre de décision $T_b$ est de partitionner, à partir d'un ensemble de prédicteurs $X = (X_1, ..., X_p)$, les $N^b$ individus (tirés par *bootstrap*) en groupes homogènes en terme de survie. Comme représenté dans la figure II.4, un arbre de décision est composé de noeuds $d \in \mathcal{D}^b$ (en bleu) et de feuilles $h \in \mathcal{H}^b$ (en rouge). Les noeuds $d$ et les feuilles $h$ sont composés respectivement de $N_d^b$ et $N_h^b$ individus, sachant que les $N^b$ individus sont répartis dans l'ensemble des feuilles $\mathcal{H}^b$, soit $\sum_{h \in \mathcal{H}^b} N_h^b = N^b$.

A chaque noeud $d$, un sous-ensemble $\overline{X}^d$ de prédicteurs est tiré aléatoirement parmi les $X$ prédicteurs. La taille de $\overline{X}^d$ dépend d'un paramètre $mtry$ défini par l'utilisateur, et commun à toute la forêt aléatoire. Pour chaque variable de $\overline{X}^d$, un ensemble de décomposition en deux groupes est créé. Dans le cas d'une variable catégorielle, les groupes peuvent




Arbre 1 Arbre 2

FIGURE II.4 – Représentation de deux arbres de décision où à chaque noeud $d \in \mathcal{D}^b$ (représenté par un cercle bleu), les individus sont récursivement répartis en deux sous-groupes selon une variable $X_p$ et un seuil $c^b_{pd}$ ou un sous-ensemble de modalités, jusqu'à un critère d'arrêt pour se retrouver dans une feuille $h \in \mathcal{H}^b$ (représenté par un cercle rouge).

être constitués à partir de toutes les combinaisons possibles des modalités. Pour une variable continue, $C$ seuils (pour $c = 1, ..., C$) sont définis pour créer $C$ variables binaires à partir de $\overline{X}^d_p$ (comme illustré dans la figure II.5A). Par exemple, les seuils peuvent être choisis à partir des déciles du prédicteur. Au final, un ensemble de variables binaires est obtenu à partir des variables appartenant à $\overline{X}^d$ sur lesquelles la différence de risque entre les groupes est testée. La taille de cet ensemble dépend du nombre combinaisons/seuils possibles mais reste, dans tous les cas, supérieur à la taille de $\overline{X}^d$.

Pour chaque variable binaire ainsi constituée, la distance en terme de survie entre les deux groupes (figure II.5B) est quantifiée à l'aide d'une statistique, le plus souvent la statistique de test du *log-rank* introduit en section II.1.4 dans le cas d'une seule cause d'évènement ou la statistique de test de Fine & Gray dans le cas d'évènements concurrents. Ainsi, la variable binaire qui maximise la statistique de test est sélectionnée. Le partitionnement optimal est défini par le couple $\{X_p, c^b_{pd}\}$, où les groupes sont constitués à partir de la combinaison/seuil $c^b_{pd}$ pour la variable $X_p$ au noeud $d$ de l'arbre $b$.

Les individus composant ces deux groupes sont répartis dans deux nouveaux noeuds fils, notés $2d$ et $2d+1$ et la même procédure est répétée jusqu'à atteindre le critère d'arrêt.

Par exemple :

$$nodesize > \sum_{i \in N_d^b} \mathbb{1}_{\delta_i = k} \tag{II.19}$$

En d'autres termes, la procédure de partitionnement est effectuée tant que le nombre d'évènement d'intérêt $k$, parmi les $N_d^b$ individus du noeud $d$, est supérieur ou égal à *nodesize*, un paramètre utilisateur fixe pour l'ensemble de la forêt aléatoire. Dans le cas où le nombre d'évènement d'intérêt est inférieur à *nodesize*, le partitionnement n'est pas réalisé et le noeud $d$ est désormais considéré comme une feuille $h$.

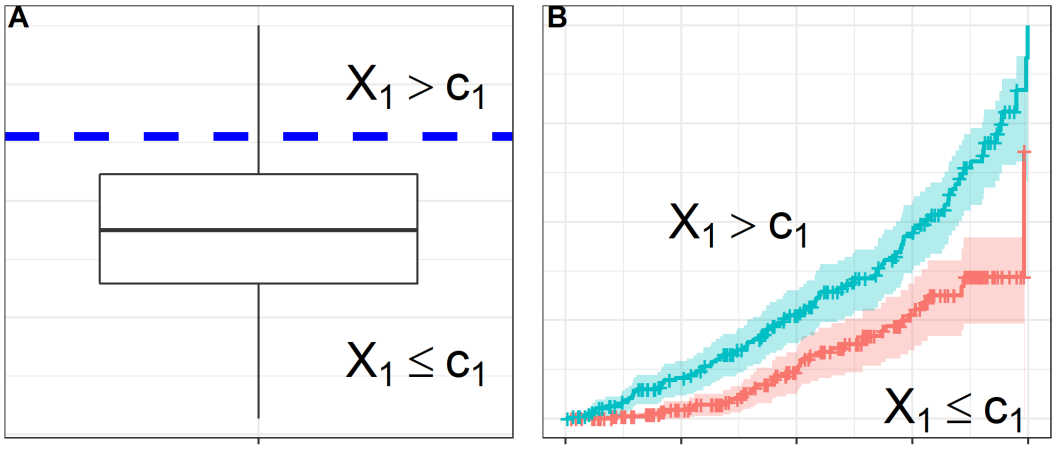



FIGURE II.5 – Illustration de la recherche de partitionnement optimal à chaque noeud $d$ dans un arbre. A partir de la variable $X_1$, les individus sont répartis en deux groupes en fonction d'un seuil $c_1$ (A). La différence de risque entre ces groupes (B) est ensuite quantifiée à l'aide d'une statistique de test (*log-rank* ou Gray). Le partitionnement optimal est obtenu en maximisant cette statistique de test.

Enfin, pour chaque feuille $h \in \mathcal{H}^b$, n'importe quelle fonction résumée de l'évènement peut être calculée. En pratique, la probabilité de subir l'évènement $k$ avant $t$, notée $\pi_k^{h^b}(t) = P(T < t, \delta = k)$, est calculée à partir de l'ensemble des individus $N_h^b$ de la feuille $h$. Cette probabilité est estimée par :

$$\widehat{\pi_k^{h^b}}(t) = \begin{cases} \widehat{F}(t) = 1 - \exp\left(-\widehat{\Lambda}_k^h(t)\right) \text{ pour } k = 1 \\ \widehat{I}_k(t) = \int_0^t \widehat{\lambda}_k(u) \exp\left(-\sum_{k=1}^K \widehat{\Lambda}_k^h(u)\right) \text{ pour } k > 1 \end{cases} \tag{II.20}$$

où $\widehat{\Lambda}_k^h(t)$ est obtenu par l'estimateur de Nelson-Aalen [Nelson, 1969, Aalen, 1976] pour

une cause, ou d'Aalen-Johansen [Aalen and Johansen, 1978] dans le cas d'évènements compétitifs.

#### II.2.2.2 Agrégation des prédictions

Pour un individu $i \in \{1, ..., N\}$, la probabilité prédite $\widehat{\pi_{ik}}(t)$ est obtenue après agrégation sur les arbres suivant :

$$\widehat{\pi_{ik}}(t) = \frac{1}{|\mathcal{O}_i|} \sum_{b \in \mathcal{O}_i} \widehat{\pi_{ik}^{h_i^b}}(t) \tag{II.21}$$

où $\mathcal{O}_i$ représente l'ensemble des arbres où l'individu $i$ est *Out-Of-Bag* (i.e. non sélectionné suite au *bootstrap*) et $|\mathcal{O}_i|$ est la taille de cet ensemble. $\widehat{\pi_{ik}^{h_i^b}}(t)$ est la probabilité de subir l'évènement $k$ avant $t$ lorsque l'individu $i$ descend l'arbre $b$, à partir de ses données $X_i$, jusqu'à tomber dans sa feuille $h_i^b$.

L'agrégation a été présentée pour un individu utilisé pour la construction de la forêt aléatoire, mais elle peut également être généralisable pour un nouvel individu où cette fois-ci, l'intégralité des $B$ arbres est utilisée pour obtenir cette probabilité.

#### II.2.2.3 Optimisation des hyperparamètres

A partir des $N$ individus ayant contribués à la construction de la forêt aléatoire, les performances prédictives de la forêt peuvent être évaluées en comparant les prédictions (obtenues dans l'équation (II.21)) avec les données observées. Pour cela, plusieurs critères peuvent être utilisés et détaillés dans la suite en section II.4.4. Le critère choisi, évalué de façon interne, est appelé erreur *Out-Of-Bag*, notée $errOOB$.

L'objectif lors de la construction de la forêt aléatoire est de minimiser l'erreur *Out-Of-Bag*. Pour cela, il est nécessaire de fournir un ensemble d'arbres les plus différents les uns des autres, tout en gardant les meilleures performances en terme de prédiction. Différents paramètres peuvent contrôler les performances prédictives dont :

1. $B$ le nombre d'arbres. Ce paramètre ne nécessite pas d'être optimisé mais doit être choisi assez grand pour réduire la variance des prédictions [Hastie et al., 2009],

conduisant à la convergence de l'erreur *Out-Of-Bag* [Probst and Boulesteix, 2017] ;

2. *mtry* le nombre de prédicteurs sélectionnés à chaque noeud. Cet hyper-paramètre est crucial, et nécessite d'être optimisé minutieusement, puisqu'il permet de régler *in fine* la quantité d'aléatoire à introduire lors de l'étape de sélection des prédicteurs. Avec un $mtry = 1$, les arbres sont fortement décorrélés puisqu'ils sont construits à partir d'un sous-ensemble de prédiction tiré complètement aléatoirement, conduisant à des prédictions sous-optimales [Probst et al., 2019]. Dans le cas extrême où $mtry = P$, la quantité d'aléatoire est quasi nulle puisque tous les prédicteurs sont systématiquement candidats. Les meilleurs prédicteurs sont sélectionnées à chaque noeud, mais seule la première étape de *bootstrap* permet d'obtenir des arbres différents. Le choix du *mtry* est donc un compromis entre la variance (par la corrélation des arbres) et la précision (par la quantité d'aléatoire) des prédictions ;
3. *nodesize* le nombre minimal d'évènements d'intérêt pouvant être contenu dans chaque feuille. Cet hyper-paramètre est également très important puisqu'il va déterminer la profondeur des arbres ; en d'autres termes, si le partitionnement des individus a été réalisé un grand nombre de fois ou non, en moyenne. En règle générale, les arbres peu profonds sont sous-optimaux et peuvent conduire à des prédictions biaisées [Ishwaran et al., 2011]. Par conséquent, un *nodesize* de petite taille est privilégié mais il faut néanmoins s'assurer que la statistique utilisée pour le partitionnement soit calculable.

Le choix de ces hyperparamètres engendre également un impact très important sur les temps de calcul. En particulier, un grand nombre d'arbre $B$ avec une profondeur élevée (i.e. *nodesize* faible) entraînent une importante augmentation des temps de calcul [Probst et al., 2019]. Ce point est donc également à prendre en compte lors de l'optimisation des hyperparamètres.

#### II.2.2.4 Importance des prédicteurs sur le risque d'évènement

L'objectif principal des forêts aléatoires en survie est de prédire la probabilité de survenue d'un évènement à partir d'un ensemble de prédicteurs. Néanmoins, cette méthodologie peut également être utilisée dans un objectif étiologique, pour déterminer les prédicteurs les plus associés au risque d'évènement. Pour cela, l'importance des variables [Breiman, 2001], notée VIMP($X_p$), peut être calculée pour une variable $X_p$ par :

$$\text{VIMP}(X_p) = \frac{1}{B}\sum_{b=1}^{B}(\widetilde{errOOB}_p^b - errOOB^b) \tag{II.22}$$

où $errOOB^b$ est l'erreur OOB associée à l'arbre $b$ en utilisant les individus OOB de cette arbre. $\widetilde{errOOB}_p^b$ désigne l'erreur OOB obtenue à partir des individus OOB de l'arbre $b$, après avoir *cassé* le possible lien entre le prédicteur $X_p$ et l'évènement. En pratique, une permutation aléatoire des observations entre individus OOB est utilisée pour *casser* ce possible lien.

Une VIMP($X_p$) élevée indique que les performances prédictives de la forêt seraient plus faibles si le prédicteur $X_p$ était retiré. A l'inverse, une VIMP($X_p$) faible ou négative indique que le prédicteur $X_p$ n'a aucun impact sur les performances prédictives.

La profondeur minimale [Ishwaran et al., 2010] est une alternative à l'importance des variables pour quantifier le pouvoir prédictif d'une variable. Le niveau de profondeur, noté $DL$ pour *depth level*, est illustré dans la figure II.6. La profondeur minimale d'un prédicteur $X_p$ se traduit par le niveau de profondeur minimal où $X_p$ est utilisé pour le partitionnement. Dans l'exemple de la figure II.6, sachant que le prédicteur $X_4$ est retrouvé au niveau 1 et 3, la profondeur minimale pour $X_4$ est de 1. Plus la profondeur minimale du prédicteur $X_p$ est faible, plus le prédicteur a été sélectionné tôt pour le partitionnement et donc sépare bien les groupes selon le critère choisi. Cette quantité peut être moyennée en utilisant l'ensemble des $B$ arbres, pour notamment réduire sa variabilité.

En pratique, l'utilisation de la profondeur minimale est rendue compliquée par le choix du $mtry$. En effet, lorsque $mtry < P$, toutes les variables ne sont pas systématiquement

candidates à chaque noeud.

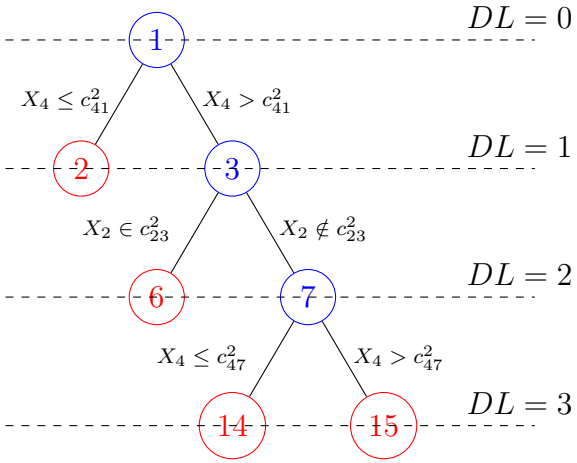



FIGURE II.6 – Illustration du niveau de profondeur, noté $DL$, lors de la construction d'un arbre. Par exemple, le prédicteur $X_4$ est utilisé pour le partitionnement au niveau 1 et 3.

## II.3 Modélisation des données longitudinales par modèles mixtes

Les données longitudinales se définissent par la collecte de mesures répétées au cours du temps pour un individu, contrairement aux données transversales recueillies à un seul temps. Ces données peuvent être collectées aux mêmes temps ou à différents temps pour tous les individus. Par exemple, la pression artérielle mesurée à différents temps pour des patients hospitalisés.

Dans la suite, nous notons $Y_{ij}$ l'observation de la variable d'intérêt $Y$ pour le sujet $i$, $i = 1, ..., N$ au temps $t_{ij}$ avec $j = 1, ..., n_i$.

Pour prendre en compte la répétition des mesures, et par conséquent la corrélation intra-sujets, il est nécessaire d'utiliser des modèles particuliers, et notamment les modèles mixtes [Laird and Ware, 1982] qui permettent de tenir compte de cette corrélation. Il existe plusieurs modèles mixtes en fonction de la nature de la variable d'intérêt, en particulier les modèles linéaires mixtes lorsque la variable $Y$ est distribuée selon une loi normale. Pour

une distribution de $Y$ appartenant à la famille de distribution exponentielle, les modèles linéaires généralisés mixtes sont alors utilisés.

## II.3.1 Modèle linéaire mixte

### II.3.1.1 Spécification du modèle linéaire mixte

Avec une variable $Y$ distribuée selon une loi normale, le modèle linéaire mixte se défini par :

$$Y_{ij} = Y_i^\star(t_{ij}) + \epsilon_{ij} = X_i^\top(t_{ij})\beta + Z_i^\top(t_{ij})b_i + \epsilon_{ij} \tag{II.23}$$

où $Y_{ij}$ est la valeur observée et définie à partir de la vraie valeur $Y_i^\star(t_{ij})$ (non observable) et d'une erreur de mesure homoscédastique $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. $X_i(t_{ij})$ et $Z_i(t_{ij})$ sont les $p$- et $q$-vecteurs des variables explicatives avec $Z_i(t_{ij}) \subseteq X_i(t_{ij})$. $\beta$ et $b_i$ sont respectivement les effets fixes et aléatoires associés aux variables explicatives $X$ et $Z$. Les effets fixes sont des paramètres de régression classiques qui décrivent l'évolution de $Y$ au niveau de la population. Les effets aléatoires sont spécifiques à chaque sujet et décrivent l'écart individuel à la trajectoire moyenne de $Y$. Ce sont les effets aléatoires qui tiennent compte de la corrélation intra-sujet entre les mesures répétées. Les effets aléatoires sont distribués tels que $b_i \sim \mathcal{N}(0, B)$, où $B$ est la matrice de covariance de dimension $q \times q$. Selon la corrélation entre les effets aléatoires, le nombre de paramètres de $B$ diffèrent. Dans le cas le plus simple où aucune corrélation n'est considérée, $B$ est une matrice diagonale composée de $q$ paramètres à estimer. Lorsque la corrélation entre les effets aléatoires est prise en compte sans structure prédéfinie, $B$ est composée de $q \times (q+1)/2$ paramètres puisque $B$ est symétrique.

Soit $Y_i = (Y_{i1}, ..., Y_{ij}, ..., Y_{in_i})$ le vecteur des observations pour l'individu $i$. La variance de $Y_i$ est définie telle que :

$$\text{var}(Y_i) = V_i = Z_i B Z_i^\top + \Sigma_i \tag{II.24}$$

où $\Sigma_i$ est la matrice de covariance des erreurs de mesures. Selon l'hypothèse où les erreurs de mesures sont indépendantes entre elles, $\Sigma_i = \sigma I_{n_i}$ est une matrice carrée diagonale de taille $n_i$. $Z_i$ est la matrice de vecteurs lignes $Z_i^\top(t_{ij})$ pour $j = 1, ..., n_i$.

#### II.3.1.2 Estimation par maximum de vraisemblance

Soit $(\beta, \phi)$ les paramètres à estimer issus d'un modèle linéaire mixte où $\phi = (Vec(B), \sigma)$ est le vecteur des paramètres de variance des effets aléatoires et des erreurs de mesure intervenant dans $V_i$. Les paramètres sont estimés à l'aide de la log-vraisemblance $\mathcal{L}(\beta, \phi)$ définie par :

$$\mathcal{L}(\beta, \phi) = -\frac{1}{2}\sum_{i=1}^{N}\{n_i \log(2\pi) + \log|V_i(\phi)| + (Y_i - X_i\beta)^\top V_i(\phi)^{-1}(Y_i - X_i\beta)\} \qquad \text{(II.25)}$$

où $|V_i(\phi)|$ est le déterminant de la matrice $V_i(\phi)$ et $X_i$ est la matrice de vecteurs lignes $X_i^\top(t_{ij})$.

Les estimateurs du maximum de vraisemblance peuvent être obtenus en maximisant cette vraisemblance. Le plus souvent, cela est fait en maximisant l'équation (II.25) par des algorithmes d'optimisation, comme par exemple l'algorithme de Marquardt-Levenberg [Marquardt, 1963].

#### II.3.1.3 Prédiction des effets aléatoires individuels

Dans le cadre de prédictions individuelles, il est nécessaire de prédire également les effets aléatoires $b_i$ spécifiques à chaque sujet. La distribution conjointe de $Y_i$ et $b_i$ est définie selon la loi normale multivariée suivante :

$$\begin{bmatrix} Y_i \\ b_i \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} X_i\beta \\ 0 \end{bmatrix}, \begin{bmatrix} V_i(\phi) & Z_iB \\ BZ_i^\top & B \end{bmatrix} \right) \qquad \text{(II.26)}$$

Les effets aléatoires $b_i$ peuvent être prédits par l'espérance *a posteriori* $E(b_i|Y_i)$ :

$$E(b_i|Y_i) = BZ_i^\top V_i(\phi)^{-1}(Y_i - X_i\beta) \tag{II.27}$$

En pratique, les paramètres $\beta$ et $\phi$ sont remplacés par leurs estimations $\widehat{\beta}$ et $\widehat{\phi}$ dans cette expression pour obtenir l'estimateur empirique bayésien $\widehat{b_i}$, également appelé BLUP pour *Best Linear Unbiased Predictor*.

#### II.3.1.4 Prédictions individuelle de la variable réponse

Dans un objectif de prédiction, nous nous intéressons à prédire la variable réponse $Y(t)$ pour tout temps $t$. Pour un individu $i$, $Y(t)$ peut être obtenu par l'espérance de $Y(t)$ conditionnellement à l'histoire des données $Y_i$, notée $E(Y(t)|Y_i)$. La définition de $E(Y(t)|Y_i)$ peut être simplifiée du fait de l'espérance des $Y$ sachant $b_i$ suivant :

$$\begin{aligned} E(Y(t)|Y_i) &= \int E(Y(t)|Y_i, b_i) f(b_i|Y_i) db_i \\ &= \int E(Y(t)|b_i) f(b_i|Y_i) db_i \end{aligned} \tag{II.28}$$

avec $E(Y(t)|b_i) = X_i(t)\beta + Z_i(t)b_i$.

A partir des paramètres estimés $\widehat{\beta}$ et $\widehat{b_i}$, les prédictions individuelles estimées $\widehat{Y_i}$ sont obtenues par :

$$\widehat{Y_i}(t) = \widehat{E}(Y(t)|b_i) = X_i(t)\widehat{\beta} + Z_i(t)\widehat{b_i} \tag{II.29}$$

Lorsque les effets aléatoires $\widehat{b_i}$ sont estimés à l'aide du BLUP suivant l'équation (II.27), les prédictions individuelles estimées $\widehat{Y_i}$ deviennent par :

$$\widehat{Y_i}(t) = X_i(t)\widehat{\beta} + Z_i(t)BZ_i^\top V_i(\phi)^{-1}(Y_i - X_i\widehat{\beta}) \tag{II.30}$$

## II.3.2 Modèle linéaire généralisé mixte

### II.3.2.1 Spécification du modèle mixte généralisé

Lorsque la variable $Y$ suit une loi de la famille exponentielle, les modèles linéaires généralisés mixtes peuvent être utilisés. La famille exponentielle est définie comme une famille de lois incluant notamment la loi normale, Bernoulli et Poisson.

Dans ce cas du modèle généralisé mixte, l'espérance de $Y$ conditionnellement aux effets aléatoires est modélisée par :

$$g\Big(E(Y(t)|b_i)\Big) = X_i^\top(t_{ij})\beta + Z_i^\top(t_{ij})b_i \tag{II.31}$$

où la variable $Y$ est reliée au prédicteur linéaire $X_i^\top(t_{ij})\beta + Z_i^\top(t_{ij})b_i$ par une fonction de lien $g$ dépendant de la nature de $Y$ (binaire, données de comptage, ...). Par exemple, si $Y$ est distribué selon une loi de Bernoulli, la fonction *logit* est choisi pour $g$. En règle générale, les effets aléatoires sont distribués selon une loi normale de moyenne 0 et de matrice de covariance $B$.

### II.3.2.2 Estimation par maximum de vraisemblance

Les paramètres $\beta$ et de variance $\phi = vec(B)$ peuvent être estimés à l'aide de la log-vraisemblance $\mathcal{L}(\beta, \phi)$ du modèle mixte généralisé. L'écriture de la log-vraisemblance exploite l'indépendance entre les sujets ainsi que l'indépendance entre les observations d'un même sujet conditionnellement aux effets aléatoires. Elle est définie par :

$$\mathcal{L}(\beta, \phi) = \sum_{i=1}^{N} \Big( \log \int \prod_{j=1}^{n_i} f_{Y_{ij}|b_i}(Y_{ij}|b_i) f_{b_i}(b_i) db_i \Big) \tag{II.32}$$

où $f_{Y|b}$ est la densité de $Y_{ij}$ définie à partir de la distribution de $Y$ et de la fonction de lien $g$ choisie.

La plupart du temps, l'intégrale sur les effets aléatoires ne possède pas de solution analytique et nécessite d'être approchée par des méthodes d'intégration numérique telles que

la quadrature gaussienne [Wulfsohn and Tsiatis, 1997] ou l'approximation de Laplace [Rizopoulos et al., 2009].

#### II.3.2.3 Prédiction des effets aléatoires individuels

Contrairement à la section II.3.1.3, la prédiction des effets aléatoires individuelles $b_i$ est plus complexe lorsque la variable $Y_i$ n'est pas distribuée selon une loi normale, car il n'y a pas d'expression analytique. Dans ce cas, une approximation de l'espérance de $b_i|Y_i$ est réalisée en utilisant le mode de la densité conditionnelle $f(b_i|Y_i)$. En utilisant la relation suivante :

$$f_{b_i|Y_i}(b_i|Y_i) = \frac{f_{Y_{ij}|b_i}(Y_{ij}|b_i)f_{b_i}(b_i)}{f_{Y_i}(Y_i)} \propto f_{Y_{ij}|b_i}(Y_{ij}|b_i)f_{b_i}(b_i) \tag{II.33}$$

l'approximation de $E(b_i|Y_i)$ est obtenue en utilisant le mode de la fonction $f_{Y_{ij}|b_i}(Y_{ij}|b_i)f_{b_i}(b_i)$. Ce mode est calculé numériquement par maximisation de la fonction à l'aide d'une méthode d'optimisation.

### II.3.3 Association avec les données de survie

Dans les cohortes, des données de survie peuvent être collectées en même temps que des données répétées de variables. Par exemple, dans la cohorte de personnes âgées des trois-cités, plusieurs tests cognitifs sont collectés tous les 2 ou 3 ans (données longitudinales) en plus de l'indicateur de démence (données de survie). Ainsi, l'étude d'un évènement clinique à partir des données répétées d'une variable dépendante du temps devient possible.

Pour analyser ces données, une approche naïve consiste à inclure la variable dépendante du temps dans un modèle à risque proportionnel, comme un modèle de Cox. Cependant, cette approche n'est pas valide et entraîne une estimation biaisée [Prentice, 1982] car la variable est mesurée avec erreur et à des temps discrets. En effet, la variable nécessite d'être mesurée pour tous les temps d'évènements. De plus, l'utilisation d'un modèle de Cox avec variable dépendante du temps nécessite une variable dite exogène (i.e. la distribution de la

variable n'est pas modifiée par la survenue des évènements) ce qui n'est pas souvent le cas lors de l'utilisation de prédicteurs collectés dans les cohortes [Kalbfleisch and Prentice, 2011]. Une alternative consiste à utiliser uniquement la dernière mesure [Keogh et al., 2019] pour décrire la dynamique de la variable mais cette approche est trop simpliste puisqu'elle fait l'hypothèse que la mesure est constante au cours du suivi.

Pour correctement tenir compte des caractéristiques d'une variable dépendante du temps endogène (i.e. où la distribution de la variable peut être modifiée par la survenue des évènements), le modèle conjoint peut être utilisé. Ce modèle consiste à introduire en variable explicative du modèle de survie des caractéristiques de la dynamique de $Y$. Dans le cas Gaussien, ces modèles sont définis par :

$$\begin{cases} Y_{ij} = Y_i^\star(t_{ij}) + \epsilon_{ij} = X_{ij}^\top \beta + Z_{ij}^\top b_i + \epsilon_{ij} \text{ avec } b_i \sim \mathcal{N}(0, B) \text{ et } \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ \lambda_i(t) = \lambda_0(t) \exp\left(W_i^\top \gamma + \Gamma_i(t, \beta, b_i, B)^\top \nu\right) \end{cases} \tag{II.34}$$

où $W_i$ est un sous-vecteur des variables explicatives avec $\gamma$ les coefficients associés. $\Gamma_i(t)$ est un vecteur de variables résumées caractérisant la dynamique $Y^\star$ et $\nu$ les coefficients associés. Les notations pour le modèle linéaire mixte sont les mêmes que décrites dans l'équation (II.23).

Pour décrire la dynamique de $Y_i$ dans $\Gamma_i(t)$, des résumés peuvent être calculé à partir d'une fonction des effets aléatoires [Tsiatis and Davidian, 2004] tels que :

- la déviation individuelle à la trajectoire moyenne : $\Gamma_i(t) = b_i$
- le niveau courant $\Gamma_i(t) = Y_i^\star(t) = X_i(t)^\top \beta + Z_i(t)^\top b_i$ ;
- la pente courante $\Gamma_i(t) = Y_i^{*\prime}(t) = \frac{\partial Y_i^*(t)}{\partial t}$ ;

Ces résumés sont les plus courants et ont déjà été utilisés dans l'étude des maladies cardio-vasculaires [Rizopoulos et al., 2017, Sweeting et al., 2017].

Deux approches ont été développées pour estimer ces modèles, soit par une estimation en deux étapes [Ye et al., 2008, Van Houwelingen, 2007] ou par une estimation conjointe [Tsiatis and Davidian, 2004, Proust-Lima and Taylor, 2009, Rizopoulos, 2012]. Ces deux approches sont détaillées dans la suite de cette section.

#### II.3.3.1 Estimation en deux étapes

L'approche en deux étapes consiste à modéliser indépendamment les données répétées et les données de survie. Tout d'abord, les paramètres du modèle mixte sont estimés à l'aide du maximum de vraisemblance décrit dans l'équation (II.25). Puis, les résumés $\Gamma_i$ sont prédits à partir des BLUP comme décrit en section II.3.1.3. Ils sont ensuite inclus dans un modèle de Cox où les paramètres sont estimés en maximisant la vraisemblance partielle comme décrit en section II.1.5. Parmi les approches en deux étapes, il existe la méthode séquentielle [Tsiatis et al., 1995], de *regression calibration* [Ye et al., 2008] et l'approche *landmark* [Van Houwelingen, 2007] qui se différencient en particulier sur la quantité d'information utilisée.

La méthode séquentielle consiste à estimer un modèle mixte pour chaque temps d'évènement $t_k$ à partir des données collectées jusqu'à $t_k$ pour les individus encore à risque en $t_k$. $\Gamma_i(t_k, b_i)$ est estimé en utilisant l'espérance de $b_i$ conditionnellement aux données collectées jusqu'en $t_k$ à partir des modèles mixtes estimés avant l'apparition de l'évènement en $T_i \leq t_k$. Cependant, cette méthode entraîne une estimation plus variable de $\Gamma_i(t_k, b_i)$ car elle est très dépendante du nombre de mesures répétées jusqu'en $t_k$ et de la taille d'échantillon. Au final, cette méthode est globalement peu utilisée puisqu'elle est particulièrement lourde numériquement.

La méthode de *regression calibration* estime indépendamment les modèles de l'équation (II.34) en incluant toute l'information disponible. Néanmoins, cette approche possède l'inconvénient d'estimer le modèle mixte sous l'hypothèse de données manquantes aléatoires (i.e. la probabilité d'être manquant peut être prédite par les observations). Dans le cas où les résumés $\Gamma_i$ de la dynamique sont associés au risque de subir l'évènement, les données manquantes sont probablement informatives (i.e. la probabilité d'observation peut dépendante des données manquantes) et génèrent un biais lors de l'estimation du risque [Albert and Shih, 2010].

L'approche *landmark* limite ce problème en n'utilisant qu'une partie des données. En effet, un temps *landmark* $s$ est défini *a priori* compte-tenu de l'évènement clinique d'in-

térêt et des données disponibles. Les modèles de l'équation (II.34) sont ensuite estimés, toujours indépendamment en deux étapes, mais cette fois uniquement à partir des données de $Y$ collectées jusqu'en $s$, pour les individus encore à risque en $s$. Dans cette approche, les résumés des dynamiques de $Y^\star$ inclus dans $\Gamma_i(t)$ peuvent alors être calculés uniquement pour un $t \leq s$. En n'utilisant qu'une partie de l'information disponible, l'approche *landmark* souffre d'un possible manque de puissance statistique mais reste néanmoins compétitif dans le cas de la prédiction [Ferrer et al., 2019]. En revanche, cette approche doit être estimée pour chaque temps *landmark* différent, pouvant rendre son utilisation compliquée.

Ces approches en deux étapes sont particulièrement flexibles. En effet, dans la littérature récente, des extensions de la méthode par *regression calibration* ont été proposées en utilisant l'analyse en données fonctionnelles [Yao et al., 2005] pour la modélisation des données longitudinales, la régression pénalisée [Signorelli et al., 2021] ou les forêts aléatoires [Jiang et al., 2021, Lin et al., 2021] pour les données de survie. Cependant, ces techniques souffrent potentiellement des mêmes biais que ceux provenant de la méthode de *regression calibration*.

En approche *landmark*, d'autres extensions ont été développées à partir de pseudo-observations [Zhao et al., 2020] ou d'une agrégation de plusieurs méthodes à travers un *superlearner* où la survie est approchée par du binaire répété [Tanner et al., 2021].

#### II.3.3.2 Estimation conjointe

Une alternative à l'estimation en deux étapes consiste à estimer tous les paramètres du modèle conjoint simultanément en une seule étape par la maximisation de la vraisemblance jointe. A partir des modèles mixte et de survie de l'équation (II.34), la log-vraisemblance jointe $\mathcal{L}(\theta)$ est définie par :

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \Big( \log \int_{\mathcal{R}^q} \prod_{j=1}^{n_i} f_{Y_{ij}|b_i}(Y_{ij}|b) S_i(T_i|b) \lambda_i(T_i|b)^{\delta_i} f_{b_i}(b) db \Big) \tag{II.35}$$

avec

$$S_i(T_i|b) = \exp\Big( - \int_0^{T_i} \lambda_i(s|b)ds \Big) \tag{II.36}$$

où $\theta$ représente l'ensemble des paramètres du modèle linéaire mixte et du modèle de survie et $q$ la dimension des effets aléatoires. Néanmoins, le calcul de la vraisemblance est complexe car l'intégrale sur les effets aléatoires n'a pas de solution analytique. Cette intégrale peut être calculée par une méthode de quadrature gaussienne [Wulfsohn and Tsiatis, 1997], Monte-Carlo [Lin et al., 2002], quasi Monte-Carlo [Philipson et al., 2020] ou par l'approximation de Laplace [Rizopoulos et al., 2009].

La modélisation conjointe est l'approche privilégiée pour estimer le risque de survenue d'un évènement en fonction de variables dépendantes du temps endogènes. Néanmoins, cette approche est très lourde numériquement, et la faisabilité dépend de la dimension des effets aléatoires $q$. Lorsque cette dimension est trop grande, l'intégrale sur les effets aléatoires ne peut pas toujours être correctement estimée [Ferrer et al., 2019].

Le modèle conjoint peut également être étendu pour prendre en compte plusieurs variables longitudinales [Rizopoulos and Ghosh, 2011, Philipson et al., 2020], modélisées par autant de modèles mixtes. Mais cette extension reste très limitée car l'estimation de l'intégrale sur les effets aléatoires devient très vite impossible en pratique avec plus de 2 ou 3 variables longitudinales [Ferrer et al., 2019]. Des techniques d'estimations alternatives sont alors requises [Rustand et al., 2022].

## II.4 Développement et évaluation d'outils de prédictions

### II.4.1 Définition de la prédiction dynamique individuelle

Dans le cadre des données de survie, nous allons nous intéresser à la probabilité $\pi_{\star k}(s, w)$ de survenue de l'évènement d'intérêt $k$ défini pour un temps *landmark* $s$ et à un temps d'horizon $w$ pour un nouvel individu $\star$ par :

$$\pi_{\star k}(s,w) = P(s < T_\star \leq s+w, \delta_\star = k \mid T_\star > s, \mathcal{Y}_\star(s), \mathcal{X}_\star) \tag{II.37}$$

où $\mathcal{Y}_\star(s)$ et $\mathcal{X}_\star$ représentent l'historique des données collectées jusqu'en $s$. Cette probabilité peut être calculée pour plusieurs temps *landmark* $s$ définissant le principe de prédictions dynamiques.

A partir de nombreux modèles et méthodes comme les modèles paramétriques, semi-paramétriques ou les méthodes d'apprentissage automatique, la probabilité peut être estimée par :

$$\widehat{\pi}_{\star k}(s,w) = P(s < T_\star \leq s+w, \delta_\star = k \mid T_\star > s, \mathcal{Y}_\star(s), \mathcal{X}_\star, \widehat{\theta}) \tag{II.38}$$

où $\widehat{\theta}$ représente les paramètres/hyperparamètres estimés/optimisés à partir d'un échantillon d'apprentissage. Pour les hyperparamètres, ils nécessitent d'être optimisés pour augmenter les performances prédictives du modèle à partir d'un critère d'évaluation des performances.

### II.4.2 Création d'un outil de prédiction

Le développement d'un outil de prédiction est scindé en deux étapes :

1. l'apprentissage de la méthode où les paramètres/hyperparamètres sont estimés/optimisés ;
2. la validation de la méthode où les prédictions individuelles sont calculées et leurs performances prédictives évaluées.

Il est essentiel que les étapes d'apprentissage et de validation soient réalisées sur des échantillons indépendants pour éviter un sur-apprentissage de la méthode. En effet, si les mêmes données sont utilisées pour l'apprentissage et la validation, les performances prédictives de la méthode risquent d'être surestimées par rapport à ses performances réelles. Pour s'assurer que les étapes d'apprentissage et de validation soient réalisées de manière indépendante, il existe la validation externe ou interne.

### II.4.3 Validation externe/interne

La validation externe est la méthode par excellence pour évaluer les performances d'une méthode, et est primordiale avant de diffuser un outil de prédiction. Cependant, la validation externe est la plus difficile à mettre en oeuvre puisque qu'elle nécessite de disposer d'un nouvel échantillon provenant de la même population d'intérêt. Le cas le plus courant de validation externe survient lorsque les données sont recueillies dans le cadre d'une cohorte multicentrique. Ainsi, l'apprentissage de la méthode peut être réalisé sur un centre et la validation sur un autre. Néanmoins, il faut s'assurer que les variables, nécessaires à la prédiction des nouveaux individus, sont présentes dans l'échantillon de validation.

Quand la validation externe n'est pas possible, il faut alors avoir recours à la validation interne. Parmi les types de validations internes existantes, nous présentons la validation croisée $K$-blocs qui est la validation interne la plus couramment utilisée. Le principe de la validation croisée $K$-blocs est représenté en figure II.7, et consiste à diviser aléatoirement les données en $K$ blocs. Pour chaque bloc $k$ ($k = 1, ..., K$), la méthode est :

- entraînée à partir des données d'apprentissage issue des $(K-1)$ blocs excluant le bloc $k$
- validée à partir des données de validation provenant du $k$-ème bloc

Les probabilités individuelles de l'ensemble des individus sont obtenues par l'agrégation des probabilités calculées pour chacun des $K$ blocs. En pratique, le nombre de blocs est couramment fixé entre $K = 5$ et $K = 10$.

### II.4.4 Évaluation des performances prédictives

Les performances prédictives d'un outil de prédiction sont évaluées à partir des prédictions individuelles obtenues dans l'échantillon de validation qu'il soit interne ou externe. Il s'agit de quantifier les capacités de l'outil à prédire correctement la survenue de l'évènement et à ordonner correctement les risques. En effet, des décisions cliniques basées sur

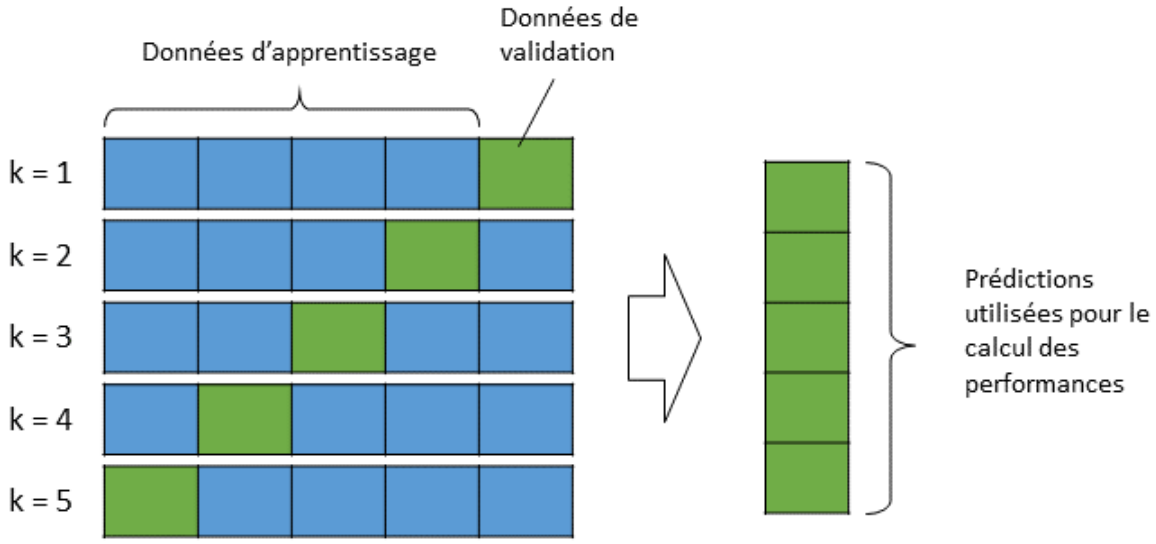



FIGURE II.7 – Principe de la validation croisée 5-blocs. Les blocs de données en bleu sont utilisés pour l'étape d'apprentissage. Les blocs de données en vert sont utilisés pour l'étape de validation. Les prédictions issues de l'étape de validation sont ensuite agrégées pour le calcul des performances.

une méthode non performante peuvent paradoxalement entraîner une augmentation du risque. Pour quantifier ce niveau de performance, et *in fine*, choisir la meilleure méthode de prédiction, deux critères sont classiquement utilisés : la calibration et la discrimination.

La calibration évalue la différence entre le risque prédit par la méthode et le risque observé sur l'ensemble des individus. Par exemple, si le modèle prédit un risque de 10% de décès pour un profil donné, nous nous attendons à observer environ 10 décès parmi 100 patients ayant ce profil. La discrimination est la capacité de la méthode à différencier les individus avec l'évènement de ceux ne l'ayant pas.

De nombreux critères de performances ont été développés [Steyerberg et al., 2010] mais tous ne sont pas transposables dans le cas des données de survie pour les prédictions dynamiques. Nous nous intéresserons dans ce travail plus particulièrement au Brier Score [Brier et al., 1950, Gerds and Schumacher, 2006] et à l'aire sous la courbe ROC [Heagerty and Zheng, 2005, Zheng and Heagerty, 2007], qui sont les critères les plus couramment utilisés. Ils sont tous deux définis pour chaque couple de temps *landmark*/horizon $(s, w)$, et peuvent être estimés à partir des prédictions $\widehat{\pi}_\star(s, w)$ définies dans l'équation (II.38).

#### II.4.4.1 Brier Score

Le Brier Score est un critère évaluant à la fois la calibration et la discrimination d'un modèle [Blanche et al., 2015]. Le Brier Score est l'erreur quadratique moyenne entre la prédiction $\pi^k(s,w)$ et $D^k(s,w)$ le statut de l'individu vis à vis de l'évènement d'intérêt $k$ au temps d'horizon $w$ chez les sujets à risque de l'évènement au temps *landmark* $s$. Il est défini par :

$$BS^k(s,w) = E\Big[\Big(D^k(s,w) - \pi^k(s,w)\Big)^2 | T > s\Big] \tag{II.39}$$

où $\pi^k(s,w)$ représente les probabilités de subir l'évènement d'intérêt $k$ dans l'intervalle de temps $(s, s+w)$ à partir des données collectées jusqu'au temps *landmark* $s$.

Ce critère est défini entre 0 et 1 où une valeur plus faible indique une meilleure performance prédictive de l'outil, sachant que 0,25 est le seuil correspondant à des probabilités individuelles complètement aléatoires.

Sachant que $D^k(s,w)$ n'est pas toujours observable en présence de données censurées, nous définissons $\widetilde{D}^k(s,w) = \mathbb{1}_{(s<\widetilde{T}\leq s+w, \delta=k)}$. Lorsque l'évènement d'intérêt $k$ survient entre $s$ et $s+w$, nous avons $\widetilde{D}^k(s,w) = 1$, sinon $\widetilde{D}^k(s,w) = 0$. Alors le Brier Score doit être estimé à l'aide d'une technique basée sur l'*Inverse Probability of Censoring Weighting* (IPCW) [Gerds and Schumacher, 2006]. Le Brier Score estimé $\widehat{BS}^k(s,w)$ est obtenu par :

$$\widehat{BS}^k(s,w) = \frac{1}{\sum_{i=1}^N \mathbb{1}_{\widetilde{T}_i > s}} \sum_{i=1}^N \widehat{W}_i(s,w)\big(\widetilde{D}_i^k(s,w) - \widehat{\pi}_i^k(s,w)\big)^2 \tag{II.40}$$

où $\widehat{W}_i(s,w)$ représente les poids pour prendre en compte la censure des données, et est défini par :

$$\widehat{W}_i(s,w) = \frac{\mathbb{1}_{(\widetilde{T}_i > s+w)}}{\widehat{G}(s+w)/\widehat{G}(s)} + \frac{\mathbb{1}_{(s<\widetilde{T}_i\leq s+w)}\delta_i}{\widehat{G}(\widetilde{T}_i)/\widehat{G}(s)} \tag{II.41}$$

avec $\widehat{G}$ l'estimateur de Kaplan-Meier.

Le Brier Score peut également être calculé pour un ensemble de temps, appelé le Brier Score Intégré (IBS) [Sène et al., 2016]. L'IBS est défini pour un temps *landmark* $s$ et un

temps d'horizon $w$ par :

$$\widehat{IBS}^k(s,w) = \int_s^{s+w} \widehat{BS}^k(s,t)dt \tag{II.42}$$

#### II.4.4.2 L'aire sous la courbe ROC

L'aire sous la courbe ROC (AUC) est un critère très populaire mesurant la qualité de discrimination d'un outil de prédiction. L'AUC correspond à la probabilité de concordance entre l'outil de prédiction et les observations. Par exemple, une AUC de 75% veut dire que la probabilité que le risque prédit soit plus grand pour une personne ayant l'évènement plus tard est de 75%.

L'AUC est définie pour un évènement $k$, pour un temps *landmark* $s$ et un temps d'horizon $w$ par :

$$AUC^k(s,w) = P\Big(\pi_i^k(s,w) > \pi_j^k(s,w) \mid D_i^k(s,w) = 1, D_j^k(s,w) = 0, T_i > s, T_j > s\Big) \tag{II.43}$$

Pour des données de type survie, l'AUC est estimée par une méthode IPCW comme le Brier Score [Blanche et al., 2015] par :

$$\widehat{AUC}^k(s,w,\widehat{\theta}) = \frac{\sum_{i=1}^N \sum_{j=1}^N \mathbb{1}_{(\widehat{\pi}_i^k(s,w) > \widehat{\pi}_j^k(s,w))} \widetilde{D}_i^k(s,w)\big(1 - \widetilde{D}_j^k(s,w)\big)\widehat{W}_i(s,w)\widehat{W}_j(s,w)}{\sum_{i=1}^N \sum_{j=1}^N \widetilde{D}_i^k(s,w)\big(1 - \widetilde{D}_j^k(s,w)\big)\widehat{W}_i(s,w)\widehat{W}_j(s,w)} \tag{II.44}$$

où les poids pour prendre en compte le censure des données $\widehat{W}(s,w)$ et l'indicateur d'évènement d'intérêt $k$ entre $s$ et $s+w$ $\widetilde{D}_j(s,w)$ sont définis de la même manière que pour le Brier Score de l'équation (II.40).

L'AUC est un score entre 0 et 1, où plus la valeur est élevée, plus l'outil de prédiction permet de discriminer les individus. En conséquence, nous cherchons à maximiser ce score. Un score de 0,5 indique que l'outil de prédiction ne fait pas mieux que l'aléatoire.

---

CONCLUSION

A travers ce chapitre, nous avons défini la modélisation des données

de survie (en particulier en grande dimension), la modélisation des données longitudinales et le principe d'un outil de prédiction. Pour répondre à notre objectif, ces différentes notions ont été combinées pour développer une approche *landmark* dans le chapitre III et une approche par forêt aléatoire en survie dans le chapitre IV.

# Chapitre III

# Approche *landmark* pour multiple données répétées

**Sommaire**

INTRODUCTION

Ce chapitre introduit l'approche *landmark* pour multiples données répétées à travers deux sections. La première section est dédiée à la présentation générale de la méthode, suivi d'une étude de simulation et de deux illustrations pour prédire (i) le décès chez les patients atteints de la cholangite biliaire primaire (ii) le décès parmi une population de personnes âgées. Ce travail a été publié dans le journal *BMC Medical Research Methodology* en tant que principal auteur.

Dans la deuxième section, l'approche *landmark* est étendu au contexte des risques compétitifs pour prédire la démence dans une cohorte de personnes âgées, tout en considérant le décès comme risque concurrent. Ce travail a été initié dans le stage de Master 2 de Marjorie Hitchon que j'ai co-encadré, et finalisé par Ariane Bercu, ingénieure statisticienne. Ce travail va être soumis pour publication en tant que co-auteur.

# III.1 Individual dynamic prediction of clinical endpoint from large dimensional longitudinal biomarker history : a landmark approach

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

[1]INSERM, Bordeaux Population Health, U1219, Univ. Bordeaux, Bordeaux, France

[2]INRIA Bordeaux Sud-Ouest, Talence, France

Published in *BMC Medical Research Methodology*. DOI : 10.1186/s12874-022-01660-3

**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

### III.1.1 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 [Proust-Lima and Taylor, 2009, Sène et al., 2014, Taylor et al., 2013], and markers such as diabetic status or blood pressure level over time are crucial in predicting the risk of cardiovascular disease [Paige et al., 2018, Sweeting et al., 2017]. Including longitudinal information into the prediction of a clinical event defines the framework for individual dynamic predictions [Proust-Lima and Taylor, 2009, Rizopoulos, 2011, Ferrer et al., 2019]. 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 III.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 [Proust-Lima and Taylor, 2009, Rizopoulos, 2011] and landmark models [Van Houwelingen, 2007]. Joint models simultaneously analyze the longitudinal and event time processes by assuming a structure of association built on summary variables of the marker dynamics [Tsiatis and Davidian, 2004]. 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 [Ferrer et al., 2019].

An alternative is to use partly conditional survival model [Maziarz et al., 2017] or landmark models [Van Houwelingen, 2007] 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 III.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 [Proust-Lima

and Taylor, 2009, Sweeting et al., 2017]. 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 [Ferrer et al., 2019]. 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 [Goldstein et al., 2016]. 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 [Breiman, 2001]. Their good predictive performances have been largely demonstrated in the literature [Lebedev et al., 2014]. 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.* [Simon et al., 2011] developed penalized Cox models either using Ridge, Lasso or Elastic-Net penalty, Bastien *et al.* [Bastien et al., 2015] developed a Cox model based on deviance residuals-based sparse-Partial Least Square, as an extension of sparse-Partial Least Square [Chun and Keles, 2010] for survival data, and Ishwaran *et al.* [Ishwaran et al., 2008] 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.* [Zhao et al., 2020] and Jiang *et al.* [Jiang et al., 2021] 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 [Ye et al., 2008] and may suffer from the same biases [Albert and Shih, 2010]. Finally Tanner *et al.* [Tanner et al., 2021] 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.

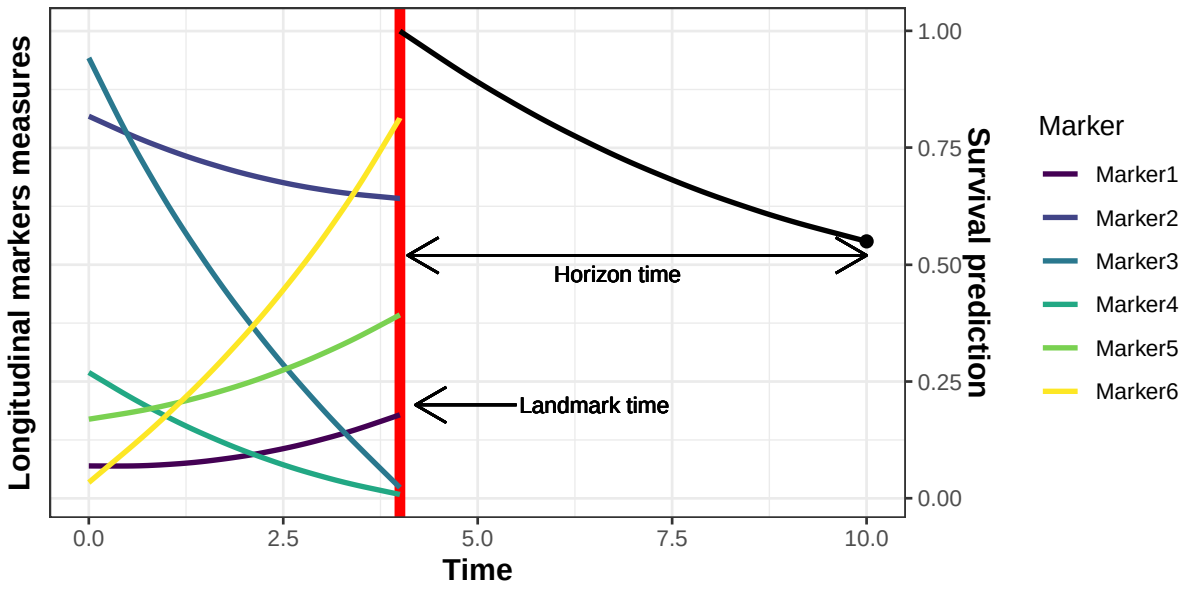



FIGURE III.1 – Illustration of individual dynamic prediction of an event computed using history of multiple repeated markers (here 6). The individual probability of event is computed from a landmark time to a horizon time by using the information on the markers trajectories collected up to the landmark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

## III.1.2 Methods

### III.1.2.1 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 $t_{ijk}$ ($j \in \{1, \ldots, n_i\}$)

and $t_{ijk} \leq t_{LM}$.

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 :

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

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 :

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

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.

#### III.1.2.2 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 [Laird and Ware, 1982]

defined as :

$$g(E(Y_{ijk}|b_{ik})) = Y^*_{ik}(t_{ijk}) = X^\top_{ik}(t_{ijk})\beta_k + Z^\top_{ik}(t_{ijk})b_{ik} \tag{III.3}$$

where $X^\top_{ik}(t_{ijk})$ and $Z^\top_{ik}(t_{ijk})$ are the $p_k$- and $q_k$-vectors associated with the fixed effects $\beta_k$ 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).

#### III.1.2.3 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 : $\widehat{b}_{ik} = \widehat{B}_k Z^\top_{ik} \widehat{V}^{-1}_{ik}(Y_{ik} - X_{ik}\widehat{\beta}_k)$ where $\widehat{V}_{ik} = Z_{ik}\widehat{B}_k Z^\top_{ik} + \widehat{\sigma}_{\epsilon k} I_{n_i}$, 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 $u \leq t_{LM}$ : $\widehat{Y}^*_{ik}(u) = X^\top_{ik}(u)\widehat{\beta}_k + Z^\top_{ik}(\tau)\widehat{b}_{ik}$ ;
- Error-free slope at time $u \leq t_{LM}$ : $\widehat{Y}^{*\prime}_{ik}(u) = \frac{\partial \widehat{Y}^*_{ik}(t)}{\partial t}|_{t=u}$ ;
- 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 $\Gamma_i$. 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.

#### III.1.2.4 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 $\Gamma_i$ 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 III.1.6.1 in supplementary material for further details.

**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 :

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

with $\lambda_0$ the baseline hazard function, $\gamma$ and $\eta$ 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*).

**Penalized-Cox models** Penalized-Cox models extend the Cox model defined in (III.4) to handle a high number of possibly correlated predictors. The partial log-likelihood is penalized with norm $\ell_2$ (Ridge penalty), norm $\ell_1$ (Lasso penalty [Goeman, 2009]) which enables covariate selection, or a mixture of both (Elastic-Net [Simon et al., 2011]). 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.

**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) [Chun and Keles, 2010] 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 [Bastien et al., 2015] (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*).

**Random Survival Forests** Random forests [Breiman, 2001] 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 [Ishwaran et al., 2008] 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 mi-

nimal 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 [Ishwaran et al., 2008]). 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*).

#### III.1.2.5 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 :

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

with $\widehat{\Lambda}_0(.)$ the Nelson-Aalen estimator, and $\widehat{\mathcal{P}}_{\star}$ the predicted linear predictor di-

rectly 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 :

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

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

#### III.1.2.6 Predictive accuracy assessment

We assessed the predictive performances of the models using the time-dependent Area Under the ROC Curve (AUC) [Blanche et al., 2013] defined as :

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

and time-dependent Brier score [Mogensen et al., 2012] defined as :

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

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 [Mogensen et al., 2012, Blanche et al., 2015] 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 III.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.

#### III.1.2.7 Combining the predictions into a single Super Learner

Each survival prediction method $m$ ($m = 1, ..., M$) provides a different individual predicted probability $\widehat{\pi}_\star^{(m)}$ (equation III.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 [van der Laan et al., 2007, Golmakani and Polley, 2020]. It consists in defining the final predicted probability as a weighted mean over the survival method-specific predictions :

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

where the weights $\omega_m$ (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 (III.8) by internal cross-validation. This lead to a three-layer cross-validation for the superlearner building and validation (see Figure III.2B).

### III.1.3 Results

#### III.1.3.1 Performances of the methodology through a simulation study

We contrasted the performances of the different survival prediction methods according to different scenarios, based on Ishwaran *et al.* [Ishwaran et al., 2014], 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.

**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 (III.3) with identity link.

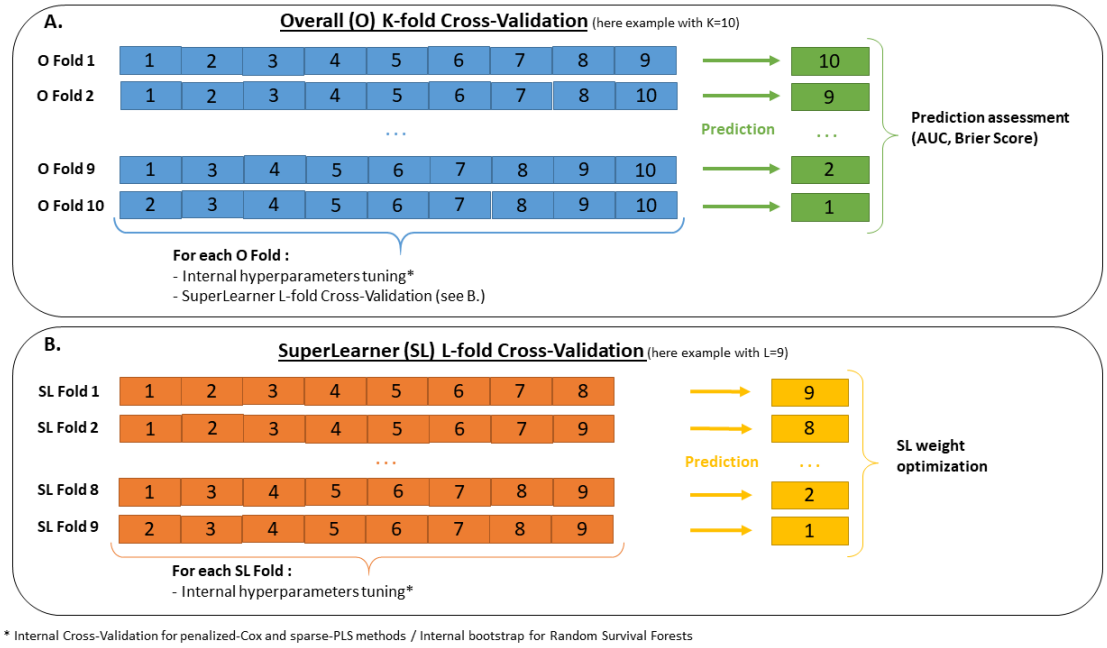



FIGURE III.2 – Multi-layer cross-validation framework : (A) Overall cross-validation to assess the predictive performances on independent samples, (B) Intermediate-layer cross-validation for the superlearner only performed on the learning sample to determine the weights. A final internal cross-validation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

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 Web Figure III.7 in supplementary material). 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 $\Gamma_i^0$. 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 $\Gamma_i^0$ 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 and scenarios are respectively given in Section III.1.6.2 and Web Table III.1 of supplementary material.

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 previously introduced, 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 $\pi_i^0$ over all individuals.

**Results** Predictive performances for scenarios with 18 summaries are summarized in Figure III.3. The same figure for scenarios with 4 summaries is given in Web Figure III.8 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.

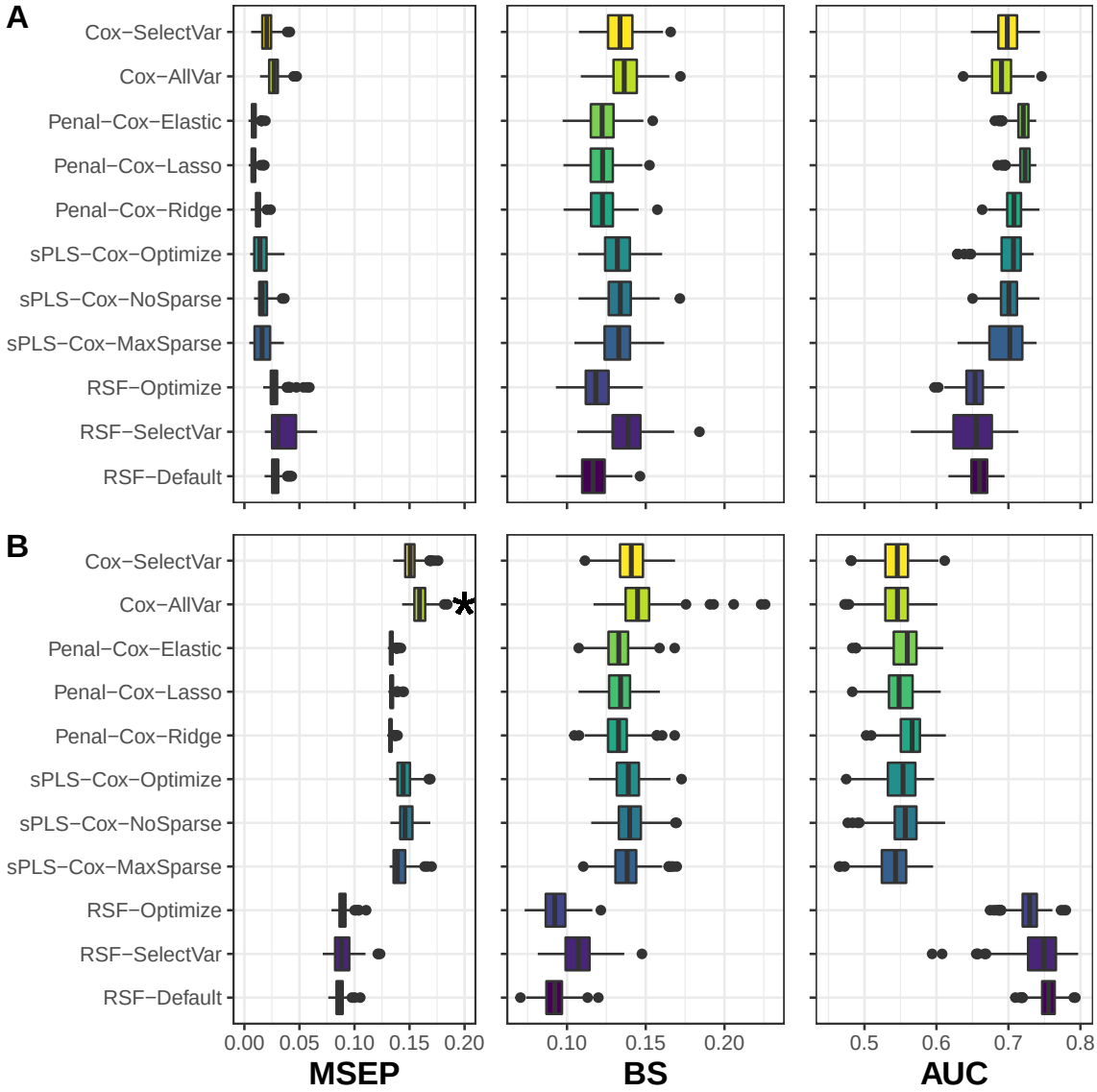



FIGURE III.3 – Simulation results over 250 replicates when considering 18 summaries associated to the event either assuming a linear form (figure A) or non-linear form (figure B). Methods are assessed using Mean Square Error of Prediction (MSEP), Brier Score (BS) and Area Under the ROC Curve (AUC). (∗)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

### III.1.3.2 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 [Kaplan, 1996], 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 [Murtaugh et al., 1994] 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 Web Table III.2 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 Web Figure III.9 in supplementary material).

After a normalization for continuous markers which did not follow a gaussian distribution using splines [Proust-Lima et al., 2017], we modeled independently the markers according to generalized mixed models (see equation III.3) with natural splines on time measurements to capture potentially complex behavior over time [Perperoglou et al., 2019] (see Section III.1.6.3 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 (Web Figure III.10 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. RSF hyperparameters tuning (according to OOB error) is reported in supplementary material Web Figure III.11.

Predictive performances are displayed in Figure III.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 III.4B).

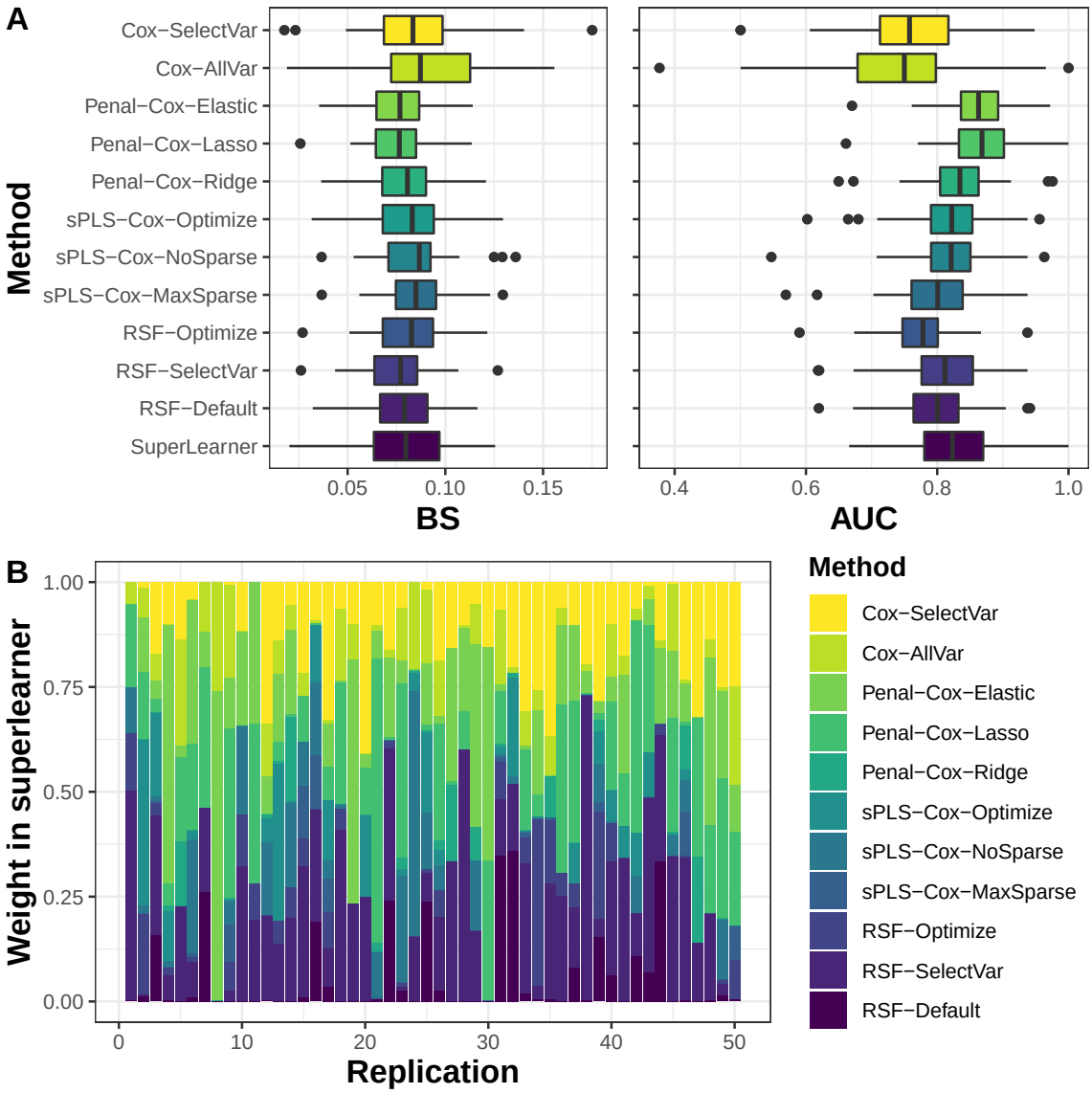



FIGURE III.4 – Assessment (figure A) and weights in superlearner (figure B) of 3-years death survival probability in primary biliary cholangitis patients using information collected up to 4 years over 50 replicates. Methods are assessed using Brier Score (BS) and Area Under the ROC Curve (AUC).

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 me-

thods) nicely illustrating the gain in updating the biomarker information over follow-up (Figure III.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.

#### III.1.3.3 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 [Helmer et al., 2001] 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 [Schneeweiss et al., 2001]. 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 sex) 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 III.1.6.3 and Web Tables III.3/III.4 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

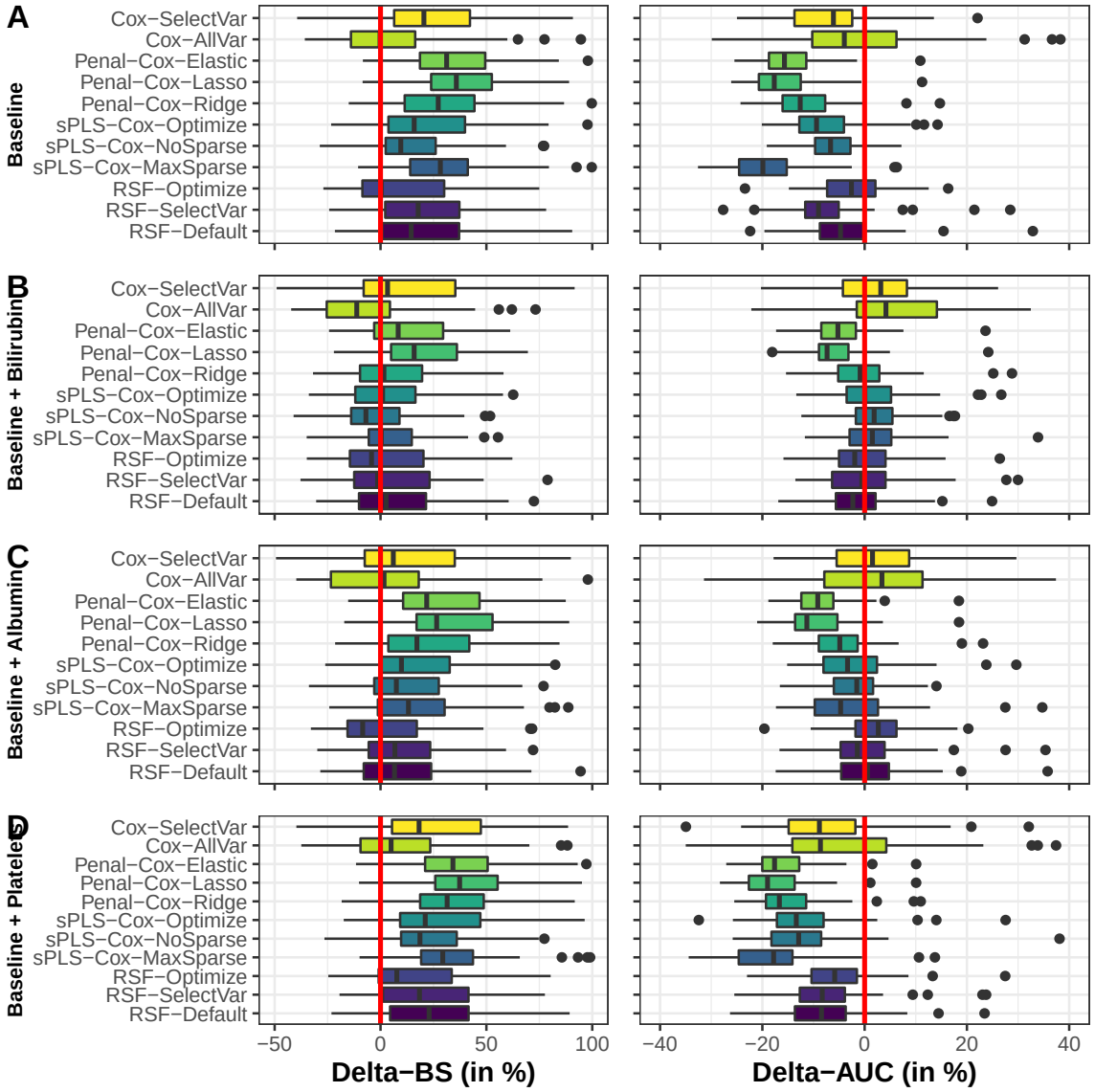



FIGURE III.5 – Assessment of 3-year survival probability in primary biliary cholangitis patients using baseline information on the 11 biomarkers and 3 covariates (figure A), baseline information and repeated measures collected up to 4 years of either serum bilirubin (figure B), albumin (figure C) or 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.

(see flowchart Web Figure III.12 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 Web Figures III.13 and III.14 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 Web Figures III.15 and III.16.

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 Web Figures III.17A and III.18A 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 Web Figures III.17B and III.18B in supplementary material) provided the lowest BS, whereas Cox with variable selection and penalized Cox models gave the highest AUC (0.66 in mean). Comparison to baseline information is also available in Web Figure III.19 in supplementary material.

This application mainly aimed at identifying and contrasting the main age-specific predictors of death at 80 and 85 years old. Figure III.6 reports the VIMP from the optimized RSF (variables selected by the Lasso are shown in supplementary material Web Figure III.20). 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, sex 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 sex.

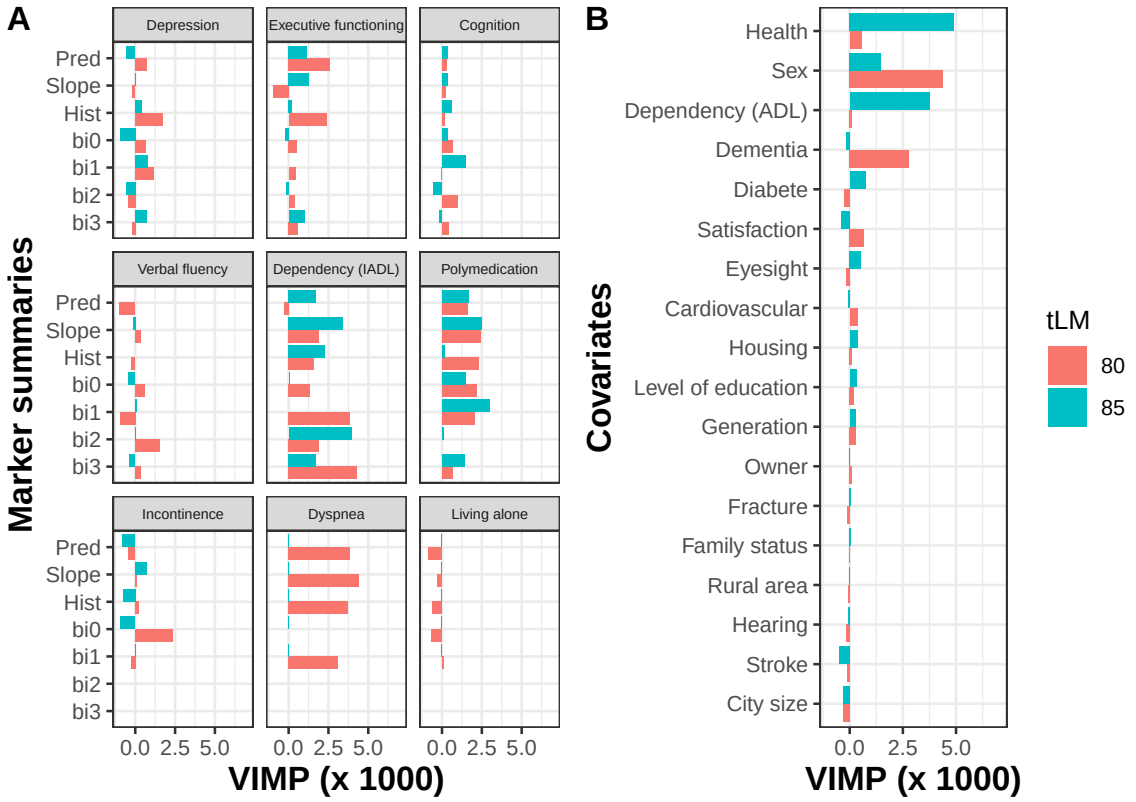



FIGURE III.6 – Variables associated with all-cause death in the *RSF-Optimize* model at 80-year and 85-year landmark age. Are displayed the VIMP value for each marker summaries (figure A) and covariate (figure B). A large VIMP value indicates that the variable is predictive of the event.

### III.1.4 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 adapted to time-to-event data, 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 monitored 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. We precise that given the wellknown overfitting when assessing the predictive accuracy on the same sample as used for the training, we systematically assessed the predictive accuracy on an external sample in the simulations. However, in the real data analyses, in the absence of available external data, we used K-fold cross-validations (repeated 50 times to account for the variability due to the cross-validation partitioning).

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 [Proust-Lima and Taylor, 2009, Rizopoulos, 2011], some authors have recently begun to tackle the problem of large dimension of longitudinal markers [Zhao et al., 2020, Jiang et al., 2021, Tanner et al., 2021]. 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 [Jiang et al., 2021], (ii) proposing the use of many summaries from the biomarkers as individual posterior computation from the longitudinal model (compared for instance to [Tanner et al., 2021] who only include one or two summaries), (iii) exploiting the time-continuous information from survival data rather than discretized scale as in [Tanner et al., 2021], and (iv) considering a vast variety of machine learning techniques as well as a superlearner rather than focusing only on one specific technique [Zhao et al., 2020]. 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 [Yao et al., 2005] 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 [Katzman et al., 2018] or random forests based on pseudo-observations [Zhao et al., 2020]. 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 [Ishwaran et al., 2014].

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, our methodology primarily relies on prediction methods (random forests, penalized regressions, dimension reduction regressions) that were shown to scale very well in high-dimensional context [Hastie et al., 2009]. The preliminary step we added to determine the set of summary features is a univariate mixed model performed independently on each marker. Therefore, it isn't affected by the number of markers. However, in high-dimensional contexts (highly large number of subjects and/or highly large number of markers), we anticipate the method to become computationally very intensive. Reducing the computational time in such high-dimensional contexts remains a future direction of research.

Our work presents the same limitations as any landmark approach. First, only the subjects at risk at the landmark time are considered which can induce a loss of efficiency. In addition, predictions from landmark approaches are not consistent since the time-varying covariate with the time-to-event are not linked at all times [Suresh et al., 2017]. However, the landmark method in an extensive simulation study with one time-varying covariate, the landmark approach was shown to provide very good predictive performances compared to the joint modelling technique and better robustness to misspecification [Ferrer et al.,

2019]. Finally our methodology is limited to 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 [Houwelingen and Putter, 2012] but we leave such development for future research.

## III.1.5 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/hdlandmark and more details are given in software section in supplementary material) 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.

## III.1.6 Web supplementary materials

### III.1.6.1 Prediction methods for survival data

We consider the prediction methods using multiple settings, in particular the choice of the hyperparameters and/or the variable select, resulting to 11 sub-methods in total. In the following, we describe the difference between those sub-methods.

**Cox models** Two cox models are estimated using either all predictors (*Cox-AllVar*) or some predictors (*Cox-SelectVar*) selected using a stepwise backward procedure based on

the AIC statistic.

**Penalized-Cox models** Penalized-cox models require the tuning of 2 parameters : the norm mixing parameter $\alpha$ and the penalty $\lambda$. The penalty $\lambda$ is chosen by minimizing the partial likelihood deviance for a given $\alpha$ using an internal 10-folds cross-validation. We define 3 sub-models according to norm mixing parameter : lasso penalty with $\alpha = 1$ (*Penal-Cox-Lasso*), Ridge penalty with $\alpha = 0$ (*Penal-Cox-Ridge*) or elastic-net penalty with $\alpha \in [0, 1]$ (*Penal-Cox-Elastic*). For the elastic-net penalty, multiple $\alpha$ are evaluated according to a grid from 0 to 1 with a 0.1 step. As a final *Penal-Cox-Elastic* model, we retain the model with the lower partial likelihood deviance over all $\alpha$ from the grid.

**Deviance residuals-based sparse-Partial Least Square** Deviance residuals-based sparse-Partial Least Square models require the tuning of 2 parameters : the number of components $C$ and the sparsity controled by the lasso penalty parameter for each component $\eta$. The number of components $C$ are chosen by maximizing the $iAUCsurvROC$ criteria (from the `plsRcox` R package) for a given $\eta$ using an internal 5-folds cross-validation. We define 3 sub-models according to the lasso penalty parameter : no sparsity with $\eta = 0$ (*sPLS-NoSparse*), maximum sparsity with $\eta = 0.9$ (*sPLS-MaxSparse*) or mixing sparsity with $\eta \in [0, 0.9]$ (*sPLS-Optimize*). For the mixing sparsity, multiple $\eta$ are evaluated according to a grid from 0 to 0.9 with a 0.1 step. As a final *sPLS-Optimize* model, we retain the model with the higher $iAUCsurvROC$ over all $\eta$ from the grid.

**Random Survival Forests** Random survival forest methods require the tuning of 2 parameters : the number of predictors drawn at each node $M$ and the minimal node size $S$. We defined 3 sub-methods according to these parameters. The *RSF-Default* method uses the default parameters with $M$ equals to the square root of the number of predictors and $S = 15$. In the *RSF-Optimize* and the *RSF-SelectVar* methods, the parameters are tuned according to a grid of values, $M$ from 5 to the maximum of predictors with a 5 step and $S \in \{3, 15\}$. The best parameters are chosen by minimizing the out-of-bag error

based on the $1 - C$, where C is Harrell's concordance index. For the *RSF-SelectVar*, a final random survival forest is computed using the predictors with $VIMP > 0.005$.

#### III.1.6.2 Simulations

**Model formulation for time-dependent markers** For all scenarios described in the main manuscript, the 17 time-dependent markers were generated up to $t_{LM} = 4$ according to a linear mixed model defined as :

Marker 1

$$Y_{i1}(t_{ij}) = \beta_0 + \beta_1 * t_{ij} + \beta_2 * t_{ij}^2 + b_{i0} + b_{i1} * t_{ij} + b_{i2} * t_{ij}^2 + \epsilon_{ij} \tag{III.10}$$

where the fixed coefficients $\beta_0 = 1.5$, $\beta_1 = 2.0$, $\beta_2 = -1.2$, the random effects $b_i = (b_{i0}, b_{i1}, b_{i2})^\top \sim \mathcal{N}(0, D)$ with $D = \begin{pmatrix} 1.5^2 & 0 & 0 \\ 0 & 0.8^2 & 0 \\ 0 & 0 & 0.5^2 \end{pmatrix}$ and the measurement error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 1.1$.

Marker 2

$$Y_{i2}(t_{ij}) = \beta_0 + \beta_1 * \log(t_{ij} + 0.1) + b_{i0} + b_{i1} * \log(t_{ij} + 0.1) + \epsilon_{ij} \tag{III.11}$$

where the fixed coefficients $\beta_0 = 5.5$, $\beta_1 = -1.5$, the random effects $b_i = (b_{i0}, b_{i1})^\top \sim \mathcal{N}(0, D)$ with $D = \begin{pmatrix} 1.4^2 & 0 \\ 0 & 0.6^2 \end{pmatrix}$ and the measurement error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 0.9$.

Marker 3

$$Y_{i3}(t_{ij}) = \beta_0 + \beta_1 * \sqrt{(t_{ij} + 0.1)} + b_{i0} + b_{i1} * \sqrt{(t_{ij} + 0.1)} + \epsilon_{ij} \tag{III.12}$$

where the fixed coefficients $\beta_0 = 2.5$, $\beta_1 = 1.8$, the random effects $b_i = (b_{i0}, b_{i1})^\top \sim \mathcal{N}(0, D)$ with $D = \begin{pmatrix} 0.7^2 & 0 \\ 0 & 0.7^2 \end{pmatrix}$ and the measurement error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 1.1$.

Marker 4

$$Y_{i4}(t_{ij}) = \beta_0 + \beta_1 * t_{ij} + b_{i0} + b_{i1} * t_{ij} + \epsilon_{ij} \tag{III.13}$$

where the fixed coefficients $\beta_0 = 3.0$, $\beta_1 = 1.2$, the random effects $b_i = (b_{i0}, b_{i1})^\top \sim \mathcal{N}(0, D)$ with $D = \begin{pmatrix} 0.7^2 & 0 \\ 0 & 0.5^2 \end{pmatrix}$ and the measurement error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 1.3$.

Marker 5

$$Y_{i5}(t_{ij}) = \beta_0 + \beta_1 * t_{ij}^2 + b_{i0} + b_{i1} * t_{ij}^2 + \epsilon_{ij} \tag{III.14}$$

where the fixed coefficients $\beta_0 = 0.0$, $\beta_1 = 0.7$, the random effects $b_i = (b_{i0}, b_{i1})^\top \sim \mathcal{N}(0, D)$ with $D = \begin{pmatrix} 0.3^2 & 0 \\ 0 & 0.2^2 \end{pmatrix}$ and the measurement error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 0.9$.

Marker 6

$$Y_{i6}(t_{ij}) = \beta_0 + \beta_1 * t_{ij} + \beta_2 * t_{ij}^2 + b_{i0} + b_{i1} * t_{ij} + b_{i2} * t_{ij}^2 + \epsilon_{ij} \tag{III.15}$$

where the fixed coefficients $\beta_0 = 3.5$, $\beta_1 = -1.2$, $\beta_2 = 0.8$, the random effects $b_i = (b_{i0}, b_{i1}, b_{i2})^\top \sim \mathcal{N}(0, D)$ with $D = \begin{pmatrix} 1.5^2 & 0 & 0 \\ 0 & 0.7^2 & 0 \\ 0 & 0 & 0.5^2 \end{pmatrix}$ and the measurement error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 1.1$.

Marker 7

$$Y_{i7}(t_{ij}) = \beta_0 + \beta_1 * t_{ij} + b_{i0} + b_{i1} * t_{ij} + \epsilon_{ij} \tag{III.16}$$

where the fixed coefficients $\beta_0 = 1.1$, $\beta_1 = 0.8$, the random effects $b_i = (b_{i0}, b_{i1})^\top \sim \mathcal{N}(0, D)$ with $D = \begin{pmatrix} 0.3^2 & 0 \\ 0 & 0.5^2 \end{pmatrix}$ and the measurement error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 0.5$.

Marker 8

$$Y_{i8}(t_{ij}) = \beta_0 + \beta_1 * \exp(t_{ij}) + b_{i0} + b_{i1} * \exp(t_{ij}) + \epsilon_{ij} \tag{III.17}$$

where the fixed coefficients $\beta_0 = 1.1$, $\beta_1 = -0.2$, the random effects $b_i = (b_{i0}, b_{i1})^\top \sim$

$\mathcal{N}(0, D)$ with $D = \begin{pmatrix} 0.3^2 & 0 \\ 0 & 0.1^2 \end{pmatrix}$ and the measurement error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 0.5$.

Marker 9

$$Y_{i9}(t_{ij}) = \beta_0 + \beta_1 * \log(t_{ij} + 0.1) + \beta_2 * t_{ij}^2 + b_{i0} + b_{i1} * \log(t_{ij} + 0.1) + b_{i2} * t_{ij}^2 + \epsilon_{ij} \quad \text{(III.18)}$$

where the fixed coefficients $\beta_0 = 6.5$, $\beta_1 = 4.5$, $\beta_2 = -1.0$, the random effects $b_i = (b_{i0}, b_{i1}, b_{i2})^\top \sim \mathcal{N}(0, D)$ with $D = \begin{pmatrix} 1.2^2 & 0 & 0 \\ 0 & 2.5^2 & 0 \\ 0 & 0 & 0.3^2 \end{pmatrix}$ and the measurement error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 0.9$.

Marker 10

$$Y_{i10}(t_{ij}) = \beta_0 + \beta_1 * t_{ij} + \beta_2 * t_{ij}^2 + b_{i0} + b_{i1} * t_{ij} + b_{i2} * t_{ij}^2 + \epsilon_{ij} \quad \text{(III.19)}$$

where the fixed coefficients $\beta_0 = 4.1$, $\beta_1 = -2.0$, $\beta_2 = 0.9$, the random effects $b_i = (b_{i0}, b_{i1}, b_{i2})^\top \sim \mathcal{N}(0, D)$ with $D = \begin{pmatrix} 1.5^2 & 0 & 0 \\ 0 & 1.1^2 & 0 \\ 0 & 0 & 0.4^2 \end{pmatrix}$ and the measurement error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 1.1$.

Marker 11

$$Y_{i11}(t_{ij}) = \beta_0 + \beta_1 * t_{ij} + \beta_2 * t_{ij}^2 + b_{i0} + b_{i1} * t_{ij} + b_{i2} * t_{ij}^2 + \epsilon_{ij} \quad \text{(III.20)}$$

where the fixed coefficients $\beta_0 = 9.4$, $\beta_1 = -1.2$, $\beta_2 = -0.7$, the random effects $b_i = (b_{i0}, b_{i1}, b_{i2})^\top \sim \mathcal{N}(0, D)$ with $D = \begin{pmatrix} 0.9^2 & 0 & 0 \\ 0 & 0.7^2 & 0 \\ 0 & 0 & 0.8^2 \end{pmatrix}$ and the measurement error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 1.1$.

Marker 12

$$Y_{i12}(t_{ij}) = \beta_0 + \beta_1 * \sqrt{(t_{ij} + 0.1)} + b_{i0} + b_{i1} * \sqrt{(t_{ij} + 0.1)} + \epsilon_{ij} \quad \text{(III.21)}$$

where the fixed coefficients $\beta_0 = 5.2$, $\beta_1 = 4.7$, the random effects $b_i = (b_{i0}, b_{i1})^\top \sim \mathcal{N}(0, D)$ with $D = \begin{pmatrix} 0.9^2 & 0 \\ 0 & 1.1^2 \end{pmatrix}$ and the measurement error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 0.9$.

Marker 13

$$Y_{i13}(t_{ij}) = \beta_0 + \beta_1 * \sqrt{(t_{ij} + 0.1)} + b_{i0} + b_{i1} * \sqrt{(t_{ij} + 0.1)} + \epsilon_{ij} \quad \text{(III.22)}$$

where the fixed coefficients $\beta_0 = 8.2$, $\beta_1 = -3.2$, the random effects $b_i = (b_{i0}, b_{i1})^\top \sim \mathcal{N}(0, D)$ with $D = \begin{pmatrix} 1.3^2 & 0 \\ 0 & 1.6^2 \end{pmatrix}$ and the measurement error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 1.3$.

Marker 14

$$Y_{i14}(t_{ij}) = \beta_0 + \beta_1 * t_{ij} + \beta_2 * t_{ij}^3 + b_{i0} + b_{i1} * t_{ij} + b_{i2} * t_{ij}^3 + \epsilon_{ij} \quad \text{(III.23)}$$

where the fixed coefficients $\beta_0 = 3.6$, $\beta_1 = -0.9$, $\beta_2 = 0.4$, the random effects $b_i = (b_{i0}, b_{i1}, b_{i2})^\top \sim \mathcal{N}(0, D)$ with $D = \begin{pmatrix} 1.5^2 & 0 & 0 \\ 0 & 0.5^2 & 0 \\ 0 & 0 & 0.1^2 \end{pmatrix}$ and the measurement error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 1.1$.

Marker 15

$$Y_{i15}(t_{ij}) = \beta_0 + \beta_1 * t_{ij} + \beta_2 * t_{ij}^3 + b_{i0} + b_{i1} * t_{ij} + b_{i2} * t_{ij}^3 + \epsilon_{ij} \quad \text{(III.24)}$$

where the fixed coefficients $\beta_0 = 8.6$, $\beta_1 = 4.9$, $\beta_2 = -0.4$, the random effects $b_i =$

$(b_{i0}, b_{i1}, b_{i2})^\top \sim \mathcal{N}(0, D)$ with $D = \begin{pmatrix} 1.5^2 & 0 & 0 \\ 0 & 0.7^2 & 0 \\ 0 & 0 & 0.2^2 \end{pmatrix}$ and the measurement error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 1.1$.

Marker 16

$$Y_{i16}(t_{ij}) = \beta_0 + \beta_1 * \exp(t_{ij}) + b_{i0} + b_{i1} * \exp(t_{ij}) + \epsilon_{ij} \tag{III.25}$$

where the fixed coefficients $\beta_0 = 4.1$, $\beta_1 = -0.2$, the random effects $b_i = (b_{i0}, b_{i1})^\top \sim \mathcal{N}(0, D)$ with $D = \begin{pmatrix} 0.6^2 & 0 \\ 0 & 0.2^2 \end{pmatrix}$ and the measurement error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 0.5$.

Marker 17

$$Y_{i17}(t_{ij}) = \beta_0 + \beta_1 * t_{ij}^3 + b_{i0} + b_{i1} * t_{ij}^3 + \epsilon_{ij} \tag{III.26}$$

where the fixed coefficients $\beta_0 = 3.2$, $\beta_1 = 0.3$, the random effects $b_i = (b_{i0}, b_{i1})^\top \sim \mathcal{N}(0, D)$ with $D = \begin{pmatrix} 0.6^2 & 0 \\ 0 & 0.2^2 \end{pmatrix}$ and the measurement error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 1.0$.

We display an example of individual trajectories for all markers in Fig. III.7. In the following for the sake of simplicity, we denote the summaries $Y_k^{pred}$, $Y_k^{slope}$, $Y_k^{hist}$ for respectively the current level, current slope and history level on landmark time for the marker $k$.

**Model formulation for time-to-event** We generated the hazard function $\lambda_i$ according to a proportional hazard model defined as :

$$\lambda_i(t) = \lambda_0(b, c, t) \exp(\mathcal{P}_i) \tag{III.27}$$

With $\lambda_0(b, c, t) = cb^c t^{c-1}$ the baseline hazard function from a Weibull distribution with parameters $b$ and $c$, and $\mathcal{P}_i$ the linear predictor. Using the standard uniform distribution

$u \sim \mathcal{U}(0,1)$ [Bender et al., 2005], we generated time-to-event $T_i$ defined as :

$$T_i = \frac{1}{b} * \left( -\frac{\log u}{\exp\left(\mathcal{P}_i\right)} \right)^{1/c} \tag{III.28}$$

In the following, we detail the parameters used to build the hazard function from equation III.27 in each of the scenarios. A recap of the summaries used in each scenario is available in the table III.1.

Scenario 1 : few summaries with linear association

$$\mathcal{P}_i = \gamma_1 * Y_{1i}^{pred} + \gamma_2 * Y_{3i}^{hist} + \gamma_1 * Y_{8i}^{pred} + \gamma_2 * Y_{15i}^{pred} \tag{III.29}$$

With $\gamma_1 = -0.5$ and $\gamma_2 = 0.5$ the coefficients associated to the summaries and $b = 0.3$ and $c = 6.5$ to build the base hazard function $\lambda_0$. Results from this scenario are given in Fig. III.8A.

Scenario 2 : few summaries with linear association with interaction

$$\begin{aligned} \mathcal{P}_i &= \gamma_1 * Y_{1i}^{pred} + \gamma_2 * Y_{3i}^{hist} + \gamma_1 * Y_{8i}^{pred} + \gamma_2 * Y_{15i}^{pred} + \gamma_1 * Y_{1i}^{pred} * Y_{3i}^{hist} \\ &+ \gamma_2 * Y_{8i}^{pred} * Y_{15i}^{pred} + \gamma_1 * Y_{1i}^{pred} * Y_{8i}^{pred} + \gamma_2 * Y_{3i}^{hist} * Y_{15i}^{pred} \end{aligned} \tag{III.30}$$

With $\gamma_1 = -0.5$ and $\gamma_2 = 0.5$ denote the coefficients associated to the summaries and $b = 0.3$ and $c = 6.5$ to build hazard function $\lambda_0$. Results from this scenario are given in Fig. III.8B.

Scenario 3 : few summaries with non-linear association

$$\mathcal{P}_i = \gamma_1 * (Y_{1i}^{pred})^2 + \gamma_1 * (Y_{1i}^{slope})^2 + \gamma_1 * (Y_{4i}^{slope})^2 + \gamma_1 * (Y_{10i}^{pred})^2 \tag{III.31}$$

With $\gamma_1 = 0.5$ denotes the coefficients associated to the summaries and $b = 0.25$ and $c = 6.5$ to build hazard function $\lambda_0$. Results from this scenario are given in Fig. III.8C.

Scenario 4 : many summaries with linear association

$$\begin{aligned}\mathcal{P}_i &= \gamma_1 * Y_{1i}^{pred} + \gamma_2 * Y_{1i}^{slope} + \gamma_1 * Y_{3i}^{pred} + \gamma_2 * Y_{3i}^{hist} + \gamma_1 * Y_{4i}^{pred} + \gamma_2 * Y_{4i}^{slope} \\ &+ \gamma_1 * Y_{5i}^{pred} + \gamma_2 * Y_{5i}^{slope} + \gamma_1 * Y_{5i}^{hist} + \gamma_2 * Y_{10i}^{pred} + \gamma_1 * Y_{10i}^{hist} + \gamma_2 * Y_{13i}^{pred} \\ &+ \gamma_1 * Y_{13i}^{slope} + \gamma_2 * Y_{15i}^{pred} + \gamma_1 * Y_{15i}^{slope} + \gamma_2 * Y_{15i}^{hist} + \gamma_1 * Y_{17i}^{pred} + \gamma_2 * Y_{17i}^{slope}\end{aligned} \quad \text{(III.32)}$$

With $\gamma_1 = -0.5$ and $\gamma_2 = 0.5$ denote the coefficients associated to the summaries and $b = 0.3$ and $c = 6.5$ to build hazard function $\lambda_0$. Results from this scenario are given in the main manuscript.

Scenario 5 : many summaries with non-linear association

$$\begin{aligned}\mathcal{P}_i &= \gamma_1 * (Y_{1i}^{pred})^2 + \gamma_1 * (Y_{1i}^{slope})^2 + \gamma_2 * (Y_{2i}^{slope})^2 + \gamma_2 * (Y_{3i}^{hist})^2 + \gamma_1 * (Y_{4i}^{pred})^2 \\ &+ \gamma_2 * (Y_{4i}^{slope})^2 + \gamma_1 * (Y_{10i}^{pred})^2 + \gamma_2 * (Y_{11i}^{hist})^2 + \gamma_1 * (Y_{12i}^{slope})^2 + \gamma_2 * (Y_{13i}^{pred})^2 \\ &+ \gamma_1 * (Y_{14i}^{slope})^2 + \gamma_2 * \mathbb{1}(Y_{15i}^{pred} > \overset{\sim}{Y}_{15}^{pred}) + \gamma_1 * \mathbb{1}(Y_{16i}^{hist} > \overset{\sim}{Y}_{16}^{hist}) \\ &+ \gamma_2 * \mathbb{1}(Y_{17i}^{pred} > \overset{\sim}{Y}_{17}^{pred}) + \gamma_1 * \mathbb{1}(Y_{5i}^{slope} > \overset{\sim}{Y}_{5}^{slope}) + \gamma_2 * \mathbb{1}(Y_{6i}^{hist} > \overset{\sim}{Y}_{6}^{hist}) \\ &+ \gamma_1 * \mathbb{1}(Y_{9i}^{slope} > \overset{\sim}{Y}_{9}^{slope}) + \gamma_2 * \mathbb{1}(Y_{9i}^{hist} > \overset{\sim}{Y}_{9}^{hist})\end{aligned} \quad \text{(III.33)}$$

With $\gamma_1 = -0.5$ and $\gamma_2 = 0.5$ denote the coefficients associated to the summaries and $b = 0.28$ and $c = 5.5$ to build hazard function $\lambda_0$. $\overset{\sim}{Y}_k$ represents the median for marker $k$. Results from this scenario are given in the main manuscript.

#### III.1.6.3 Applications

**Prediction of death in primary billiary cholangitis** To estimate the probability of death on primary biliary cholangitis patients, we used 7 continuous time-dependent markers measuring bilirubin, cholesterol, albumin, alkaline, SGOT, platelets and prothrombin and 4 binary time-dependent markers measuring the presence of ascites, hepatomegaly, spiders and edema. Except albumin, all continuous variables were normalized using splines to follow a gaussian distribution [Philipps et al., 2014].

Except ascites and edema, we modeled the variables using generalized mixed model

defined as :

$$g(E(Y_{ij}|b_i)) = \beta_0 + \sum_{l=1}^{L} \beta_l * f_l(t, L) + b_{i0} + \sum_{l=1}^{L} b_{il} * f_l(t, L) \tag{III.34}$$

With $g(.)$ the link function taking into account the nature of the marker, $\beta_0$ and $\beta_l$ the fixed coefficients and the random effects $b_{il} = (b_{i0}, b_{il})^\top \sim \mathcal{N}(0, D)$ with $D$ a covariance matrix. $f(t, L)$ denotes the natural splines function with $L$ knots.

To model continuous markers, we use $g(.)$ as the identity function, $D$ an unstructured covariance matrix and $L = 3$ for the natural splines function with two internal knots placed at $t = 0.5$ and $t = 2.0$ and boundary knots at $t = 0$ and $t = 4$. To model binary markers, we use $g(.)$ as the logit function, $D$ a diagonal independent covariance matrix and $L = 2$ for the natural splines function with a single internal knot placed at $t = 1.0$ and boundary knots at $t = 0$ and $t = 4$.

Finally, to avoid convergence issues, ascites and edema are defined as :

$$g(E(Y_{ij}|b_i)) = \beta_0 + \beta_1 * t_{ij} + b_{i0} + b_{i1} * t_{ij} \tag{III.35}$$

With $g(.)$ the logit function, $\beta_0$ and $\beta_l$ the fixed coefficients and the random effects $b_{il} = (b_{i0}, b_{i1})^\top \sim \mathcal{N}(0, D)$ with $D$ an unstructured covariance matrix.

**Prediction of 5-years death at 80 and 85 years old**

Model specification

To estimate the probability of death at 80 and 85 years old, we used 6 continuous time-dependent markers (measuring depression, executive functioning, cognition, speed on fluency, dependency and polymedication) and 3 binary time-dependent markers (measuring the presence of incontinence, dyspnea and living alone). Except executive functioning and speed on fluency, all continuous variables were normalized using splines to follow a gaussian distribution [Philipps et al., 2014]. We modeled the variables from 5 years prior

the landmark time using generalized mixed model defined as :

$$g(E(Y_{ij}|b_i)) = \beta_0 + \sum_{l=1}^{L} \beta_l * f_l(t, L) + b_{i0} + \sum_{l=1}^{L} b_{il} * f_l(t, L) \tag{III.36}$$

With $g(.)$ the link function taking into account the nature of the marker, $\beta_0$ and $\beta_l$ the fixed coefficients and the random effects $b_{il} = (b_{i0}, b_{il})^\top \sim \mathcal{N}(0, D)$ with $D$ a covariance matrix. $f(t, L)$ denotes the natural splines function with $L$ knots.

To model continuous markers, we use $g(.)$ as the identity function, $D$ an unstructured covariance matrix and $L = 3$ for the natural splines function with two internal knots placed at $t = 1.7$ and $t = 3.4$ and boundary knots at $t = 0$ and $t = 5$.

To avoid convergence issues, we model binary markers as :

$$g(E(Y_{ij}|b_i)) = \beta_0 + \beta_1 * t_{ij} + b_{i0} + b_{i1} * t_{ij} \tag{III.37}$$

With $g(.)$ the logit function, $\beta_0$ and $\beta_l$ the fixed coefficients and the random effects $b_{il} = (b_{i0}, b_{i1})^\top \sim \mathcal{N}(0, D)$ with $D$ a diagonal independent covariance matrix.

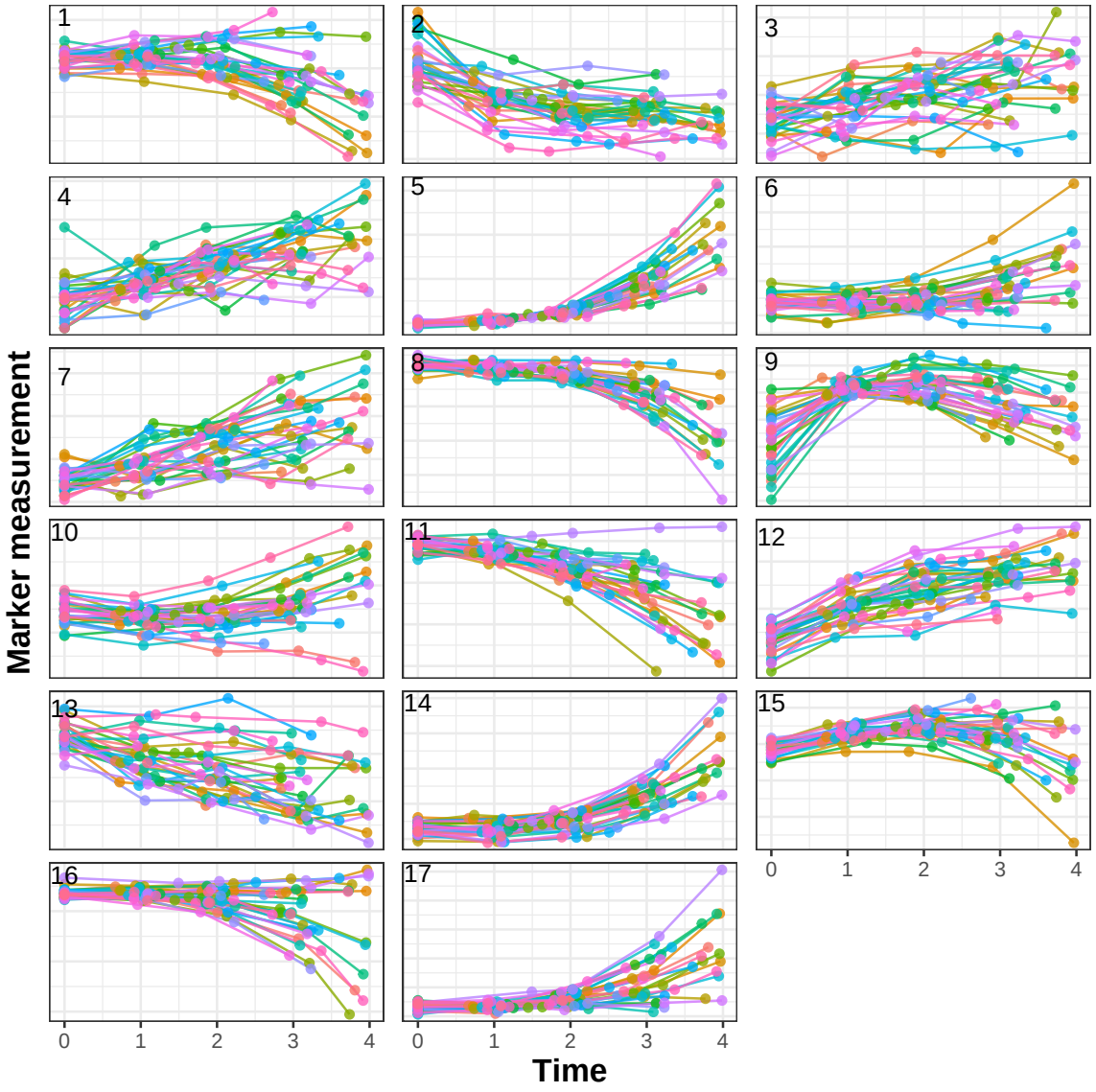
Age-specific predictors of death using Cox model with Lasso penalty

Fig. III.20 displays the predictors selected for Cox model with Lasso penalty. We can see that many time-dependent markers are associated with death, especially with $t_{LM} = 80$. In addition, among these markers, we found 3 predictors of death (measuring executive functioning, dependency and polymedication) at both landmark time 80 and 85 years old. Variables measuring dependency and polymedication are strongly predictive of death at $t_{LM} = 80$. Indeed, 3 summaries were selected by the model. For time-independent variables, sex, history of dementia and dependency are predictive of death at both landmark time 80 and 85 years old.

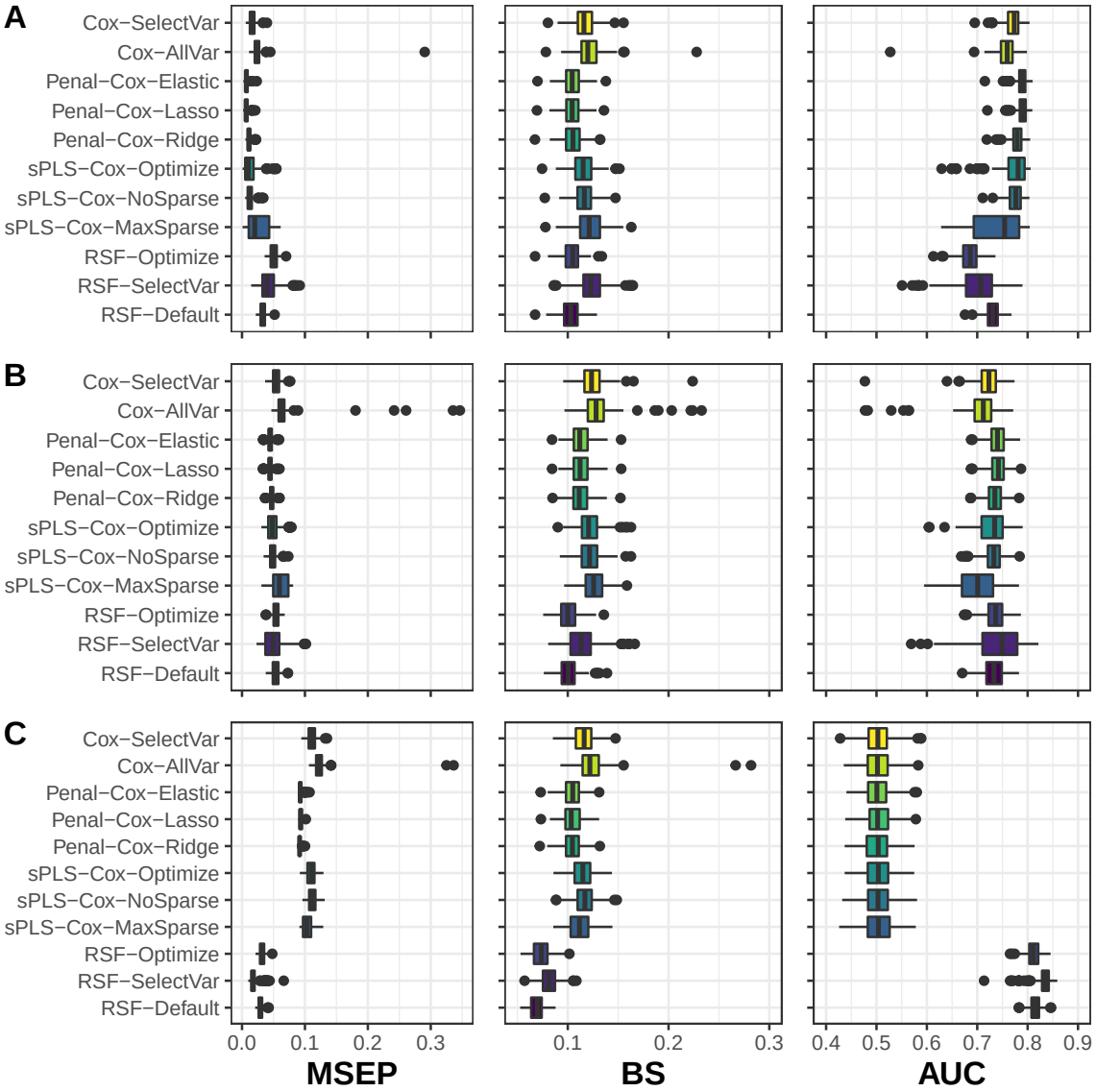
#### III.1.6.4 Software

All analysis were performed using R software version 3.6. We used `lcmm` [Proust-Lima et al., 2017] (for continuous markers) and `lme4` [Bates et al., 2015] (for binary markers) to compute generalized mixed models. Predictions are computed using `survival` [Terry M. Therneau and Patricia M. Grambsch, 2000] for Cox model, `glmnet` [Simon et al., 2011] for Cox model with penalty (with 2 hyperparameters : $\lambda$ and $\alpha$ for respectively strength of the penalty and the type of penalty), `plsRcox` [Bastien et al., 2015] for the Deviance residuals-based sparse-Partial Least Square (with 2 hyperparameters : $\eta$ and *ncomp* for respectively sparsity parameter and the number of components) and `randomForestSRC` [Ishwaran et al., 2008] for random survival forests (with 2 hyperparameters : *mtry* and *nodesize* for respectively the number of variables randomly selected as candidates for splitting a node and the forest average number of unique cases in a terminal node). R code detail and example can be found on https ://github.com/anthonydevaux/hdlandmark.

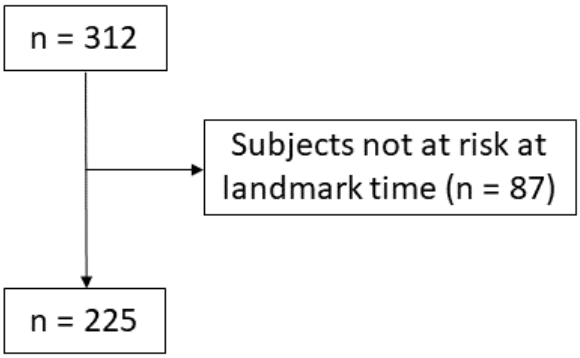
### III.1.6.5 Figures



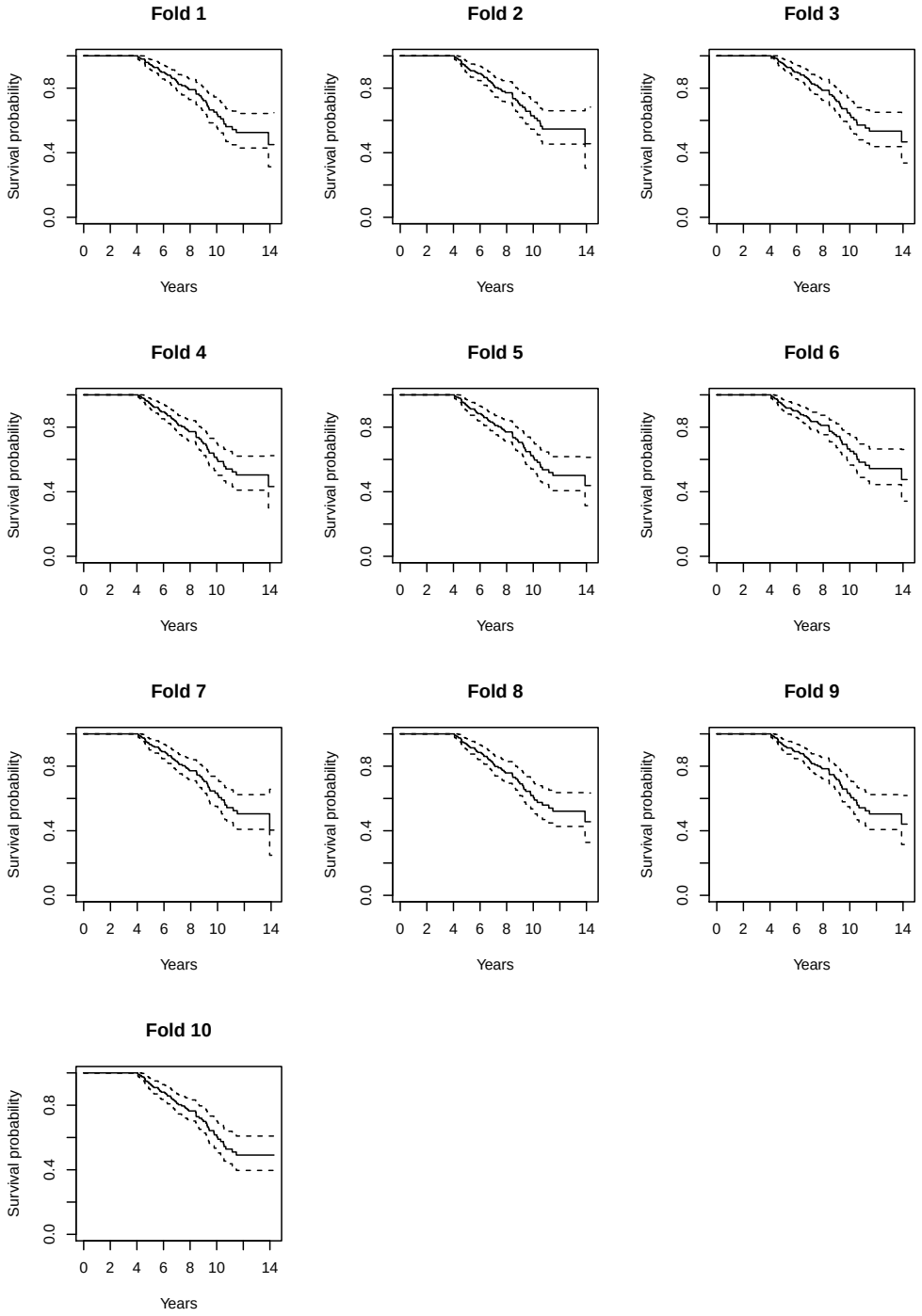

WEB FIGURE III.7 – Illustration of 30 randomly selected individual trajectories chosen randomly for the 17 markers generated up to $t_{LM} = 4$ in the simulation study.



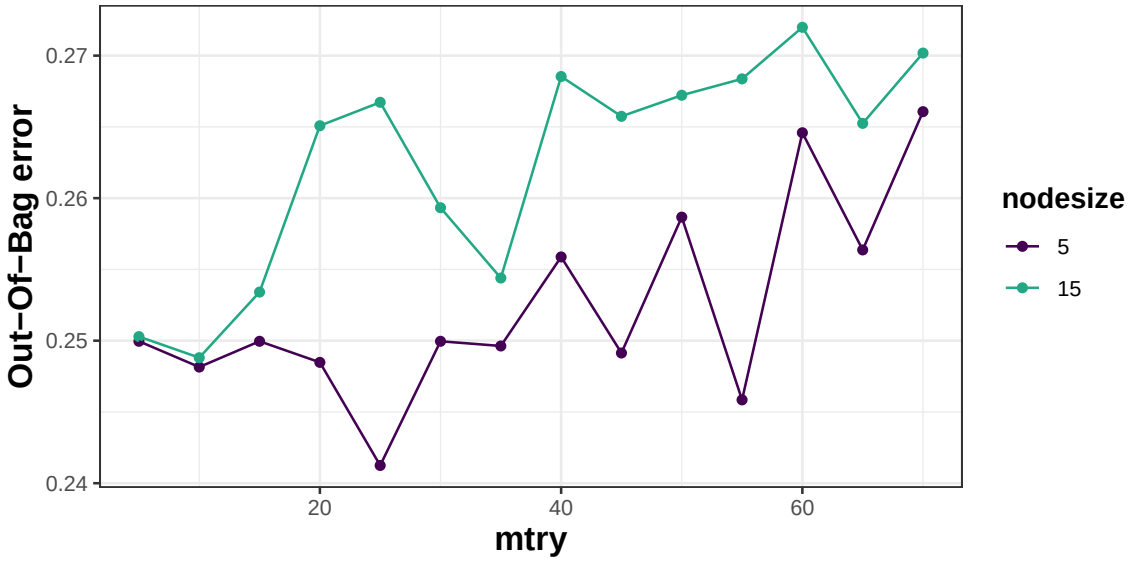

WEB FIGURE III.8 – Scenario results with 4 summaries associated to the event with linear form (figure A), linear form with interaction (figure B) and non-linear form (figure C) over 250 replicates. Methods are assessed using at 3 years Mean Square Error of Prediction (MSEP), Brier Score (BS) and Area Under the ROC Curve (AUC).



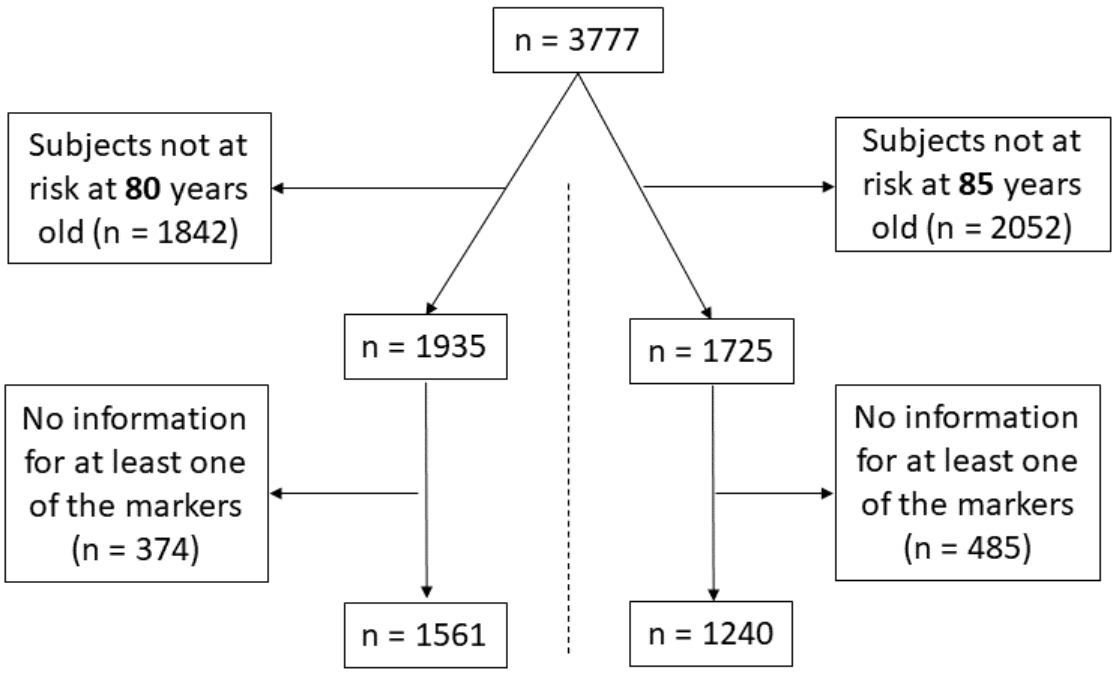

Web Figure III.9 – PBC data flowchart



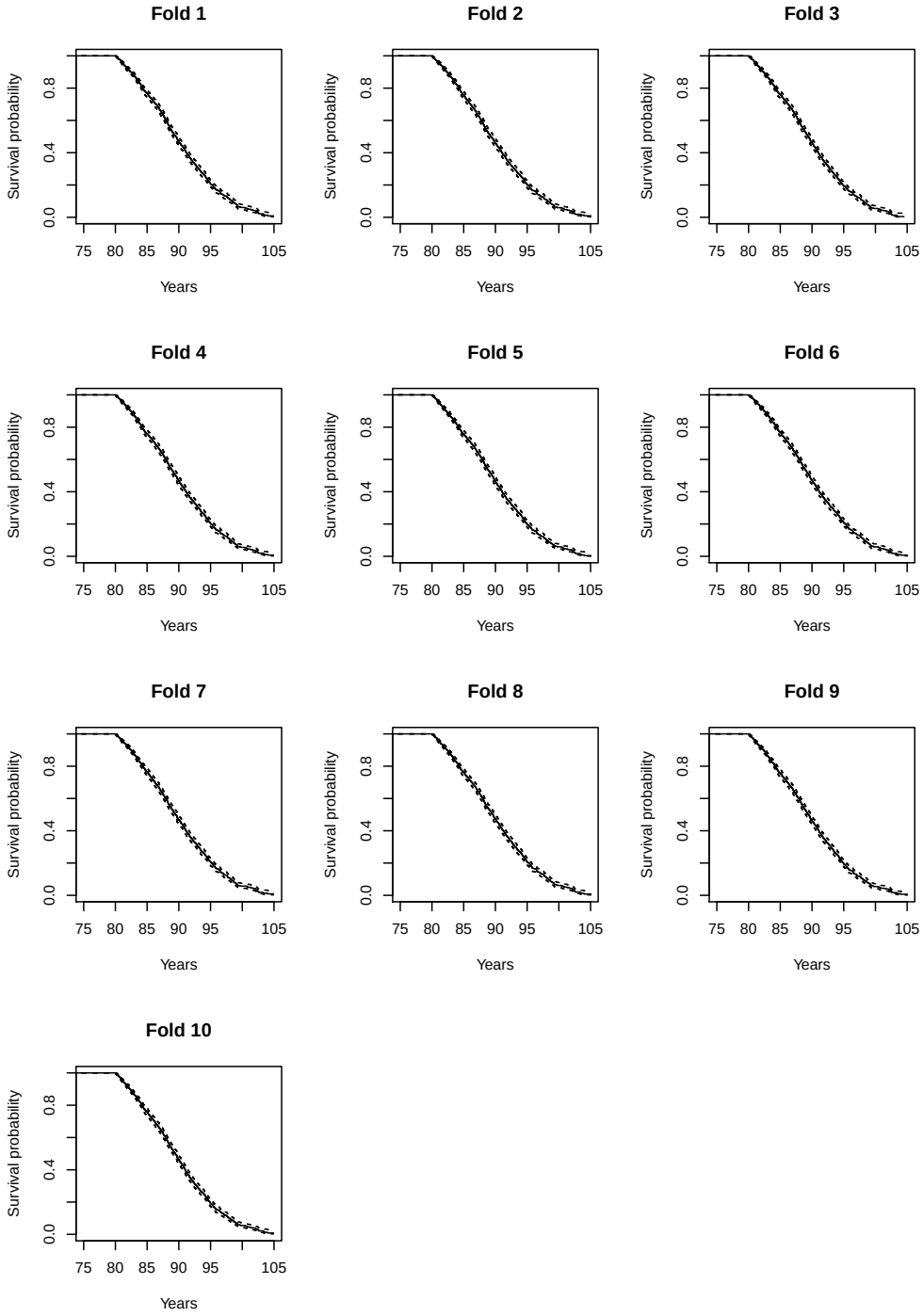

WEB FIGURE III.10 – 3-year survival probability estimated by Kaplan-Meier in each of the 10 folds for PBC subjects still at risk at landmark time $t_{LM} = 4$.



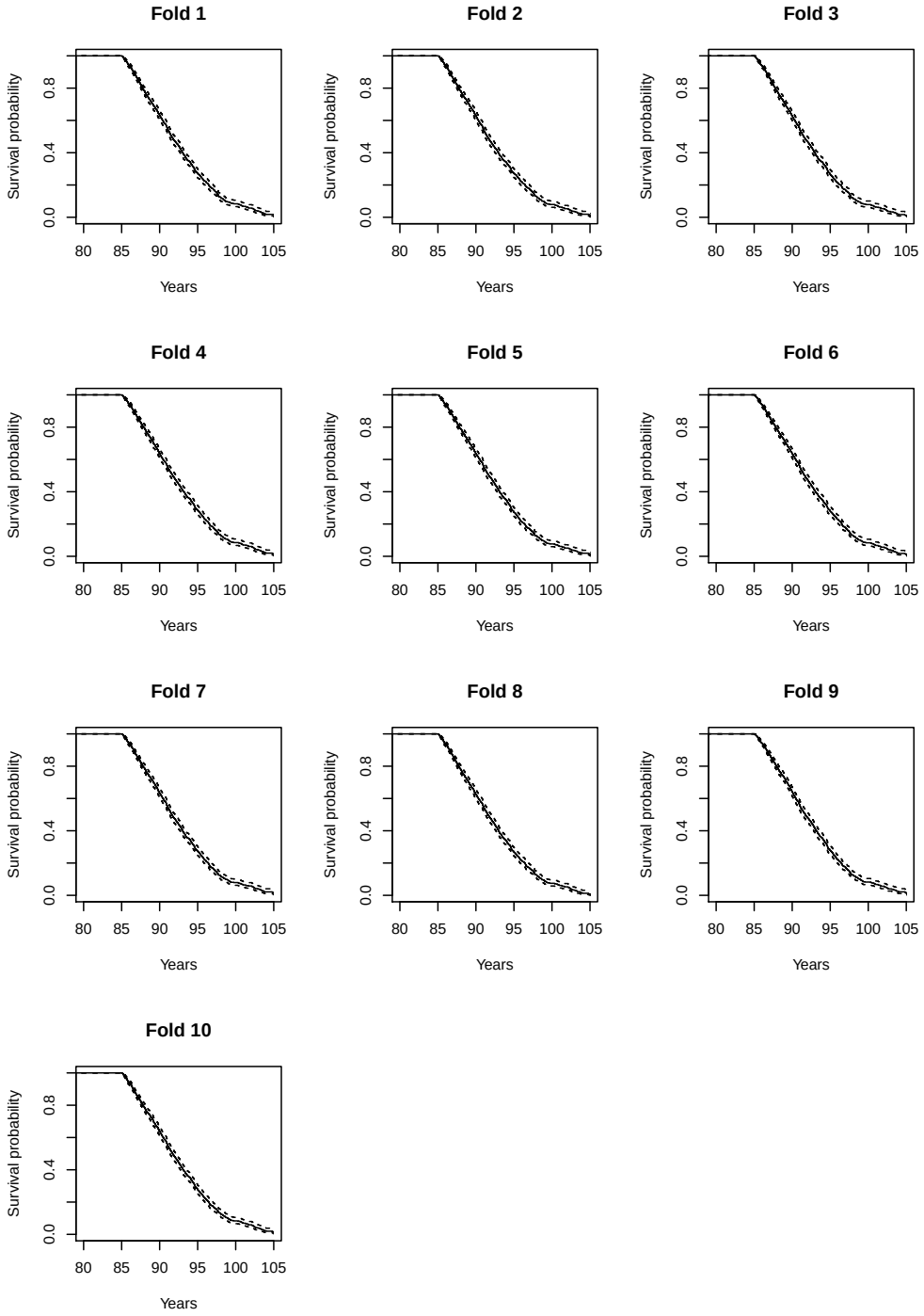

WEB FIGURE III.11 – Random survival forest hyperparameters tuning on primary biliary cholangitis patients at landmark time $t_{LM} = 4$. The best hyperparameters ($mtry = 25$ and $nodesize = 5$) are chosen by minimizing the out-of-bag error.



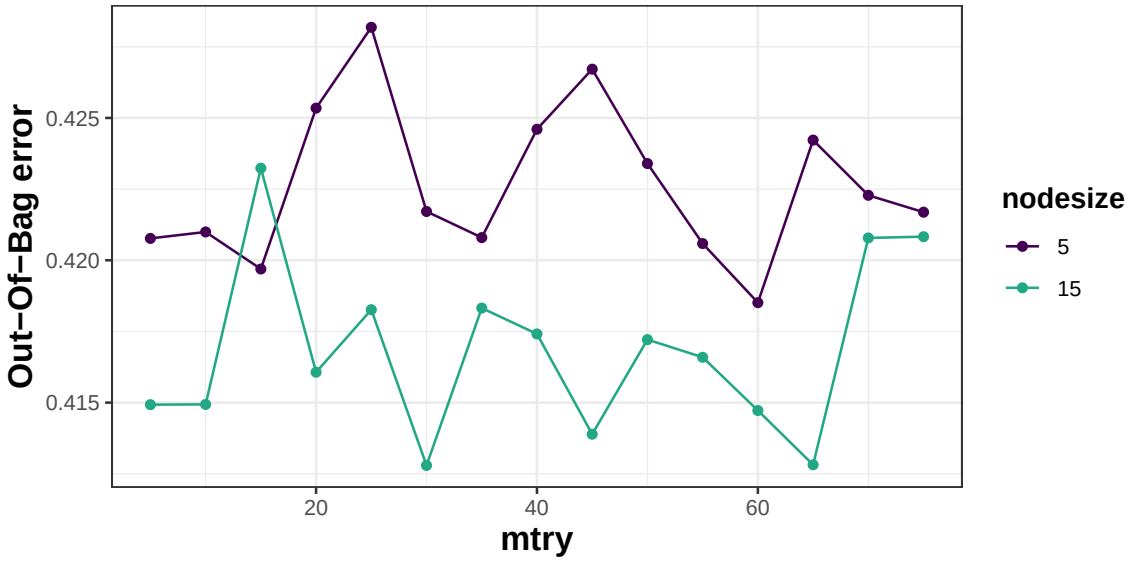

WEB FIGURE III.12 – Flowchart for Paquid application with a landmark times at 80 (left) and 85 (right) years old.



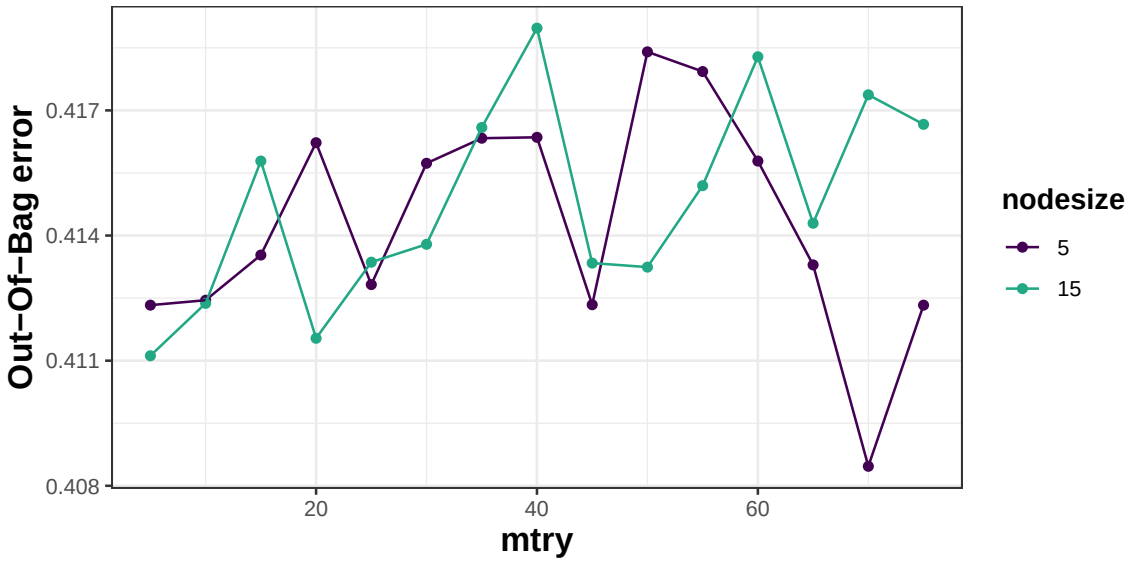

WEB FIGURE III.13 – Survival probability estimated by Kaplan-Meier over the 10 folds for elderly people still at risk at landmark time $t_{LM} = 80$.



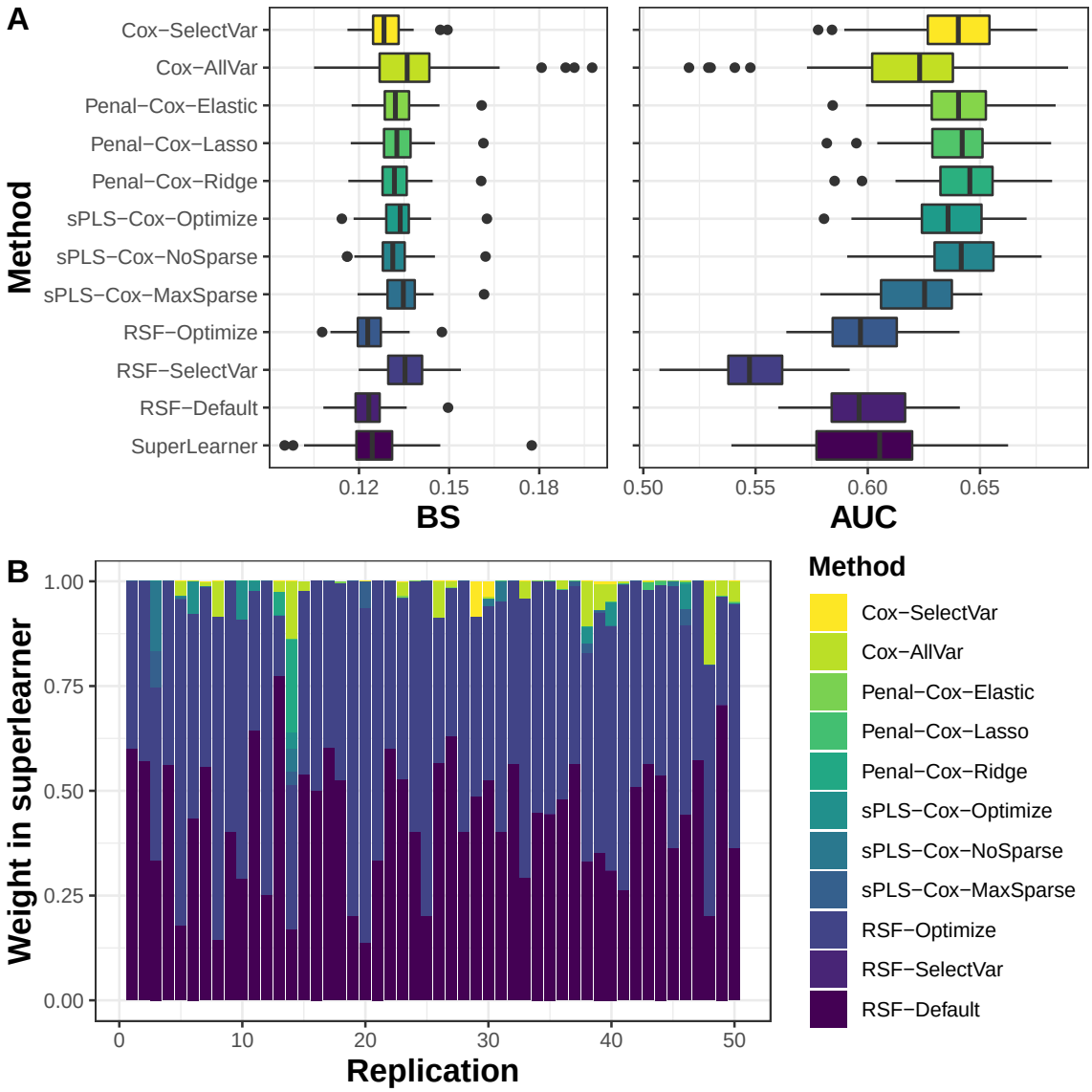

WEB FIGURE III.14 – Survival probability estimated by Kaplan-Meier over the 10 folds for elderly people still at risk at landmark time $t_{LM} = 85$.



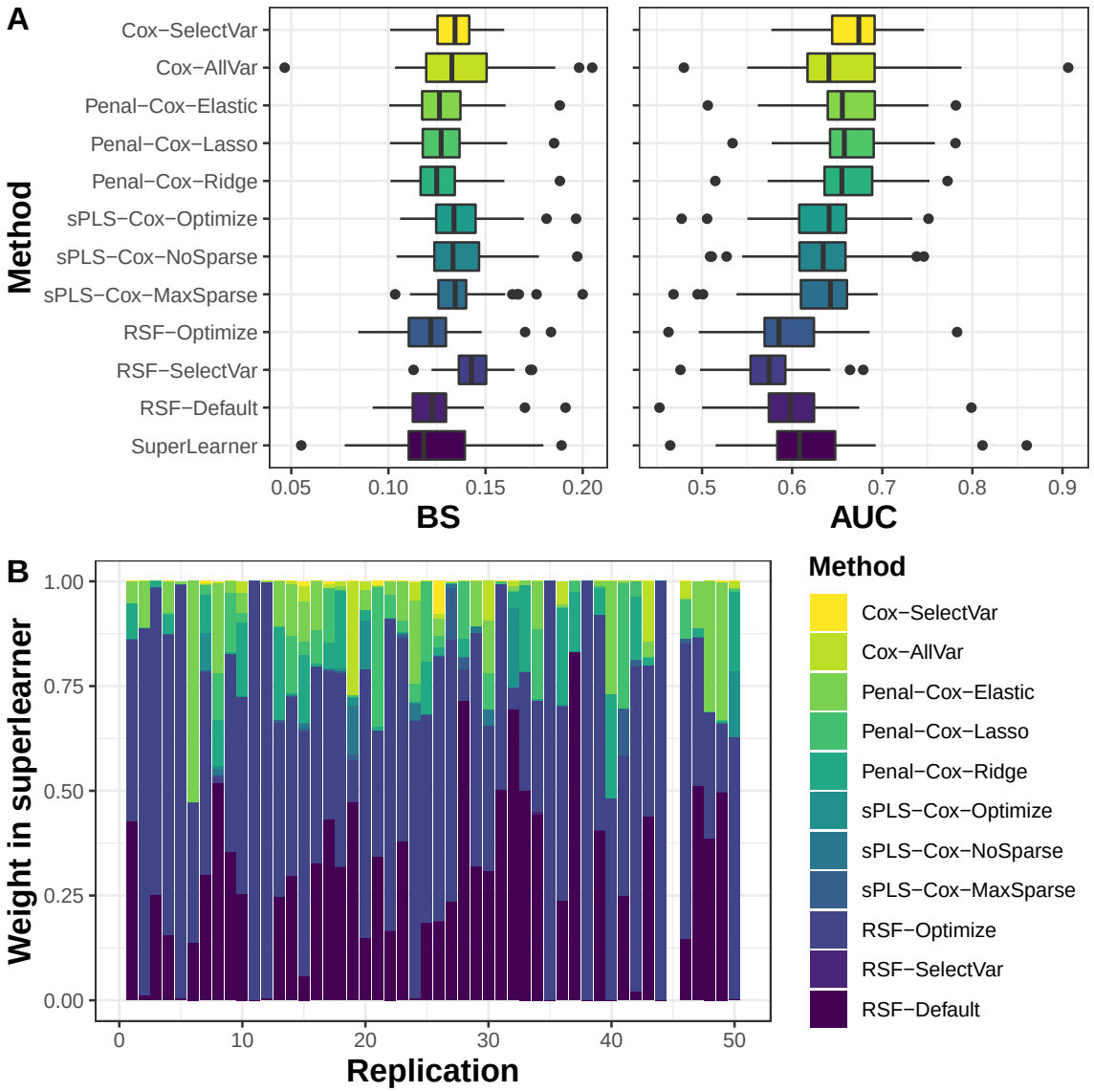

WEB FIGURE III.15 – Random survival forest hyperparameters tuning in the paquid application at landmark time $t_{LM} = 80$. The best hyperparameters ($mtry = 30$ and $nodesize = 15$) are chosen by minimizing the out-of-bag error.



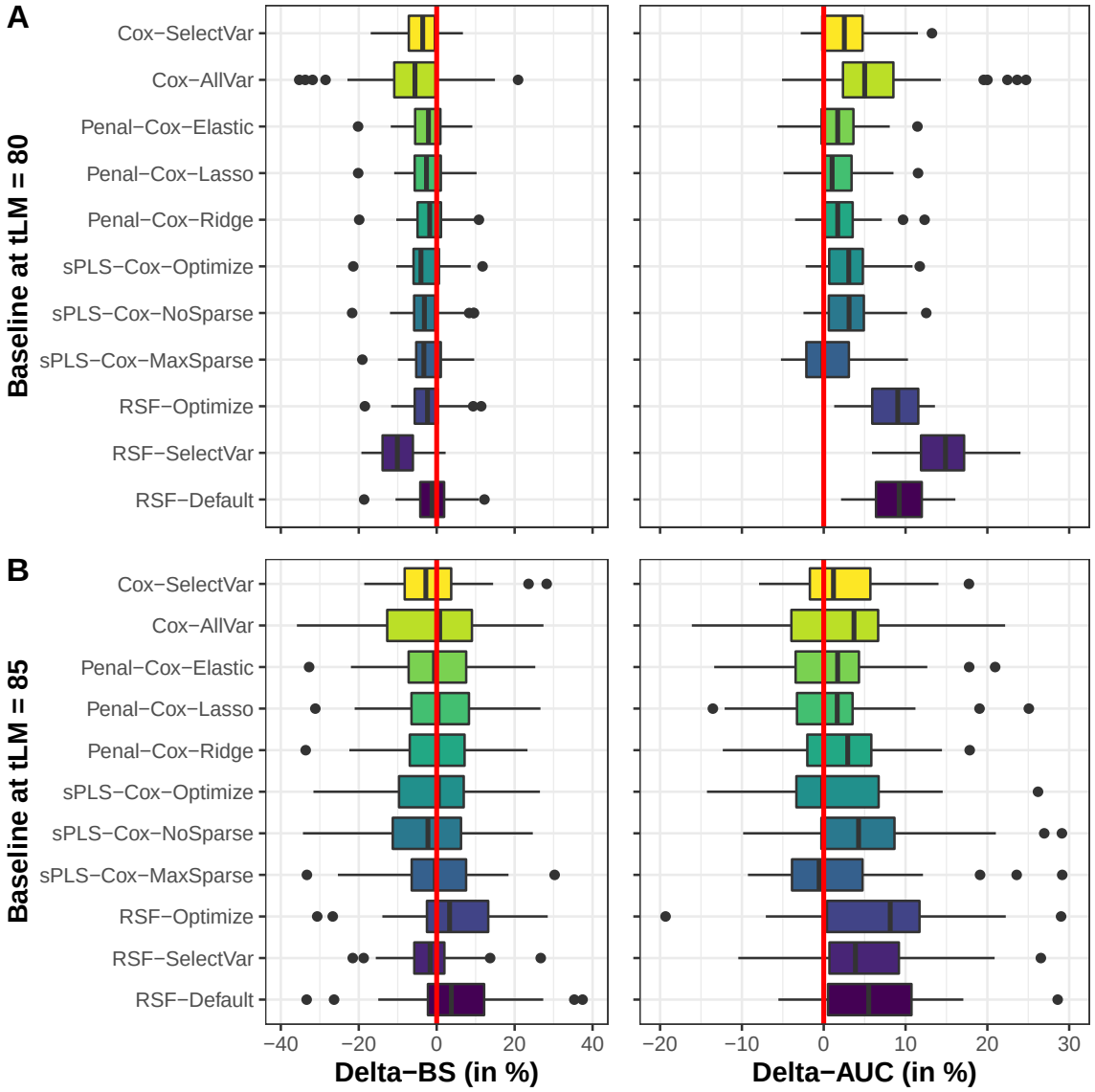

WEB FIGURE III.16 – Random survival forest hyperparameters tuning in the paquid application at landmark time $t_{LM} = 85$. The best hyperparameters ($mtry = 70$ and $nodesize = 5$) are chosen by minimizing the out-of-bag error.



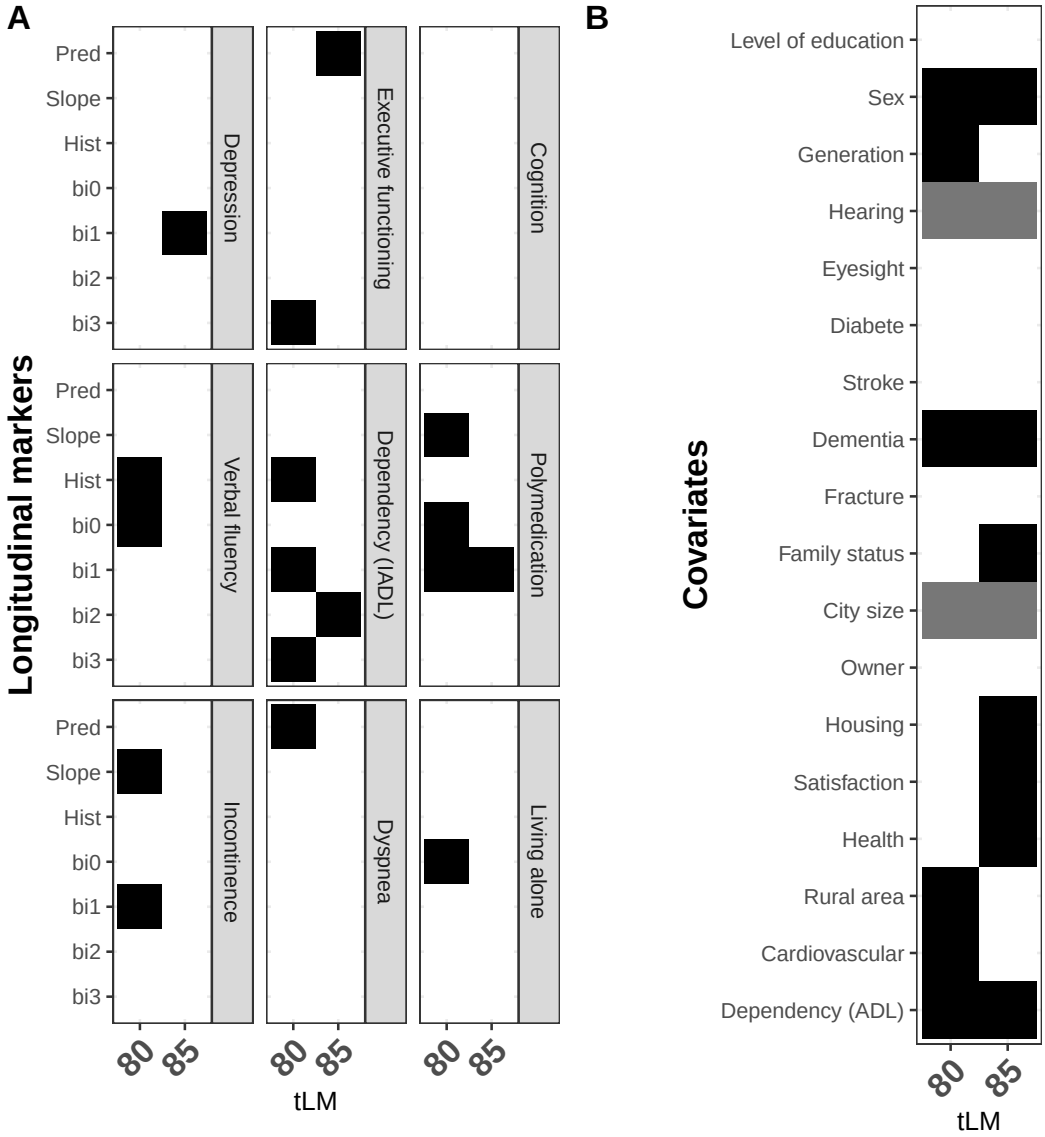

WEB FIGURE III.17 – Predictive performances (figure A) and weights in superlearner (figure B) of 5-year survival prediction tool that uses information collected from the last 5 years before landmark time $t_{LM} = 80$ over 50 replicates. Methods are assessed using Brier Score (BS) and Area Under the ROC Curve (AUC).




WEB FIGURE III.18 – Predictive performances (figure A) and weights in superlearner (figure B) of 5-year survival prediction tools that use information collected from the last 5 years before landmark time $t_{LM} = 85$ over 50 replicates. Methods are assessed using Brier Score (BS) and Area Under the ROC Curve (AUC).




WEB FIGURE III.19 – Predictive performances of 5-year survival prediction tools that use information collected at landmark time $t_{LM} = 80$ (figure A) and $t_{LM} = 85$ (figure B) over 50 replicates. Methods are assessed using Brier Score (BS) and Area Under the ROC Curve (AUC).




WEB FIGURE III.20 – Variables associated with the event for Cox model with Lasso penalty. The heatmaps show which summaries (figure A) and covariates (figure B) have been selected in the model. The black color indicates that the variable has been selected, while the color grey indicates that at least one modality of the variable has been selected, otherwise white for no selection.

### III.1.6.6 Tables

WEB TABLE III.1 – Type of summaries used in scenario 1 (in red), scenario 2 (in blue), scenario3 (in green), scenario 4 (in yellow) or scenario 5 (in orange).

| | $Y_i^{pred}$ | $Y_i^{slope}$ | $Y_i^{hist}$ | $b_{i0}$ | $b_{i1}$ | $b_{i2}$ |
|---|---|---|---|---|---|---|
| Marker 1 | ✔ ✔ ✔ | ✔ ✔ ✔ | ✔ ✔ | | | |
| Marker 2 | | ✔ | | | | |
| Marker 3 | ✔ | | ✔ ✔ ✔ ✔ | | | |
| Marker 4 | ✔ | ✔ ✔ ✔ | ✔ | | | |
| Marker 5 | ✔ | ✔ ✔ | ✔ | | | |
| Marker 6 | | | ✔ | | | |
| Marker 7 | | | | | | |
| Marker 8 | ✔ ✔ | | | | | |
| Marker 9 | | ✔ | ✔ | | | |
| Marker 10 | ✔ ✔ ✔ | | ✔ | | | |
| Marker 11 | | | ✔ | | | |
| Marker 12 | | ✔ | | | | |
| Marker 13 | ✔ ✔ | ✔ | | | | |
| Marker 14 | | ✔ | | | | |
| Marker 15 | ✔ ✔ ✔ ✔ | ✔ | ✔ | | | |
| Marker 16 | | | ✔ | | | |
| Marker 17 | ✔ ✔ | ✔ | | | | |

WEB TABLE III.2 – Summaries of predictors used to predict survival probability in primary biliary cholangitis patients.

| **Predictor** | **Description** | **Type** | **Time-dependent Yes/No** |
|---|---|---|---|
| Bilirubin | Level of serum bilirubin | Continuous | Yes |
| Cholesterol | Level of serum cholesterol | Continuous | Yes |
| Albumin | Level of albumin | Continuous | Yes |
| Alkaline | Level of alkaline phosphatase | Continuous | Yes |
| SGOT | Level of aspartate aminotransferase | Continuous | Yes |
| Platelets | Platelet count | Continuous | Yes |
| Prothrombin | Prothrombin time | Continuous | Yes |
| Ascites | Presence of ascites (Yes/No) | Binary | Yes |
| Hepatomegaly | Presence of hepatomegaly (Yes/No) | Binary | Yes |
| Spiders | Blood vessel malformations in the skin (Yes/No) | Binary | Yes |
| Edema | Presence of edema (Yes/No) | Binary | Yes |
| Age | Age at enrollment | Continuous | No |
| Sex | / | Binary | No |
| Treatment | Drug treatment (D-penicillmain/Placebo) | Binary | No |

Web Table III.3 – List of time-dependent variables used to predict the survival probability in the elderly

| **Predictor** | **Description** | **Type** | **Time-dependent Yes/No** |
|---|---|---|---|
| Depression | Measured using the Center for Epidemiological Studies-Depression (CES-D) providing a score from 0 to 60, with high scores indicating most depressive symptoms | Continuous | Yes |
| Executive functioning | Measured using the Wechsler code test | Continuous | Yes |
| Cognition | Measured using the Mini-Mental State Examination (MMSE) providing a score from 0 to 30, with lower score indicating suspicion of dementia | Continuous | Yes |
| Verbal fluency | Measured using Isaac set Test, which evaluates the verbal fluency in 15 secondes by reapeated a list of specific words | Continuous | Yes |
| Dependency (IADL) | Measured using Instrumental Activities of Daily Living (IADL), also called Lawton scale, with multiple questions about how well you can live on your own | Continuous | Yes |
| Polymedication | Daily number of drugs taken by the patient | Continuous | Yes |
| Incontinence | Yes/No | Binary | Yes |
| Dyspnea | Yes/No | Binary | Yes |
| Living alone | Yes/No | Binary | Yes |

WEB TABLE III.4 – List of time-dependent variables used to predict the survival probability in the elderly

| **Predictor** | **Description** | **Type** | **Time-dependent Yes/No** |
|---|---|---|---|
| Level of education | Scale from no education from higher education | 5-level factor | No |
| Sex | Male/Female | Binary | No |
| Generation | Age at enrollment | Continuous | No |
| Hearing | Hearing self-assessment (Good/Medium/Bad) | 3-level factor | No |
| Eyesight | Eyesight self-assessment (Good/Bad) | Binary | No |
| Diabete | Diabete history (Yes/No) | Binary | No |
| Stroke | Stroke history (Yes/No) | Binary | No |
| Dementia | Demantia history (Yes/No) | Binary | No |
| Fracture | Fracture history (Yes/No) | Binary | No |
| Family status | In couple/Single | Binary | No |
| City size | / | 4-level factor | No |
| Owner | Owner/Tenant | Binary | No |
| Housing | Personal residence/Other | Binary | No |
| Satisfaction | Scale measuring whether the patient is satisfied with his life | Continuous | No |
| Health | Health self-assessment (Good/Medium/Bad) | 3-level factor | No |
| Rural area | Yes/No | Binary | No |
| Cardiovascular | Cardiovascular history (Yes/No) | Binary | No |
| Dependency (ADL) | Measured using Activities of Daily Living (ADL), also called Katz scale, with multiple questions about the functional status of the patient | Binary | No |

# III.2 Prédiction de la démence avec la prise en compte du risque compétitif

Le travail méthodologique précédemment présenté a été étendu pour prendre en compte les risques compétitifs. Il a été appliqué pour prédire la démence en tenant compte du décès sans démence, où les analyses statistiques ont été réalisées par Ariane Bercu.

## III.2.1 Introduction

La démence est un syndrome causé par la dégradation du cerveau conduisant à une altération des fonctions cérébrales et une perte d'autonomie. Ce syndrome touche principalement les personnes âgées, réduisant drastiquement leur condition de vie au quotidien, et entraînant une augmentation de la mortalité chez ces personnes [Agüero-Torres et al., 1998]. La dégradation du cerveau survient plus de 10 ans avant l'apparition des premiers symptômes [Fuhrer et al., 2003], ce qui donne des pistes pour diagnostiquer la démence le plus tôt possible. Parmi les atteintes cérébrales, il existe l'atrophie de certaines régions du cerveau (hippocampe en particulier), mais également la maladie des petits vaisseaux du cerveau. Cette maladie se caractérise notamment par l'apparition de lésions vasculaires cérébrales pouvant être détectées par de l'imagerie à résonance magnétique (IRM). Plusieurs études ont mis en évidence que la maladie des petits vaisseaux du cerveau est un facteur de risque pour expliquer la survenue de la démence [Viswanathan et al., 2009, Mortamais et al., 2013].

Pour prédire au mieux la survenue de la démence, il est pertinent de prendre en compte les données issues de l'imagerie cérébrale en plus de facteurs de risque déjà identifiés comme l'âge, le sexe, le niveau d'éducation, la génétique, la dépendance, la dépression, le diabète, les maladies cardio-vasculaires ou le déclin cognitif [Tierney et al., 2005, Tierney et al., 2010, Lechevallier-Michel et al., 2004]. La prédiction de la démence, à partir de données d'IRM et divers facteurs de risque mesurés à l'inclusion, a déjà été réalisée [Stephan et al., 2015, Kuller et al., 2003]. L'objectif de ce travail est d'aller plus loin en

prenant en compte les données répétées d'IRM, de tests cognitifs, d'échelles mesurant la dépendance, de dépression, en plus de variables socio-démographiques pour prédire la survenue de la démence. Nous souhaitons en particulier identifier les facteurs associés de la démence et évaluer l'intérêt de considérer les données d'IRM en plus des données cliniques et psychométriques.

## III.2.2 Méthodologie

### III.2.2.1 La cohorte des trois-cités

Pour répondre à cet objectif, nous allons utiliser les données provenant de l'étude des trois-cités (3C). L'étude 3C est une cohorte prospective française en population générale dont l'objectif est d'étudier l'association entre les facteurs vasculaires et le risque de démence [3C Study Group, 2003]. Les individus éligibles au recrutement doivent : (i) habiter et être inscrits sur les listes électorales de Bordeaux, Dijon ou Montpellier ; (ii) être âgés de plus de 65 ans lors de l'inclusion entre mars 1999 et mars 2001 ; (iii) ne pas être institutionnalisés. La cohorte comporte 9294 individus.

Diverses mesures ont été collectées au cours du suivi après 2, 4, 7, 10, 12, 14 (uniquement Bordeaux et Montpellier) et 17 (uniquement Bordeaux) années, notamment des mesures précises de cognition, de dépression, de vie quotidienne et de santé. La fonction cognitive a été étudiée à l'aide de plusieurs tests dont : le test de mémoire visuelle de Benton (BENTON) [Benton, 1945], le *Mini-Mental State Examination* (MMSE) [Folstein et al., 1983] qui mesure le fonctionnement cognitif global, le *Trail Making Test part A and B* (TMT-A et TMT-B) [Reitan and Wolfson, 1985] qui mesure les fonctions exécutives, et le test d'Isaacs (IST) [Isaacs and Kennie, 1973] qui mesure la fluence verbale. La symptologie dépressive a été évaluée à partir du *Center for Epidemiologic Studies-Depression Scale* (CES-D) [Radloff, 1977]. Les échelles de Katz (ADL) [Katz, 1983] et Lawton (IADL) [Lawton and Brody, 1969] ont été utilisées pour évaluer la dépendance des individus dans leur quotidien. Diverses variables ont également été incluses : la consommation quotidienne de

médicaments, la pression artérielle systolique et diastolique, et l'indice de masse corporelle. Enfin, plusieurs variables ont été utilisées à l'inclusion comme le statut diabétique, la présence du gène APOE-4, le sexe et le niveau d'éducation de l'individu.

Par ailleurs, des IRM cérébrales ont été réalisées à l'inclusion puis 4 ans (Bordeaux, Dijon) et 10 ans (Bordeaux) après l'inclusion. A partir des IRM, plusieurs variables ont été extraites dont le volume moyen de l'hippocampe (HIPP), le volume intracrânien (TIV), les volumes de substance blanche (WMV) et grise (GMV), le volume d'hyper-intensité de la substance blanche dans la zone péri-ventriculaire (WMH Peri) et hors de la zone (WMH Deep) et la fraction parenchymal (BPF) calculée comme le ratio des volumes de substance blanche et grise sur le TIV. Le diagnostic de démence a été évalué lors de chaque visite de suivi. Dans le cas où l'individu est déclaré dément, le milieu de l'intervalle entre la dernière visite et la visite actuelle est défini comme temps de démence. L'indicateur de décès et la date de décès sont recueillis tout au long du suivi dans le cas où un individu est déclaré décédé (sans démence). Pour parer aux personnes décédées sans connaissance récente du statut de démence, seules les décès dans les 3 ans suivant un diagnostic de démence négatif ont été conservés. Les données des autres personnes décédées ont été censurées à la dernière visite renseignée.

#### III.2.2.2 Méthode par *landmark* avec risques compétitifs

A partir des données issues de la cohorte 3C, l'étude de la démence, avec la prise en compte du décès, nécessite une méthode spécifique. La méthode par *landmark* (décrite en section III.1) a été étendue en incorporant des méthodes de survie pour risques compétitifs. Cette extension ne concerne uniquement que la deuxième étape. La première étape visant à résumer les trajectoires des marqueurs reste inchangée.

L'objectif de ce travail est de fournir un modèle de prédiction de la démence, avec prise en compte du décès sans démence, sur un temps d'horizon de 5 ans à partir du temps *landmark* défini à 4 ans. Pour prendre en compte l'IRM dans l'analyse, seules les données de Bordeaux et Dijon ont été utilisées. Les variables TIV et IADL n'ont pas pu être

modélisés en tant que données répétées et ont été considérées respectivement à l'inclusion et au temps *landmark*. A partir de l'échelle ADL, l'item correspondant à l'incontinence a été évaluée de façon répétée, alors que les autres items de l'ADL ont été agrégés à l'aide d'un score calculé au temps *landmark*. Au total, 16 variables longitudinales ont été modélisées (voir figure III.21) et 11 variables ont été prises en compte à un seul temps (7 au temps *landmark* et 4 à l'inclusion).

La méthode nécessite au moins une mesure par sujet pour chacune des variables, conduisant à un total de 1720 sujets encore à risque à 4 ans. Dans la première étape de ce travail, les 16 variables longitudinales ont été indépendamment modélisées par des modèles mixtes de 0 à 4 ans, après une normalisation par *splines* [Proust-Lima et al., 2017]. Les valeurs et les pentes ont été prédites au temps *landmark* à l'aide des paramètres estimés des modèles mixtes. Dans la deuxième étape, les valeurs et pentes prédites au temps *landmark*, ainsi que les variables mesurées au temps *landmark* et à l'inclusion ont été utilisées en tant que prédicteurs dans ces modèles de survie, soit un total de 43 prédicteurs. Nous avons considéré plusieurs techniques de survie adaptées aux risques compétitifs d'évènements :

- les forêts aléatoires en survie : le critère pour partitionner les individus en sous-groupe peut être choisi pour cibler l'incidence cumulée de l'évènement d'intérêt. Nous avons utilisé le test de Fine & Gray [Gray, 1988] ;
- les régressions pénalisées : l'incidence cumulée de démence peut être modélisée, en prenant en compte le risque de décès sans démence, par un modèle de Fine & Gray [Fine and Gray, 1999] estimé par vraisemblance pénalisée. La pénalisation *Minimax Concave Penalty* (MCP) a été choisie pour faire l'inférence statistique car les paramètres estimés sont peu biaisés [Zhang, 2010, Fan and Li, 2001]. Cette pénalisation, notée $\mathcal{P}(\beta)$, est de la forme suivante :

$$\mathcal{P}(\beta) = \begin{cases} \lambda|\beta| - \frac{\beta^2}{2\gamma} \text{ si } |\beta| \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2 \text{ si } |\beta| > \gamma\lambda \end{cases} \tag{III.38}$$

où $\lambda$ est le paramètre contrôlant la force de la pénalisation. $\gamma$ est le paramètre pour

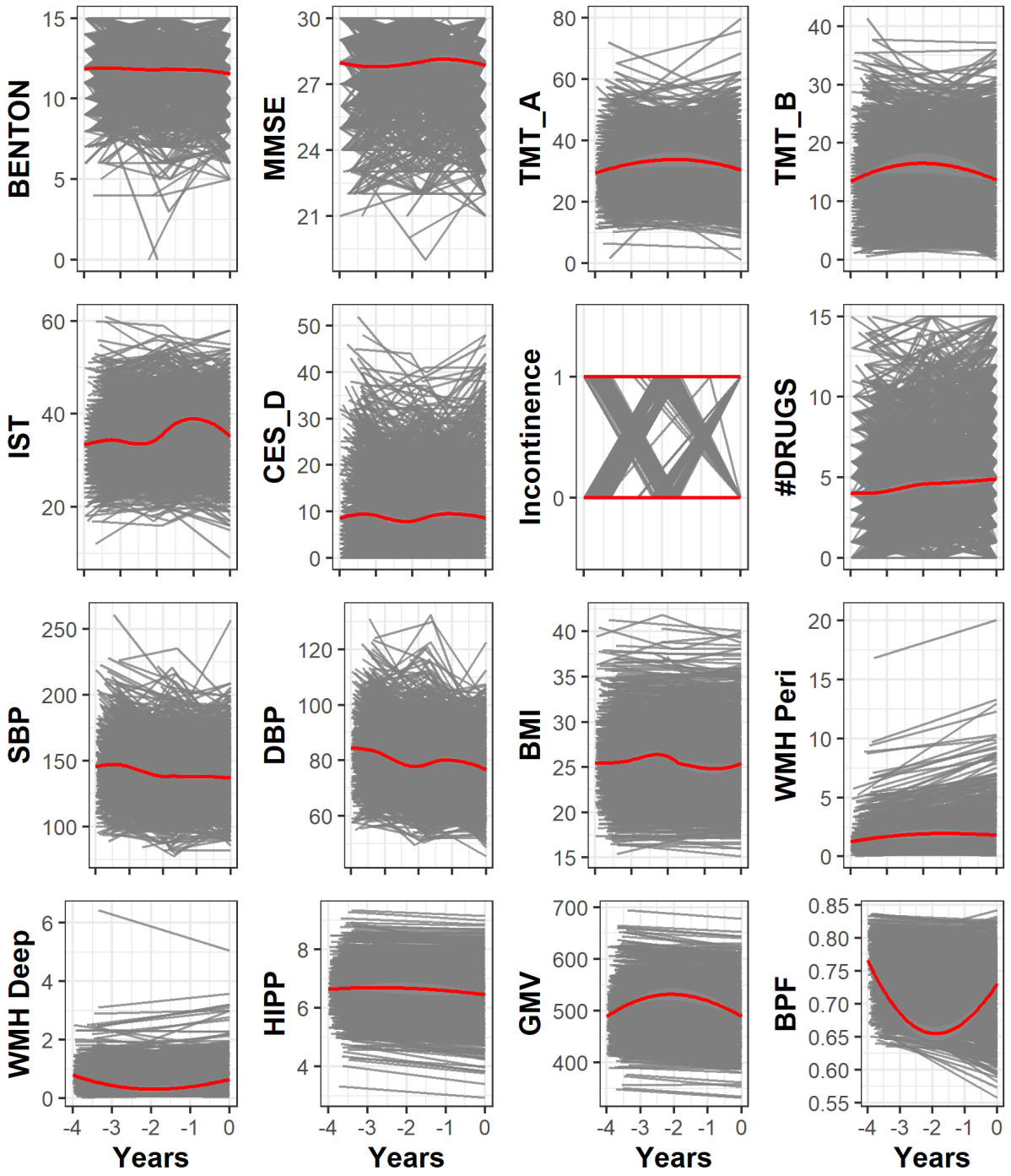



FIGURE III.21 – Trajectoires individuelles des variables longitudinales utilisées dans l'étude 3C.

définir la concavité de la pénalisation (i.e. la décroissance plus ou moins rapide de la pénalisation).

Suivant les recommandations de Zhang, $\gamma$ a été fixé à 2,7. En revanche, le paramètre

$\lambda$ a été optimisé en minimisant le critère de *generalized cross-validation* (GCV) [Wahba, 1977].

La qualité prédictive du modèle a été évaluée à l'aide du Brier Score (BS) et de l'*Area Under the ROC Curve* (AUC) après validation interne des prédictions par validation croisée 5 blocs, pour éviter un sur-apprentissage des données. La méthode a été répliquée 50 fois pour vérifier la variabilité des résultats.

### III.2.3 Résultats

Au total, 1720 personnes à risque de démence à 4 ans ont été incluses. Parmi elles, 111 ont développé une démence et 109 sont décédées sans démence dans les 5 ans.

Suite aux analyses préliminaires, la forêt aléatoire en survie pour risques compétitifs a été abandonnée et seule la régression pénalisée par MCP est décrite dans la suite. Les résultats montrent de bonnes performances prédictives en terme de discrimination et calibration pour le modèle incluant toutes les variables (voir figure III.22). En effet, l'AUC obtenue est de $0,822$ ($\text{IC}_{95\%} = [0,804; 0,840]$) et le BS est de $0,049$ ($\text{IC}_{95\%} = [0,047; 0,051]$). Ensuite, un modèle sans les données d'IRM a également été évalué pour quantifier la perte de qualité de prédiction par rapport au modèle complet. Les résultats montrent une AUC de $0,801$ ($\text{IC}_{95\%} = [0,778; 0,823]$) et un BS de $0,051$ ($\text{IC}_{95\%} = [0,050; 0,053]$). Les résultats indiquent que le modèle complet est légèrement meilleur en terme de discrimination et calibration par rapport au modèle complet sans les données IRM.

Sachant que la pénalisation MCP fournit des paramètres estimés peu biaisés, il est possible de faire des tests statistiques pour vérifier s'il existe une association entre les prédicteurs et l'incidence cumulée de démence. La figure III.23 indique que les valeurs prédites à 4 ans du volume de matière grise GMV, MMSE et HIPP et la pente prédite à 4 ans de HIPP sont les prédicteurs les plus associés à l'incidence cumulée de démence. L'utilisation de transformation *splines* complique la quantification exacte de l'augmentation ou de la diminution de l'incidence cumulée de démence en fonction d'un prédicteur. Cependant, le sens de l'association reste inchangé. Ainsi, l'augmentation de GMV (pré-

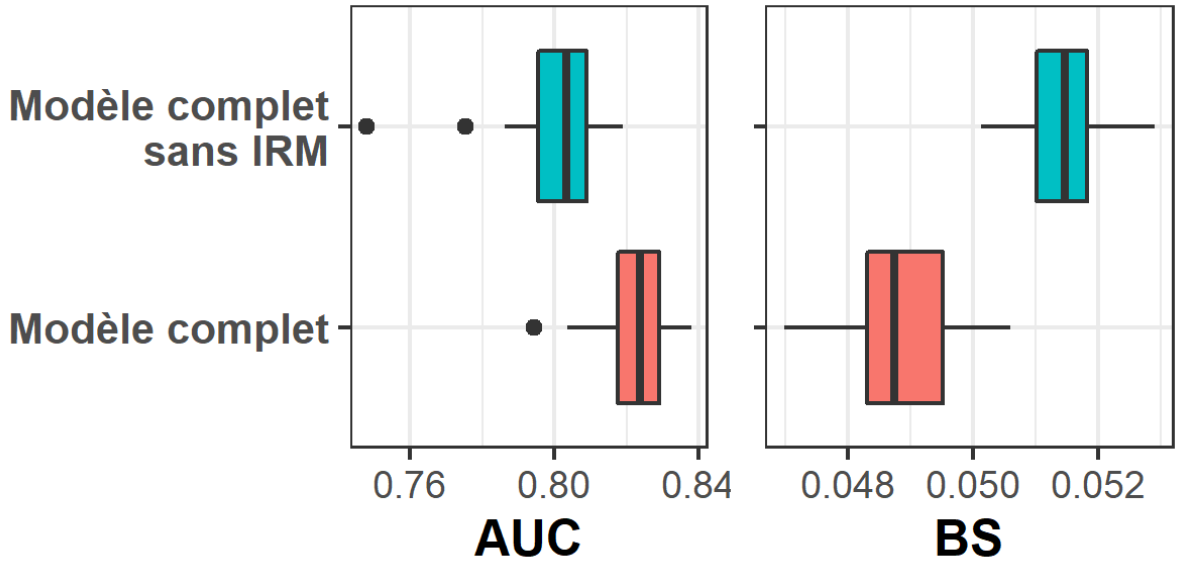



FIGURE III.22 – Évaluation des prédictions de la démence à 5 ans pour le modèle complet avec et sans IRM en utilisant l'*Area Under the ROC Curve* (AUC) et le Brier Score (BS).

dite à 4 ans) entraîne une augmentation de l'incidence. A l'inverse, l'augmentation du score MMSE, la valeur et la pente courante (prédites à 4 ans) du volume de l'hippocampe conduit à une baisse de l'incidence.

### III.2.4 Discussion

Dans ce travail, nous avons construit un modèle *landmark* pour prédire la démence à 5 ans, à partir de variables collectées jusqu'à 4 ans (IRM, tests cognitifs, échelle d'autonomie) et de variables socio-démographiques mesurées à l'inclusion sur une population de personnes âgées. Les résultats du modèle ont montré de bonnes performances prédictives à la fois en BS et AUC après validation croisée pour éviter un sur-apprentissage des données.

A l'aide de la pénalisation MCP, nous avons pu identifier les variables les plus associées à l'incidence de démence, et interpréter les différents paramètres associés. Nous avons montré que ces variables proviennent essentiellement des données IRM et des tests cognitifs. De plus, en modélisant les données longitudinales, nous avons pu mettre en évidence que la dynamique joue également un rôle essentiel, comme la pente prédite du

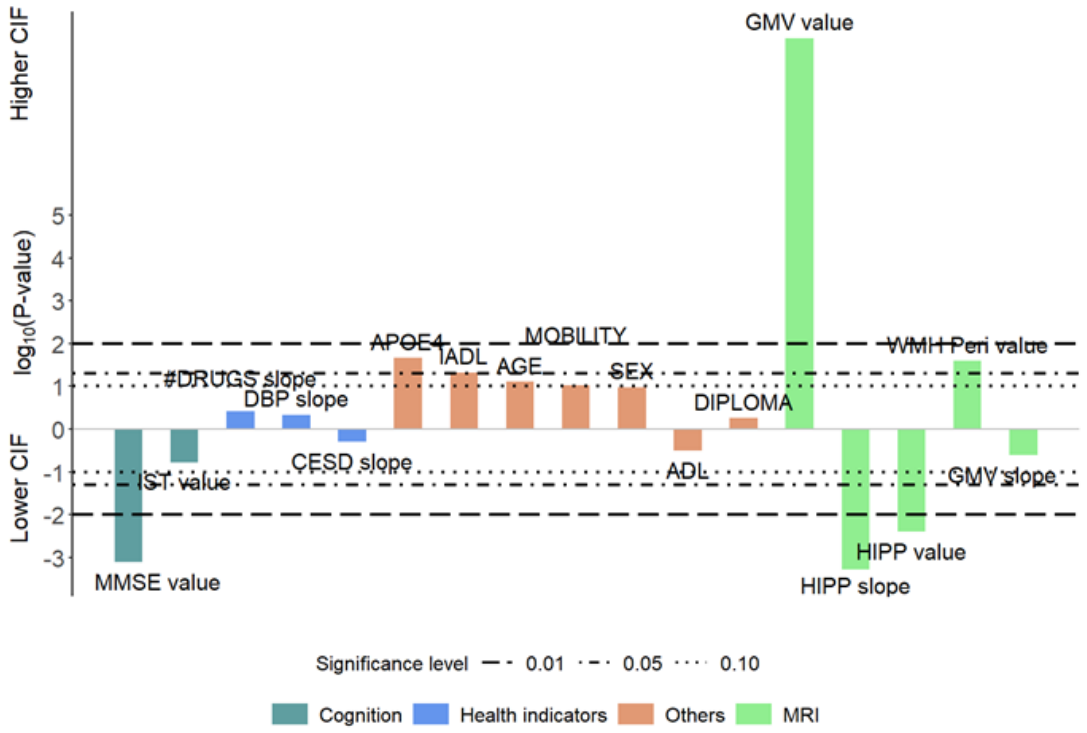



FIGURE III.23 – Association entre l'incidence cumulée (CIF) de démence et les variables sélectionnées par le modèle avec pénalisation MCP. Pour facilité la lecture du graphique, les p-valeurs sont représentées dans une échelle logarithme en base 10.

volume de l'hippocampe après 4 ans de suivi.

Dans cette approche, nous avons utilisé un modèle de Fine & Gray dans le but de construire un modèle prédictif. Néanmoins, l'interprétation des paramètres est plus complexe car le modèle estime l'incidence cumulée de démence. Dans un futur travail, il serait intéressant d'utiliser un modèle de Cox cause-spécifique pour estimer directement le risque de démence dans le but d'interpréter les paramètres de façon causale.

---

## CONCLUSION

Nous avons proposé dans ce chapitre une approche pour associer un grand nombre de marqueurs répétés avec des évènements de survie, possiblement en compétition avec un autre évènement. La méthode d'estimation a été implémentée dans le package R `hdlandmark`.

L'approche *landmark* est très efficace lorsqu'elle est utilisée avec un unique temps *landmark*. Cependant, cette approche est limitée quand plusieurs temps *landmark* sont pertinents. En effet, il est nécessaire

d'estimer un modèle pour chaque temps *landmark* rendant plus complexe l'utilisation de la méthode. Étendre cette approche pour permettre l'estimation jointe quel que soit le temps *landmark* est l'objectif du chapitre suivant. La perspective est de pouvoir évaluer les prédictions à différents temps *landmark* pour considérer plus ou moins de données longitudinales.

# Chapitre IV

# Forêt aléatoire en survie pour multiples données répétées

## Sommaire

INTRODUCTION

Ce chapitre est consacré aux forêts aléatoires en survie pour multiples données répétées, et est composé de trois sections. La première section introduit la méthodologie suivie d'une étude de simulation et d'une application pour prédire la démence parmi une population de personnes âgées. Ce travail a été soumis pour publication dans *Biostatistics*.

La deuxième section décrit précisément le package R `DynForest` développé pour l'utilisation de cette méthodologie et disponible sur le CRAN (*The Comprehensive R Archive Network*). Les différentes fonctionnalités de `DynForest` sont présentées à travers plusieurs exemples en fonction de la nature de la variable réponse à prédire. Cette section écrite comme une vignette est à soumettre au journal *Journal of Statistical Software*.

Enfin, la troisième partie est une application clinique de la méthode pour prédire la survenue du vasospasme cérébral auprès des patients atteints d'hémorragie sous-arachnoïdienne. Elle est le fruit d'une collaboration avec Hugues De Courson, praticien hospitalier en service de neuro-réanimation au CHU de Bordeaux.

# IV.1 Random survival forests for competing risks with multivariate longitudinal endogenous covariates

**Anthony Devaux**[1], **Catherine Helmer**[1], **Robin Genuer**[1,2] and **Cécile Proust-Lima**[1]

[1]INSERM, Bordeaux Population Health, U1219, Univ. Bordeaux, Bordeaux, France

[2]INRIA Bordeaux Sud-Ouest, Talence, France

*Submitted*

**Abstract :** Predicting the individual risk of a clinical event using the complete patient history is still a major challenge for personalized medicine. Among the methods developed to compute individual dynamic predictions, the joint models have the assets of using all

the available information while accounting for dropout. However, they are restricted to a very small number of longitudinal predictors. Our objective was to propose an innovative alternative solution to predict an event probability using a possibly large number of longitudinal predictors. We developed `DynForest`, an extension of random survival forests for competing risks that handles endogenous longitudinal predictors. At each node of the tree, the time-dependent predictors are translated into time-fixed features (using mixed models) to be used as candidates for splitting the subjects into two subgroups. The individual event probability is estimated in each tree by the Aalen-Johansen estimator of the leaf in which the subject is classified according to his/her history of predictors. The final individual prediction is given by the average of the tree-specific individual event probabilities. We carried out a simulation study to demonstrate the performances of `DynForest` both in a small dimensional context (in comparison with joint models) and in a large dimensional context (in comparison with a regression calibration method that ignores informative dropout). We also applied `DynForest` to (i) predict the individual probability of dementia in the elderly according to repeated measures of cognitive, functional, vascular and neuro-degeneration markers, and (ii) quantify the importance of each type of markers for the prediction of dementia. Implemented in the R package `DynForest`, our methodology provides a novel and appropriate solution for the prediction of events from any number of longitudinal endogenous predictors.

**Keywords :** Individual dynamic prediction ; Multivariate predictors ; Random survival forest ; Longitudinal data ; Survival data ; Competing risks.

### IV.1.1 Background

Quantifying the patient specific risk of disease or health events related to a disease progression based on patient's information has become a crucial issue in modern medicine. This may be done in order to monitor the disease progression, and adapt therapeutic strategies and medical choices according to their risk. One strategy is to predict the risk

of event using only the data collected at the prediction time. However, in many contexts, patients data include repeated measures of markers which trajectories are highly predictive of the event. This is the case for instance with prostate specific antigen for the risk of prostate cancer recurrence [Proust-Lima and Taylor, 2009] or serum creatinine for the risk of kidney graft failure [Fournier et al., 2019]. In other contexts, such as in cardiovascular disease, not only one specific marker but many potential markers may be relevant [Paige et al., 2018].

Longitudinal markers are endogenous variables in the sense that they may be affected by the event of interest [Rizopoulos, 2012], and are usually measured intermittently with a measurement error. This makes their statistical analysis challenging. Three approaches were proposed in the literature for the prediction of a clinical event given longitudinal endogenous information : *landmark approach* [Van Houwelingen, 2007], *joint models* [Proust-Lima and Taylor, 2009] and *regression calibration* techniques [Ye et al., 2008].

Landmark approach consists in considering only the subjects still at risk of the event at a prediction time $t$ (called landmark time) and including their information collected until $t$ to build a prediction tool for subsequent risk of event [Van Houwelingen, 2007, Ferrer et al., 2019]. The longitudinal information of intermittently measured and prone-to-error markers up to $t$ can be included after a pre-processing step by mixed models [Proust-Lima and Taylor, 2009, Paige et al., 2018, Devaux et al., 2022a, Tanner et al., 2021]. In multivariate settings, the Cox model may be replaced by more advanced techniques, coming from statistical learning and adapted to survival data [Devaux et al., 2022a, Tanner et al., 2021] to account for the possibly large dimension of the predictors and their correlation.

The landmark approach is relatively easy to implement and was shown to be robust to the misspecification of the marker trajectory or to the proportional assumption in the Cox model [Ferrer et al., 2019]. This makes it an appealing approach for extending the concept of individual dynamic prediction from a unique longitudinal marker to predictions from multivariate longitudinal markers. However, because it only relies on subjects at risk at the landmark time and exploits only the longitudinal information up to the landmark

time, it suffers from a lack of efficiency, and is restricted to pre-determined prediction times.

When longitudinal and survival processes are inter-related as assumed in the dynamic prediction context, the joint modelling framework constitutes the most appropriate approach to handle this mutual dependence [Rizopoulos, 2012]. Joint models (JM) simultaneously model the longitudinal and survival processes over time while accounting for their association using shared latent quantities ; and the posterior conditional individual probability of event given the longitudinal predictor history can be easily deduced. Initially developed for a single longitudinal predictor [Proust-Lima and Taylor, 2009], the method was then extended to a few longitudinal predictors [Hickey et al., 2018, Rizopoulos, 2016]. In contrast with landmark approaches, JM exploit all the available longitudinal information to build the prediction tool, thus leading to a better efficiency. However, their performances are very sensitive to the correct specification of the model [Ferrer et al., 2019]. Moreover, due to the complexity of their estimation, they are currently limited to a small number of longitudinal markers (usually 2 or 3) and thus cannot be used to predict individual risk of event in more complex settings [Hickey et al., 2016].

In the context of a large to high number of longitudinal predictors, regression calibration (RC) techniques were proposed as an alternative to JM. RC is a 2-stage approach which first summarizes the longitudinal predictors into time-fixed features as in the landmark approach, and then includes the features into prediction models. Several RC methods have been proposed with a first step using mixed models or functional data analysis [Yao et al., 2005] to summarize the multiple longitudinal predictors. Then, Cox model [Li and Luo, 2019], penalized regression [Signorelli et al., 2021] or random survival forests [Jiang et al., 2021, Lin et al., 2021] have been used to derive the risk prediction. Although RC techniques include all the available information on the markers and survival during the follow-up to build the prediction tool as in JM, they neglect the informative truncation of the longitudinal data due to the event, which may bias the estimates and impact the prediction accuracy [Albert and Shih, 2010].

In this work, we propose a novel methodology based on random survival forests (RSF) [Ishwaran et al., 2008] to accurately predict a risk of event from possibly large-dimensional longitudinal predictors. RSF have become popular for prediction tasks as they can handle a high number of covariates and capture potentially complex associations. However, RSF have been limited so far to time-independent predictors. To extend RSF to intermittently measured and error-prone longitudinal predictors, we model them using linear mixed model. However, in contrast to the landmark and RC techniques, we directly incorporate those computations at each recursive step of the RSF tree building to better handle the informative truncation of the longitudinal endogenous data and thus provide more appropriate and accurate individual predictions.

The rest of this article is organized as follows. Section 2 introduces our extended random survival forest methodology, called `DynForest`, and describes how it can handle time-dependent endogenous predictors to predict competing risks of event. Section 3 describes an extensive simulation study which aimed at validating `DynForest` methodology and at contrasting its performances with those of alternative approaches. In section 4, `DynForest` is applied in a large population-based French cohort to predict the probability of experiencing a dementia before death from multiple markers stemming from neuropsychological evaluation, clinical evaluation and brain Magnetic Resonance Imaging (MRI). Finally, section 5 closes the work with a discussion.

## IV.1.2 Methods

### IV.1.2.1 Framework and notations

We consider a sample of $N$ subjects. For each individual $i \in \{1, ..., N\}$, we denote $T_i$ the event time, $C_i$ the independent censoring time, and $\tilde{T}_i = \min(T_i, C_i)$ the observed time of event. We define $\delta_i$ the indicator of the cause of event with $\delta_i = k$ if subject experiences the event of cause $k \in \{1, ..., K\}$ before censoring and $\delta_i = 0$ otherwise. We observe an ensemble $\mathcal{M}_x$ of $P$ time-independent covariates $X_{ip}$ $(p = 1, ..., P)$, and an

ensemble $\mathcal{M}_y$ of $Q$ time-dependent covariates $Y_{ijm}$ for $m = 1, ...Q$ measured at subject-and-covariate-specific times $t_{ijm}$ with $j = 1...n_{im}$ the occasion and $t_{ijm} \leq \tilde{T}_i$.

Our methodology consists of a random survival forest for competing causes of event (RSF) that incorporates an internal processing for handling time-dependent covariates. A RSF is an ensemble of $B$ survival decision trees that are ultimately aggregated together. Each tree $b \in \{1, ..., B\}$ is built from a bootstrap sample of the original sample of $N$ subjects. This results, on average, in the exclusion of 37% of the subjects that constitute the out-of-bag (OOB) sample, noted $OOB^b$.

### IV.1.2.2 The tree building

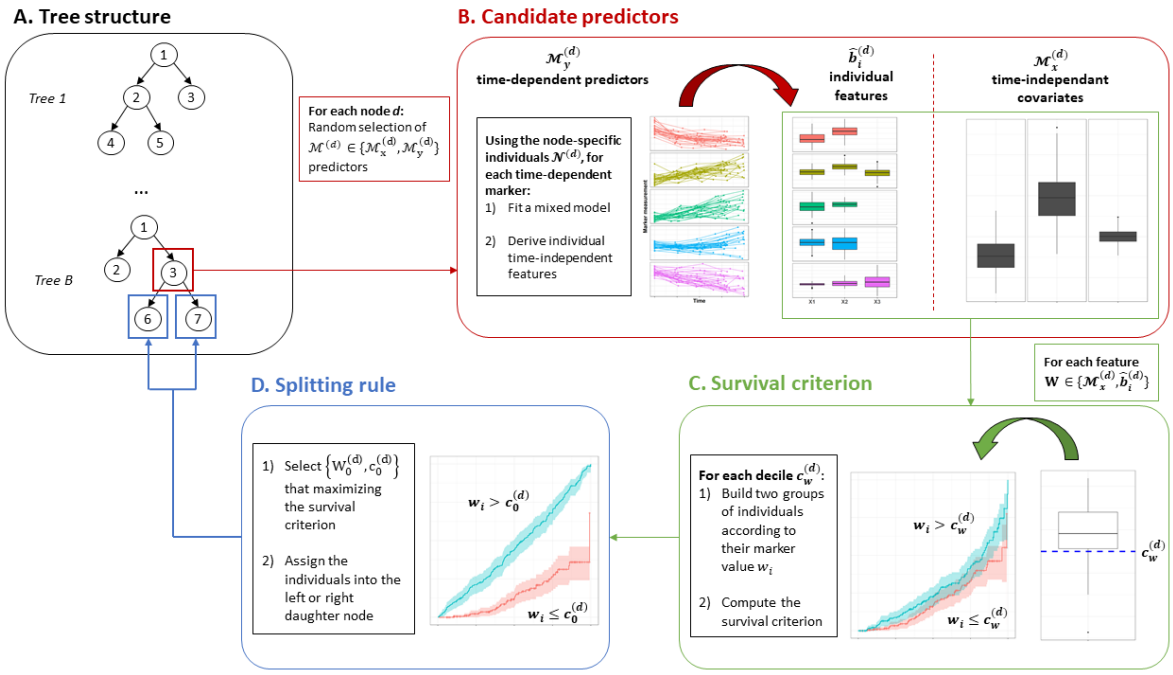



FIGURE IV.1 – Overall scheme of the tree building in `DynForest` with (A) the tree structure, (B) the node-specific treatment of time-dependent predictors to obtain time-fixed features, (C) the dichotomization of the time-fixed features, (D) the splitting rule.

A tree is a recursive procedure designed to partition the subjects into homogeneous groups regarding the outcome of interest. The overall tree building procedure is summarized in Figure IV.1. Each tree recursively splits the bootstrap sample into two subgroups at junctions called nodes until the subgroups reach a minimal size. At each node $d \in \mathcal{D}$,

the split is determined according to a dichotomized feature that maximizes the distance between the two groups ; the distance definition depends on the nature of the outcome (see IV.1.2.2 for the competing risk setting). To improve accuracy and minimize the correlation between the trees, randomness is incorporated at each node $d$ by considering only a random subset of candidate covariates $\mathcal{M}^{(d)} = \{\mathcal{M}_x^{(d)}, \mathcal{M}_y^{(d)}\} \subset \{\mathcal{M}_x, \mathcal{M}_y\}$ which size is a tuning parameter, called *mtry*.

**Internal processing for time-dependent covariates** For all the time-dependent covariates, a node-specific pre-processing is achieved to summarize the covariate dynamics into a set of time-independent features to be included in the pool of candidates for the splitting (see Figure IV.1B). At each node $d$, the trajectory of time-dependent covariate $Y_m \in \mathcal{M}_y^{(d)}$ is modeled using a flexible mixed model [Laird and Ware, 1982] as :

$$Y_{ijm} = X_{im}^{\top}(t_{ijm})\beta_l^{(d)} + Z_{im}^{\top}(t_{ijm})b_{im}^{(d)} + \epsilon_{ijm}^{(d)} \tag{IV.1}$$

where $Y_{ijm}$ is the covariate value for subject $i$ at time $t_{ijm}$, $X_{im}^{\top}(t_{ijm})$ and $Z_{il}^{\top}(t_{ijm})$ are the $p_m$- and $q_m$-vectors associated with the fixed effects $\beta_m^{(d)}$ and random effects $b_{im}^{(d)}$ (with $b_{im}^{(d)} \sim \mathcal{N}(0, B_m^{(d)})$), respectively. $\epsilon_{ijm}^{(d)}$ denotes the error measurement with $\epsilon_{ijm}^{(d)} \sim \mathcal{N}(0, \sigma_m^{2\,(d)})$.

We present the method for continuous Gaussian time-dependent covariates only. However, the pre-processing procedure could be easily adapted to other types of time-dependent covariates by replacing the linear mixed model in (IV.1) by a generalized linear mixed model.

Any specification can be considered for $X_{im}^{\top}(t_{ijm})$ and $Z_{il}^{\top}(t_{ijm})$. To allow for a flexible modeling of the trajectory over time, we consider a basis of natural cubic splines with knots to be determined in input. Although the specification of the model is similar for each node, the maximum likelihood estimation of the parameters is performed at each node on the subset of subjects present at the node (i.e., $\forall i \in \mathcal{N}^{(d)}$). When the covariate had already been selected at a parent node, estimated parameters from the closest parent

node are considered as initial values to drastically speed-up the procedure.

Time-independent features are then derived as the predicted individual deviations to the mean trajectory :

$$\hat{b}_{im}^{(d)} = \hat{B}_m^{(d)} Z_{im}^\top \hat{V}_{im}^{-1(d)} (Y_{im} - X_{im}\hat{\beta}_m^{(d)}) \tag{IV.2}$$

where $X_{im}$ and $Z_{im}$ are the matrices with $j$-row vectors $X_{im}^\top(t_{ijm})$ and $Z_{il}^\top(t_{ijm})$ (with $j = 1, ...n_{im}$), $\hat{V}_{im}^{(d)} = Z_{im}\hat{B}_m^{(d)} Z_{im}^\top + \hat{\sigma}_{\epsilon m}^{(d)} I_{n_i}$, $I_{n_i}$ the $n_i \times n_i$ identity matrix and the hat denotes the Maximum Likelihood Estimates.

At this stage, the ensemble of candidate features for the time-dependent covariates becomes $\mathcal{M}_{y\star}^{(d)} = \{\hat{b}_{im}^{(d)} \quad \forall m : Y_m \in \mathcal{M}_y^{(d)}\}$ and the total ensemble of candidate features $\mathcal{M}_\star^{(d)} = \{\mathcal{M}_x^{(d)}, \mathcal{M}_{y\star}^{(d)}\}$ is now only composed of time-independent features.

**Splitting rule** At each node $d \in \mathcal{D}$, the subjects are to be split into the two daughter nodes that are the most different possible according to the outcome (Figure IV.1D). With survival outcome, the difference was previously quantified according to the log-rank statistic. In the presence of competing risks, we propose to use instead the Fine & Gray test statistic [Gray, 1988] which directly quantifies the difference in terms of the probability of the cause of event of interest.

The splitting procedure requires that each feature $W \in \mathcal{M}_\star^{(d)}$ be dichotomized. For a continuous predictor, this is achieved by considering a dichotomization according to a threshold $c : w_i > c$ or $w_i \leq c$. We used each decile of $W$ as a candidate threshold $c$. Alternatively $c$ could be chosen according to values randomly drawn from $W$. For a non-continuous predictor, the dichotomization can be achieved as $w_i \in c$ or $w_i \notin c$ with $c$ each possible subset of $W$ modalities.

The Fine & Gray test statistic is computed for all potential dichotomized features (defined by couple $\{W, c\}$), and the dichotomized feature ($\{W_0^d, c_0^d\}$) that maximizes the test statistic is selected to create the left and right daughter nodes, denoted nodes $2d$ and $2d + 1$, respectively.

**Stopping criteria** Criteria need to be established to end the recursive procedure of a tree construction. We distinguish two criteria to pursue with the splitting of a node : (i) A minimum number of events called *minsplit* ; (ii) A minimum number of subjects in each of the daughter nodes called *nodesize*. These two parameters control the depth of the trees. They need to be carefully determined as a trade-off between the performances of the random forest (with the deeper the trees, the lower the error of prediction) and the computational time. In our examples, we often used $nodesize = 3$ and $minsplit = 5$. When a stopping criterion is reached, the node is considered as a terminal node or leaf $h \in \mathcal{H}$.

**Leaf summary** The subjects classified in the same leaf are supposed to be homogeneous in terms of their probability of event of interest. Each leaf $h^b$ of tree $b$ is thus summarized by the cumulative incidence function (CIF) for cause $k$ ($k = 1, ..., K$) :

$$\pi_k^{h^b}(t) = P(T_i < t, \delta_i = k \mid i \in h^b) \ , \ \forall t \in \mathbb{R}^+ \tag{IV.3}$$

An estimate $\hat{\pi}_k^{h^b}(t)$ of the CIF $\pi_k^{h^b}(t)$ is given by the Aalen-Johansen estimator.

### IV.1.2.3 Individual prediction of the outcome

**Out-of-bag individual prediction** Let us consider an individual $\star$ with the $P$-vector of time-independent covariates $X_\star$ and the ensemble of time-dependent covariate observations $\mathcal{Y}_\star = \{Y_{\star jm}, m = 1, ...Q, j = 1...n_{\star m}\}$. The individual-specific CIF for individual $\star$ in tree $b$ is given by :

$$\begin{aligned} \pi_{\star k}^b(t) &= P(T_\star < t, \delta_\star = k \mid \mathcal{Y}_\star, X_\star, b) \\ &= P(T_\star < t, \delta_\star = k \mid \star \in h_\star^b) \\ &= \pi_k^{h_\star^b}(t) \end{aligned} \tag{IV.4}$$

where $h_\star^b$ is the leaf in which individual $\star$ ends when dropping into tree $b$. Specifically, at each node $d$, subject $\star$ is recursively assigned to the left or right node according to

whether $w_\star > c_0^d$ or $w_\star \leq c_0^d$. In the case where $W_0^d$ is a predicted random-effect from time-dependent covariate $m$, the random-effect prediction for individual $\star$, $\hat{b}_{\star m}^{(d)}$, is computed using formula (IV.2) with the estimated parameters obtained at this specific node $d$.

An ensemble estimate of the individual CIF $\hat{\pi}_{\star k}(t)$ for cause $k$ can finally be defined by aggregating the tree-specific individual predictions $\hat{\pi}_{\star k}^b(t) = \hat{\pi}_k^{h_\star^b}(t)$ over all the trees $\mathcal{O}_\star \subset \{1, ..., B\}$ for which $\star$ is $OOB$, as :

$$\hat{\pi}_{\star k}(t) = \frac{1}{|\mathcal{O}_\star|} \sum_{b \in \mathcal{O}_\star} \hat{\pi}_k^{h_\star^b}(t) \tag{IV.5}$$

where $|\mathcal{O}_\star|$ denotes the length of $\mathcal{O}_\star$ and $\hat{\pi}_k^{h_\star^b}(t)$ is the Aalen-Johansen estimator in leaf $h_\star^b$ of the $b$-th tree.

**Individual dynamic prediction from a landmark time** The methodology described in subsection IV.1.2.3 for an out-of-bag individual can be used to provide the individual dynamic prediction of the outcome of cause $k$ from the information collected up to a landmark time $s$. Let consider a new subject $\star$ still at risk of the event at time $s$. The covariate information available at the time of prediction is the $P$-vector of time independent covariates $X_\star$ and the history of time-dependent covariates observations up to time $s$ $\mathcal{Y}_\star(s) = \{Y_{\star jm}, m = 1, ...Q, j = 1...n_{\star m}, t_{\star jm} < s\}$. The probability of experiencing cause $k$ of event at a horizon time $w$ is then defined as :

$$\begin{aligned} \pi_{\star k}(s, w) &= P(s < T_\star \leq s + w, \delta_\star = k \big| T_\star > s, \mathcal{Y}_\star(s), X_\star) \\ &= \frac{\pi_{\star k}(s + w) - \pi_{\star k}(s)}{1 - \sum_{l=1}^K \pi_{\star l}(s)} \end{aligned} \tag{IV.6}$$

where each $\pi_{\star k}(t)$ (for $k = 1, ..., K$) can be estimated using equation (IV.5) with the history of the time-dependent covariates $\mathcal{Y}_\star(s)$ up to the landmark time $s$ only, and $\mathcal{O}_\star = \{1, ..., B\}$.

#### IV.1.2.4 Error of prediction

The error of prediction can be used in RSF with two objectives :(i) tuning the hyper-parameters of the RSF (*mtry*, *minsplit* and *nodesize*) to achieve an optimal RSF. This is done by minimizing the OOB error of prediction ; (ii) assessing the predictive performances of the optimal RSF. This is achieved by computing the error of prediction for an external validation sample, that is a sample where subjects are OOB for all the trees. In this work, we considered mainly the Brier Score measure [Blanche et al., 2015], and its integrated version (IBS) between two time points $\tau_1$ and $\tau_2$ to assess the error of prediction.

**IBS for optimizing the RSF** For the optimization of the RSF, the IBS estimator is given by $I\hat{B}S(\tau_1;\tau_2) = \int_{\tau_1}^{\tau_2} \hat{BS}(t)\,dt$ with the Brier Score estimated by :

$$\hat{BS}(t) = \frac{1}{N}\sum_{i=1}^{N} \hat{\omega}_i(t)\Big\{I(\tilde{T}_i \leq t, \delta_i = k) - \hat{\pi}_{ik}(t)\Big)\Big\}^2 \tag{IV.7}$$

where $\hat{\pi}_{ik}(t)$ is the estimated probability of event of cause $k$ given $Y_i$ and $X_i$ defined in (IV.5), and $\hat{\omega}_i(t)$ are Inverse Probability of Censoring Weights (IPCW) that account for the censoring between $\tau_1$ and $\tau_2$ [Gerds and Schumacher, 2006]. We used in this work the Kaplan-Meier estimator to compute the probability of censoring in IPCW.

By default, $(\tau_1, \tau_2)$ corresponds to the span of the time to event data.

**External assessment of RSF predictive performances** For the external evaluation of the RSF performances, the IBS computation slightly differs. First, it is now computed on an external sample of size $N^*$, and the information considered is now the information up to the prediction time $s$, with $s \leq \tau_1$, so that $I\hat{B}S^s(\tau_1;\tau_2) = \int_{\tau_1}^{\tau_2} \hat{BS}^s(t)\,dt$ with :

$$\hat{BS}^s(t) = \frac{1}{N^*}\sum_{\star=1}^{N^*} \hat{\omega}^s_\star(t)\Big\{I(\tilde{T}_\star \leq t, \delta_i = k) - \hat{\pi}_{\star k}(s, t-s)\Big)\Big\}^2 \tag{IV.8}$$

where $\hat{\pi}_{\star k}(s, t-s)$ is the estimated probability of event of cause $k$ between $s$ and $t$ given

the information on $Y_\star$ and $X_\star$ up to $s$ (see definition in (IV.6)), and $\hat{\omega}^s_\star(t) = \hat{\omega}_\star(t)I(\tilde{T}_\star > s)$.

This external validation step can be incorporated into a k-fold cross-validation strategy. This is what was done in the application in the absence of an actual external dataset, and repeated 50 times to account for the k-fold cross-validation variability.

#### IV.1.2.5 Importance of the predictors

Beyond the overall predictive performance of the approach, one can be interested in identifying which predictors are the most predictive. We propose to evaluate the association between event and predictors through two measures : the variable importance and the minimal depth.

**Variable importance** The variable importance (VIMP) measures the variable prediction ability by computing the increase in OOB error obtained after breaking the link between a given variable and the event. Such a link is broken by permuting the values of variable $p$ at the individual level when $p$ is time-fixed and at the observation level when $p$ is time-dependent. Then, the VIMP statistic for covariate $p$, called $VIMP^{(p)}$, is the difference between the mean over the trees of OOB errors obtained after permuting the values of covariate $p$ ($I\hat{B}S_b(\tau_1,\tau_2)^{(p)}$ for $b = 1, ..., B$) and the mean over the trees of the OOB errors ($I\hat{B}S_b(\tau_1,\tau_2)$ for $b = 1, ..., B$) :

$$VIMP^{(p)}(\tau_1,\tau_2) = \frac{1}{B}\sum_{b=1}^{B} I\hat{B}S_b(\tau_1,\tau_2)^{(p)} - \frac{1}{B}\sum_{b=1}^{B} I\hat{B}S_b(\tau_1,\tau_2) \tag{IV.9}$$

where $I\hat{B}S_b(\tau_1,\tau_2)^{(p)}$ and $I\hat{B}S_b(\tau_1,\tau_2)$ are defined similarly as the IBS by computing the Brier Score (in equation (IV.7)) only on $b$-tree OOB subjects and using the estimate of $b$-tree individual prediction $\hat{\pi}^{h^b_\star}_k(t)$ defined under equation (IV.5).

Large VIMP value indicates a loss of predictive ability when removing covariate $p$ whereas null VIMP value indicates no predictive ability. Due to the permutation procedure, negative VIMP may be obtained. They are interpreted as null VIMP.

**Grouped variable importance** Because of the potential correlation between variables, the VIMP computed at the variable level may not always indicate the correct variable-specific predictive ability. To assess the predictive ability of correlated variables, Gregorutti *et al.* [Gregorutti et al., 2015] proposed the grouped variable importance (gVIMP) statistic in standard random forest. It consists in simultaneously noising-up all the variables of a given group. We considered the same methodology for our RSF. The overall gVIMP statistic for group $g \in \{1, ..., G\}$ is defined as $gVIMP^{(g)} = \frac{1}{B}\sum_{b=1}^{B} I\hat{B}S_b^{(g)} - \frac{1}{B}\sum_{b=1}^{B} I\hat{B}S_b$ where $I\hat{B}S_g^{(g)}$ denotes the OOB error obtained when noising-up all the variables from group $g$.

**Average minimal depth** The minimal depth of a variable in a tree corresponds to the distance between the root node and the first node that used the variable for splitting the data. The minimal depth can be averaged across all the trees allowing to rank the predictors. Indeed, during the tree building, the most predictive variables are expected to be chosen for the first splits so the closer the average minimal depth from 1, the better the predictive ability of the variable. When only a random subset of variables are considered at each node ($mtry < P+Q$), the interpretation of the minimal depth may be blurred. We thus recommend to compute this statistic only for the maximal $mtry = P+Q$. Moreover, because the depth of the trees may vary and some predictors may not be systematically used in the tree building process, we recommend to report the number of trees where the predictor was selected along with the average minimal depth. Note that, compared to the VIMP, the minimal depth can be computed both at the summary feature and at the covariate level allowing to fully understand the tree building process.

### IV.1.3 Simulation study

We carried out a simulation study to illustrate the behaviour of `DynForest` in comparison with alternative methods under two scenarios :

- repeated data of two longitudinal predictors. We compared the performances of

`DynForest` with a JM estimated using `JMBayes` R package [Rizopoulos, 2016]. For `JMBayes`, we considered the same specification for the linear mixed models as in `DynForest` and modelled the association with the event with a proportional hazard model (baseline risk function approximated by 4 cubic splines) that included the current levels and current slopes of the two predictors as covariates.

- repeated data of 20 longitudinal predictors. We could not compare with a JM anymore. Instead, we compared `DynForest` with a RC technique in which the exact same specification for the linear mixed models and the exact same strategy for the RSF were considered. The difference in the RC was that the linear mixed models were estimated once and for all prior to the application of standard RSF.

For both scenarios, we additionally included two time-fixed predictors unrelated to the event. Finally, we compared the predictive performance of the techniques in predicting the clinical event occurrence at two horizon times $w = 1, 2$ from two landmark times $s = 2, 4$. We measured the performance with both Brier Score (defined in section IV.1.2.4) and Area Under the ROC curve (AUC) with estimators adapted to dynamic prediction [Blanche et al., 2015].

#### IV.1.3.1 Design

For both scenarios, $R = 250$ samples of $N = 500$ individuals were built for the learning step and a single external validation sample of $N = 500$ individuals was generated for evaluating the predictive performance. The generation procedure is detailed in supplementary material (simulation section and Web Tables IV.1/IV.2) and summarized below. Individual trajectories are also displayed in supplementary material (Web Figure IV.7).

For each subject, we generated two time-fixed covariates (one continuous according to a standard Gaussian distribution and one binary according to a Bernoulli with probability 0.5). We also generated repeated data of 2 or 20 continuous time-dependent predictors, for small and large dimension scenarios, respectively. Times of measurement were at baseline and then randomly drawn (using an exponential departure) around theoretical annual

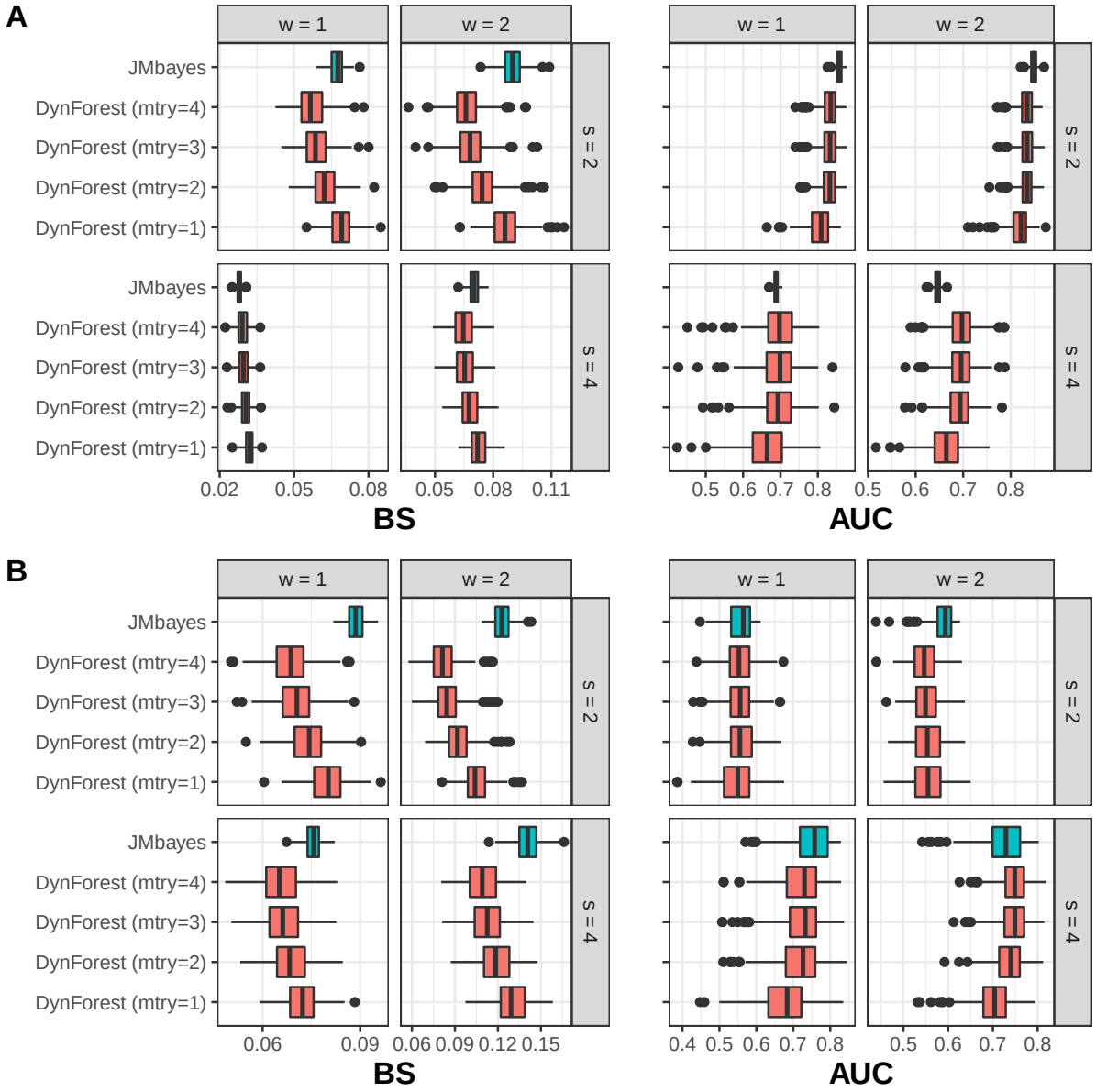



FIGURE IV.2 – External predictive performances of `DynForest` and `JMbayes` in the small dimension scenario of simulations (2 predictors) for the 250 replications. Are reported the Brier Score (BS) and the Area Under the ROC Curve (AUC) at two landmark times $s = 2, 4$ and two horizons $w = 1, 2$. The generated joint model included non-linear association between the markers and the event, using random-effects with two-by-two interaction (A) or latent class membership (B). In these scenarios, `JMbayes` is considered as misspecified. For `DynForest`, we fixed $nodesize = 3$ and $minsplit = 5$, and their results are reported for all mtry values to underline the importance of this tuning parameter.

visits up to 10 years. Each marker trajectory followed a latent class linear mixed model [Proust-Lima et al., 2014] with 4 classes and either class-specific linear individual trajectories or class-specific nonlinear individual trajectories. The risk of event was then generated

using a proportional hazard model with a Weibull baseline hazard with shape and scale parameters equal to 0.1 and 2, respectively. For both scenarios, we considered two sub-scenarios according to the form of the dependence function between the predictors and the risk of event using in the linear predictor of the survival model : (i) random-effects and two-by-two interactions between random-effects ; (ii) latent class membership directly.

#### IV.1.3.2 Results

**Small dimension scenario** Predictive performances on the external dataset are reported in terms of BS and AUC in Figure IV.2. For `DynForest`, we fixed $nodesize = 3$ and $minsplit = 5$ whereas we chose to report the results with each possible value of $mtry$ to underline the importance of this tuning parameter. As expected, the results varied substantially according to its value. The best performances in terms of BS (minimal BS) was systematically obtained with the largest $mtry$, that is 4 (2 time-dependent and 2 time-fixed predictors), and the worse with $mtry = 1$. For the AUC, the differences were less visible. In comparison with the JM estimated with `JMbayes`, `DynForest` showed overall better predictive abilities, in particular in terms of Brier Score. Again, conclusions were more tempered when considering AUC. It should be noted that JM in this simulation was misspecified in terms of both the individual deviation from the mean trajectory (since simulated using latent classes), and of dependence association since it has been generated as a non-linear function. The objective here was to illustrate that even in small dimension scenarios, `DynForest` could already be of interest and constituted a competing alternative for individual dynamic prediction purpose.

**Large dimension scenario** In the large dimension scenario, we report in Figure IV.3 the predictive performances on the external dataset of `DynForest` and its RC counterpart (or 2-stage counterpart). For both techniques, we fixed $nodesize = 3$ and $minsplit = 5$ whereas $mtry$ parameter was tuned. Regarding the range of possible values of $mtry$ (from 1 to 22 for `DynForest` and from 1 to 62 for RC method), tuning this parameter on each

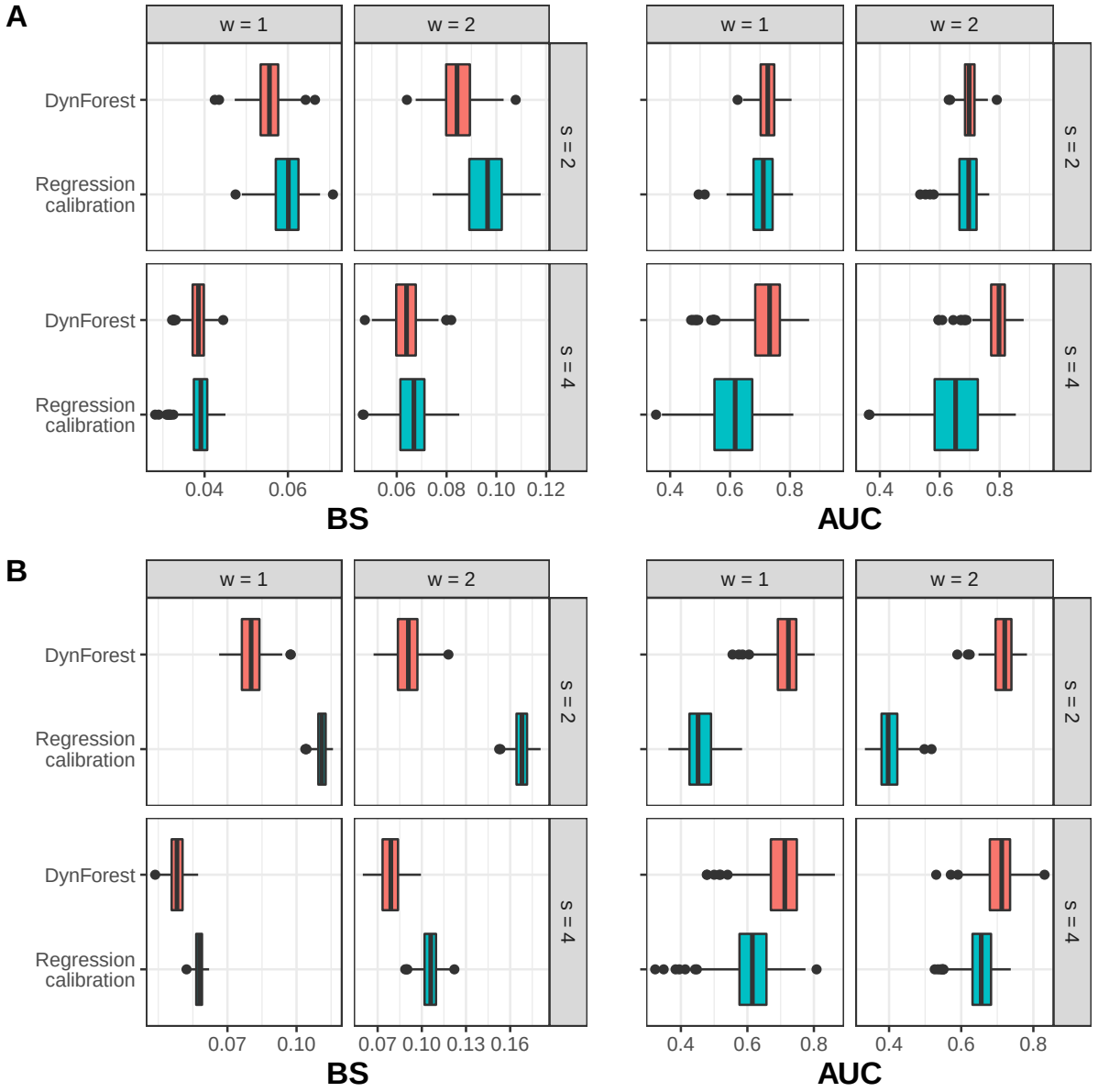



FIGURE IV.3 – External predictive performances of `DynForest` and its regression calibration version in the large dimension scenario of simulations (20 predictors) for the 250 replications. Are reported the Brier Score (BS) and the Area Under the ROC Curve (AUC) at two landmark times $s = 2, 4$ and two horizons $w = 1, 2$. Non-linear association between the markers and the event was displayed using random-effects with two-by-two interactions (A) or latent class membership (B). The regression calibration version of `DynForest` consisted in summarizing the time-dependent markers into time-fixed features once for all prior to inclusion in the RSF. We fixed $nodesize = 3$ and $minsplit = 5$. $mtry$ parameter was also fixed for all replications after tuning process on an unique dataset.

dataset was too computationally intensive. Thus, we decided to tune this parameter using an unique dataset, and used the optimal value for all the replications. With non-linear

association using random-effects plus interactions, optimal values were $mtry = 9$ and $mtry = 46$ for `DynForest` and RC, respectively. They were $mtry = 5$ and $mtry = 6$ when linking the markers to the event through the latent class membership. `DynForest` outperformed the RC technique for both BS and AUC with non-linear association using latent class membership (Figure IV.3B). Under non-linear association using random-effects with interactions (Figure IV.3A), the results were still slightly in favor of `DynForest`. This underlines the substantial impact of not including the time-dependent predictor modeling step within the survival tool construction to correctly account for the correlation between the longitudinal and survival processes and the informative dropout.

## IV.1.4 Application

We aimed at predicting the individual probability of dementia in the elderly in the presence of competing death by leveraging the history of repeated data on clinical exam, neuropsychological battery and brain Magnetic Resonance Imaging (MRI) exam. We relied for this on the Three-City (3C) cohort study [3C Study Group, 2003].

### IV.1.4.1 The 3C study

The 3C study is a French prospective population-based cohort study which enrolled individuals aged 65 years and older from electoral rolls in three French cities (Bordeaux, Dijon and Montpellier). Extensive follow-up interviews were conducted at baseline and then 2, 4, 7, 10, 12, 14 and 17 years after the enrollment including an extensive clinical and neuropsychological exam done in-person at home by a trained psychologist. At 1, 4 and 10 years, a subsample underwent an additional MRI exam. The diagnosis of dementia relied on a 2-step procedure with suspected cased of dementia examined by a clinician and validated by an independent expert committee of neurologists and geriatricians. Deaths were continuously recorded but were considered as a competing event for dementia only in the 3 years after a negative diagnosis. Our analytical sample included all the individuals free of dementia at baseline and with at least 1 measure at each of the 29 predictors

under study during the follow-up in Bordeaux and Dijon cities. This lead to a sample of $N = 2140$ subjects (with 10766 observations) among which 234 were diagnosed with an incident dementia and 311 died before any dementia (Web Figure IV.11 in supplementary material).

We considered a total of 24 time-dependent and 5 time-fixed predictors structured into 9 groups : socio-demographic (time-fixed age at baseline, education, gender), cardio-metabolic factors (3 time-dependent markers with body mass index, diastolic and systolic blood pressure, and 1 time-fixed with diabetes status at baseline), medication (time-dependent number of medication), depressive symptomatology (1 time-dependent scale of depressive symptomatology), cognition (4 time-dependent cognitive tests), functional dependency (1 time-dependent scale of instrumental activities of daily living), genetic (time-fixed APOE4 allele carrier status), neurodegeneration (8 time-dependent brain MRI markers including regional volumes and global measures) and vascular brain lesions (6 time-dependent markers of white matter hyperintensities). Complete information on the predictors are provided in Web Table IV.3 to IV.8 in supplementary material. For longitudinal predictors, individual trajectories are displayed in Web Figures IV.8, IV.9 and IV.10 in supplementary material.

#### IV.1.4.2 `DynForest` specification

The probability of dementia was predicted according to time from the enrollment. MRI data were collected 1.7 times on average and modeled using quadratic and linear trajectories at the population and at the individual level, respectively. Other time-dependent predictors were measured 5.1 times on average. Their trajectories according to time in the study were modeled using natural splines with one internal knot both at the population and individual level. To satisfy the normality assumption of the linear mixed model, all time-dependent predictors were previously normalized using splines transformations [Proust-Lima et al., 2019].

In the absence of an external dataset available with the same longitudinal predictors

and the same target population, predictive abilities were assessed using a 10-fold cross-validation procedure to avoid over-fitting. For each of the 10 folds, `DynForest` was trained on the sample that excluded the fold (learning step on 90%) and individual probabilities of dementia were computed on the fold (prediction step on the remaining 10%). The cross-validation procedure was repeated $R = 50$ times to appreciate the variability of the results. During the learning step, we systematically fixed parameters $minsplit = 5$ and $nodesize = 3$ to favor deep trees. The $mtry$ parameter was tuned within the range of possible value (from 1 to 29 predictors) to minimize the OOB IBS. On the total sample, we first observed that the OOB IBS decreased rapidly with increasing $mtry$ until a stabilization around $mtry = 15$ (Web Figure IV.12 in supplementary material). So for each fold, we ran `DynForest` twice with $mtry = 15$ and $mtry = 20$ and selected the optimal $mtry$ according to the OOB IBS. For the prediction step, individual dementia probabilities were computed for the remaining fold following section IV.1.2.3.

#### IV.1.4.3 Results

To better understand the importance of each predictor, we report the VIMP statistics in Figure IV.4A. The VIMP statistics were computed 10 times and averaged across the replications to reduce the variability due to the permutation procedure. IADL (functional dependency) was the marker the most associated to dementia with a mean gain in IBS of 4.5%, followed by neuro-degeneration markers with the right hippocampus and lobe medio-temporal volumes (gains of 4.2% and 3.1%, respectively), and cognitive tests with the Isaacs Set Test and Benton test (gains of 3.4% and 2.6%, respectively). Since the VIMP may not correctly translate the importance of correlated variables, we also reported in Figure IV.4B the gVIMP grouped by dimensions. The 8 neuro-degeneration predictors reached a mean gain of 10.3% of IBS, and the 4 cognitive tests a mean gain of 9.2%. Then, we observed less importance for the unique marker of functional dependency (mean gain of 4.5%) followed by the 6 markers of vascular brain lesions (mean gain of 3.6%).

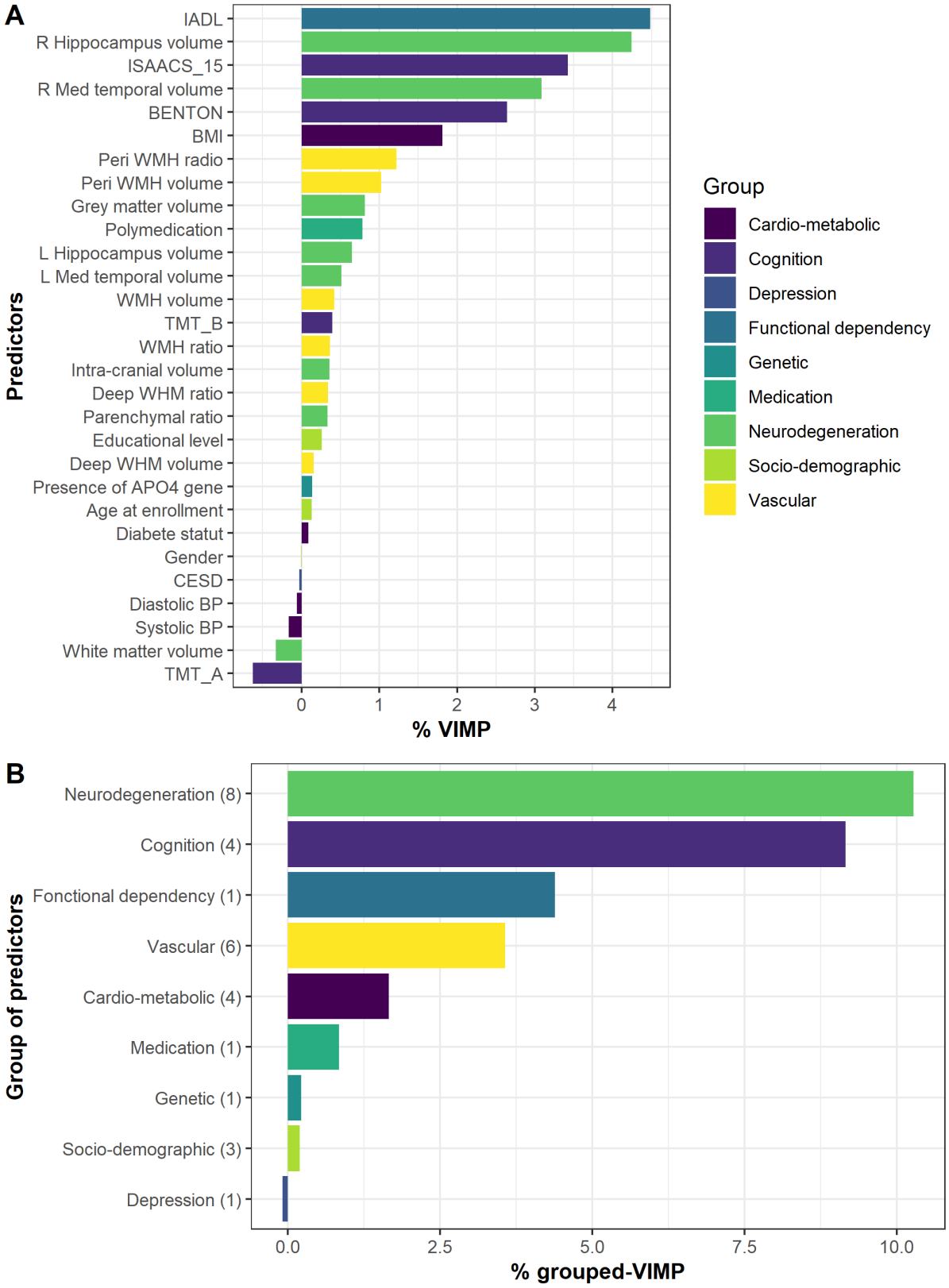



FIGURE IV.4 – (A) Importance variable (VIMP) and (B) grouped importance variable (gVIMP) averaged over 10 permutation procedures for each dementia predictor or group of dementia predictors. Application in the 3C study.

We also computed the minimal depth when using the largest *mtry* hyper-parameter (i.e. $mtry = 29$) (Figure IV.5). IADL (functional dependency) and cognition tests (Isaacs Set Test, Benton test and Trail Making Test A) were the predictors with the lowest average minimal depth, and were selected 100%, 100%, 98% and 97% among the trees, respectively. It means that these predictors were the most effective to split the individuals into homogeneous subgroups according to their risk difference. Except for Trail Making Test A, these results were in accordance with those obtained using the VIMP statistic.

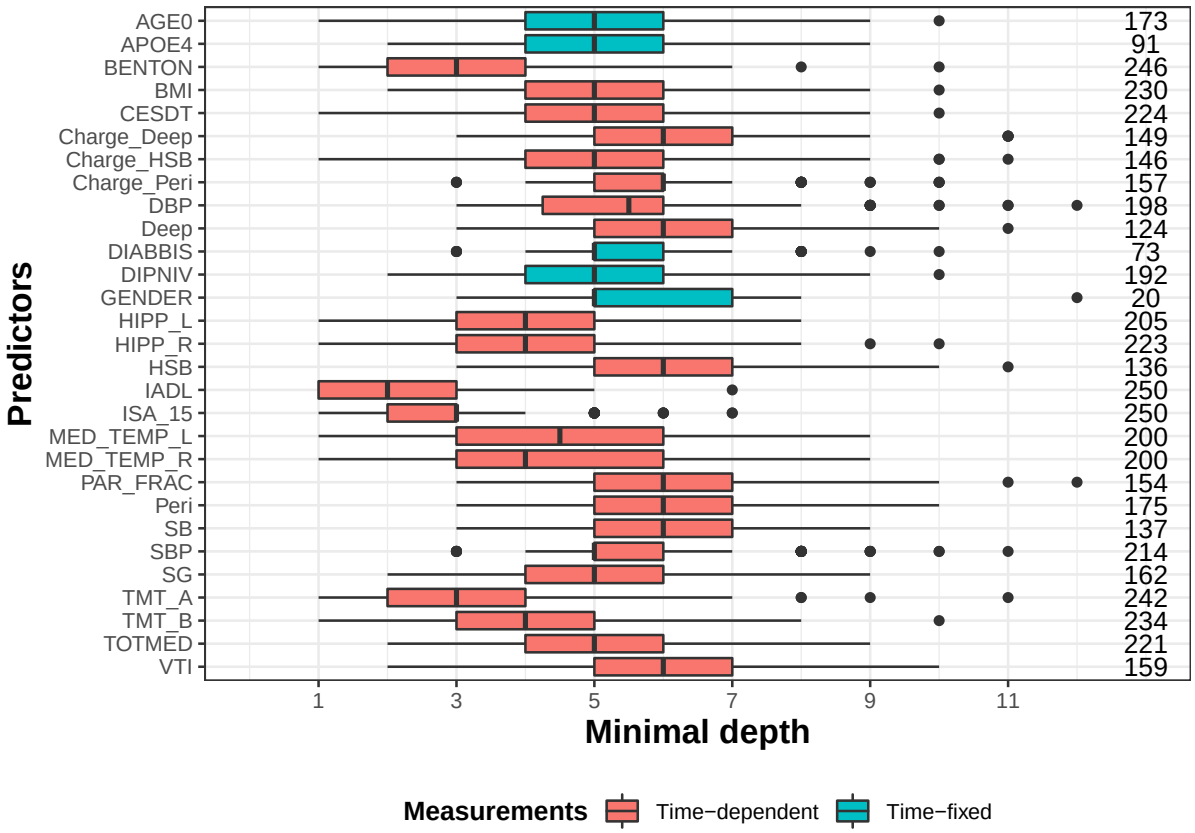



FIGURE IV.5 – Minimal depth computed with the largest *mtry* hyper-parameter (i.e. $mtry = 29$) for each predictor of dementia. We display on the right of the graph the amount of tree where the predictor is found among the 250 trees used to build the random forest.

We then considered two landmark times $s = 5, 10$ years to assess the predictive abilities of `DynForest` to predict dementia between $s$ and $s + w$ (horizon times $w = 3, 5$ years) from individual history up to time $s$. This resulted in 1727 and 1150 individuals still at risk of dementia, respectively. The cross-validated AUC and BS (Figure IV.6) varied from

0.78 to 0.80 and from 0.048 to 0.086 depending on the landmark and the horizon times.

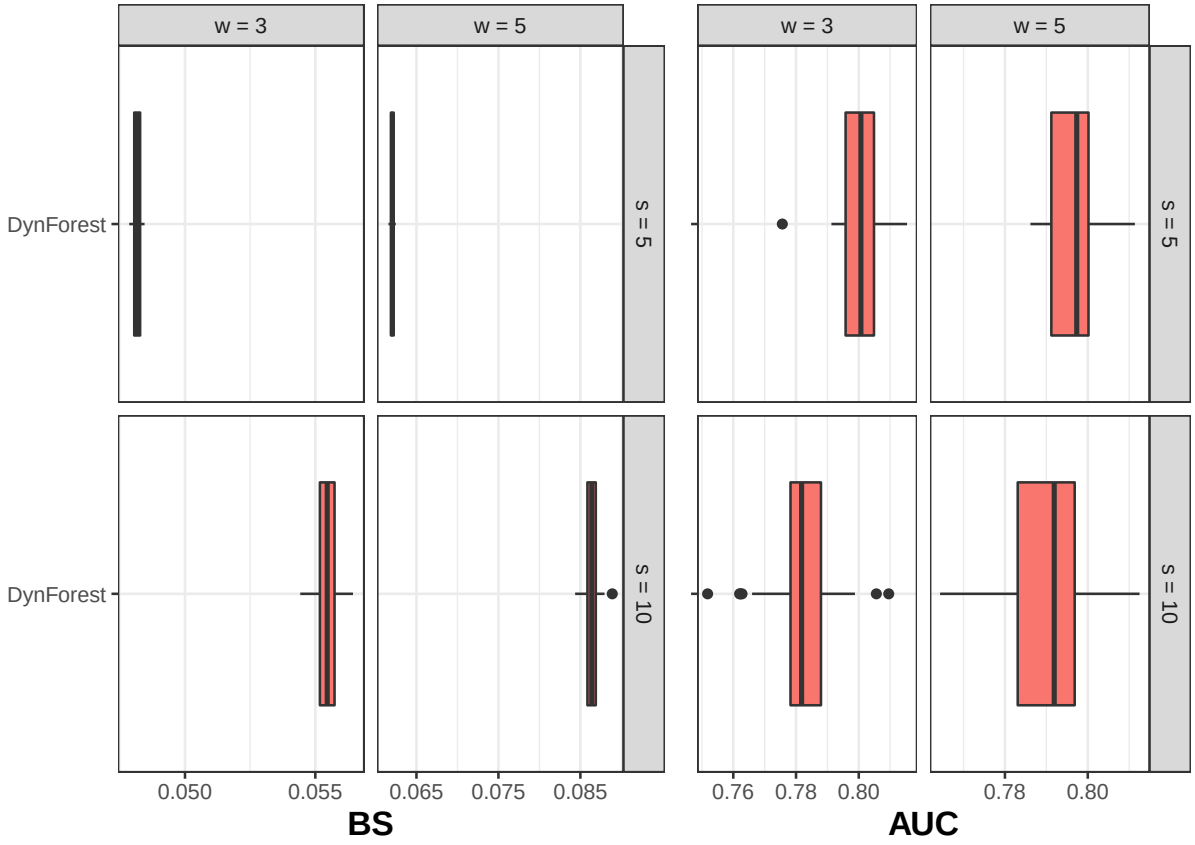



FIGURE IV.6 – Predictive assessment of dementia at landmark times $s = 5, 10$ years and horizon times $w = 3, 5$ years using Brier Score (BS) and Area Under the ROC Curve (AUC). Application in the 3C study.

We finally explored the predictive ability of each predictor in this landmark context by computing the VIMP and gVIMP using only the information prior to 5 years and considering a short span from 5 to 10 years (Web Figure IV.13 in supplementary material). Again, IADL had the largest VIMP value, followed by the Isaacs Set Test and the right hippocampus volume.

### IV.1.5 Discussion

We developed an original methodology, called `DynForest`, to compute individual dynamic predictions from multiple longitudinal predictors. We extended the RSF for competing risks (which were limited so far to time-fixed predictors) [Ishwaran et al., 2008]

to handle endogenous longitudinal predictors. This was achieved by including in the tree building a node-specific internal processing to translate the longitudinal predictors into time-fixed features. `DynForest` can be used to compute individual dynamic predictions of events as well as quantify the importance of the longitudinal predictors using VIMP and grouped-VIMP adapted to longitudinal data.

Through a simulation study, we first showed in a small dimensional context that `DynForest` could be a relevant alternative to the JM reference technique. Indeed, in contrast with JM, `DynForest` does not need to pre-specify the association structure with the event, and may account for nonlinear associations and interactions. In the second scenario, we considered a larger dimensional context, with 20 longitudinal markers, for which JM could not be estimated anymore. We showed, in this larger dimensional scenario, that `DynForest` outperformed the RC alternative proposed in the literature [Li and Luo, 2019, Jiang et al., 2021, Lin et al., 2021]. Indeed, in contrast with RC technique, `DynForest` accounts for the truncation of the repeated data due to the event by re-estimating the mixed models at each node on the node-specific subsample. Since these subsamples become more and more homogeneous regarding the event, the missing at random assumption of the mixed models becomes more and more valid.

Compared with the other methodologies adapted to the large dimensional and longitudinal context, our methodology has the assets of (i) using all available information when landmark approaches [Devaux et al., 2022a, Tanner et al., 2021] only include subjects still at risk at landmark time, resulting in a lack of efficiency [Ferrer et al., 2019] ; (ii) simultaneously analyzing the longitudinal and time-to-event processes when the other methods based on 2-step RC [Li and Luo, 2019, Jiang et al., 2021, Lin et al., 2021] neglect the association leading to a potential bias in the prediction ; (iii) allowing for complex and nonlinear association structures between the predictors and the event ; (iv) allowing the analysis of potentially high dimensional data (i.e. hundreds/thousands of predictors). Indeed, the longitudinal markers are independently modeled so that the method could be easily applied no matter the number of longitudinal markers. Finally, we introduced

two stopping criteria defining the minimum number of events and of subjects required to proceed to a subsequent split. This allows some leaves to have an homogeneous subsample with no events.

Our methodology has also drawbacks. First, although it may be applied whatever the number of predictors, the computation time may become extremely long in high dimensional settings, in particular with a large number of candidates $mtry$. Indeed, mixed models are to be estimated at each node of each tree even though we managed to fasten the estimation by using the estimates previously obtained as initial values. Second, we only considered continuous longitudinal markers only. However, other natures of repeated markers (e.g. binary, categorical, counts) could be considered using generalized mixed models instead. Third, we relied on linear mixed models for deriving time-fixed features. Functional principal components analysis [Yao et al., 2005] could be considered instead. We leave such development for future research. Finally, although we were able to provide the strength of the association between the predictors and the event using the VIMP and gVIMP statistics, these tools do not inform on the sign of the association.

To conclude, using the framework of the random survival forests combined with mixed models for internally processing longitudinal predictors, we tackled the challenge of predicting an event from a potential high number of longitudinal endogenous predictors. `DynForest` offers an innovative solution accompanied by a user-friendly R package.

## IV.1.6 Web supplementary materials

### IV.1.6.1 Simulations

**Generating models** For all the scenarios, we generated repeated data for $m \in \{1, ..., M\}$ longitudinal markers using latent class linear mixed models [Proust-Lima et al., 2014] with $G = 4$ latent classes. The latent class-membership was first defined as a multinomial variable $c_{mi}$. Then, in each latent class $g$, the repeated measures at times $t_{ij}$ of marker $Y_{im}$ were generated according to the following model :

$$Y_{im}(t_{ij})|c_{mi} = g = \beta_{0gm} + \sum_{l=1}^{L-1} \beta_{lgm} * f_{ml}^{L}(t) + b_{i0m} + \sum_{l=1}^{L-1} b_{iml} * f_{ml}^{L}(t) + \epsilon_{ij}$$

The shape of the trajectory over time was defined according to a basis of natural cubic splines with $L$ knots $(f_{ml}^{L}(t))_{l=1,\dots,L-1}$. Depending on the marker and scenario, we chose among : (i) $L = 2$ resulting in a linear trajectory ; (ii) $L = 3$ with one internal knot placed at $t = 5$ ; (iii) $L = 4$ with two internal knots placed at $t = 3$ and $t = 6$. The boundary knots were systematically placed at $t = 0$ and $t = 10$. The basis of splines was associated with fixed effects $\beta_{lgm}$ to define the mean shape of the marker over time, and random effects $b_{im} = (b_{i0m}, b_{ilm})^{\top} \sim \mathcal{N}(0, D)$ to define individual departures from the class-specific mean trajectory. The measurement errors $\epsilon_{ij}$ were zero-mean independent Gaussian variables with variance $\sigma^2$.

Measurements times $t_{ij}$ were generated at baseline and were then randomly drawn around theoretical visits every year up to 10 years. Each year, a departure from the theoretical time was generated according an exponential distribution $\mathcal{E}(5)$.

We assumed a proportional hazard model for the time-to-event with instantaneous risk defined as :

$$\lambda_i(t) = \lambda_0(b, c, t) \exp(\mathcal{P}_i)$$

With $\lambda_0(b, c, t) = cb^c t^{c-1}$ the baseline hazard function from a Weibull distribution with parameters $b$ and $c$, and $\mathcal{P}_i$ the linear predictor. We fixed $b = 0.1$ and $c = 2$ for all scenarios.

**Scenarios** A total of 4 scenarios were considered : 2 with $M = 2$ longitudinal predictors (called Small1, Small2) and 2 with $M = 20$ longitudinal predictors (called Large1, Large2). Details on the association between the predictors and event and associated parameters are given in table IV.1 for small scenarios and table IV.2 for large scenarios. An illustration of the longitudinal trajectories are given in figure IV.7 for the scenario with 2 predictors.

### IV.1.6.2 Applications

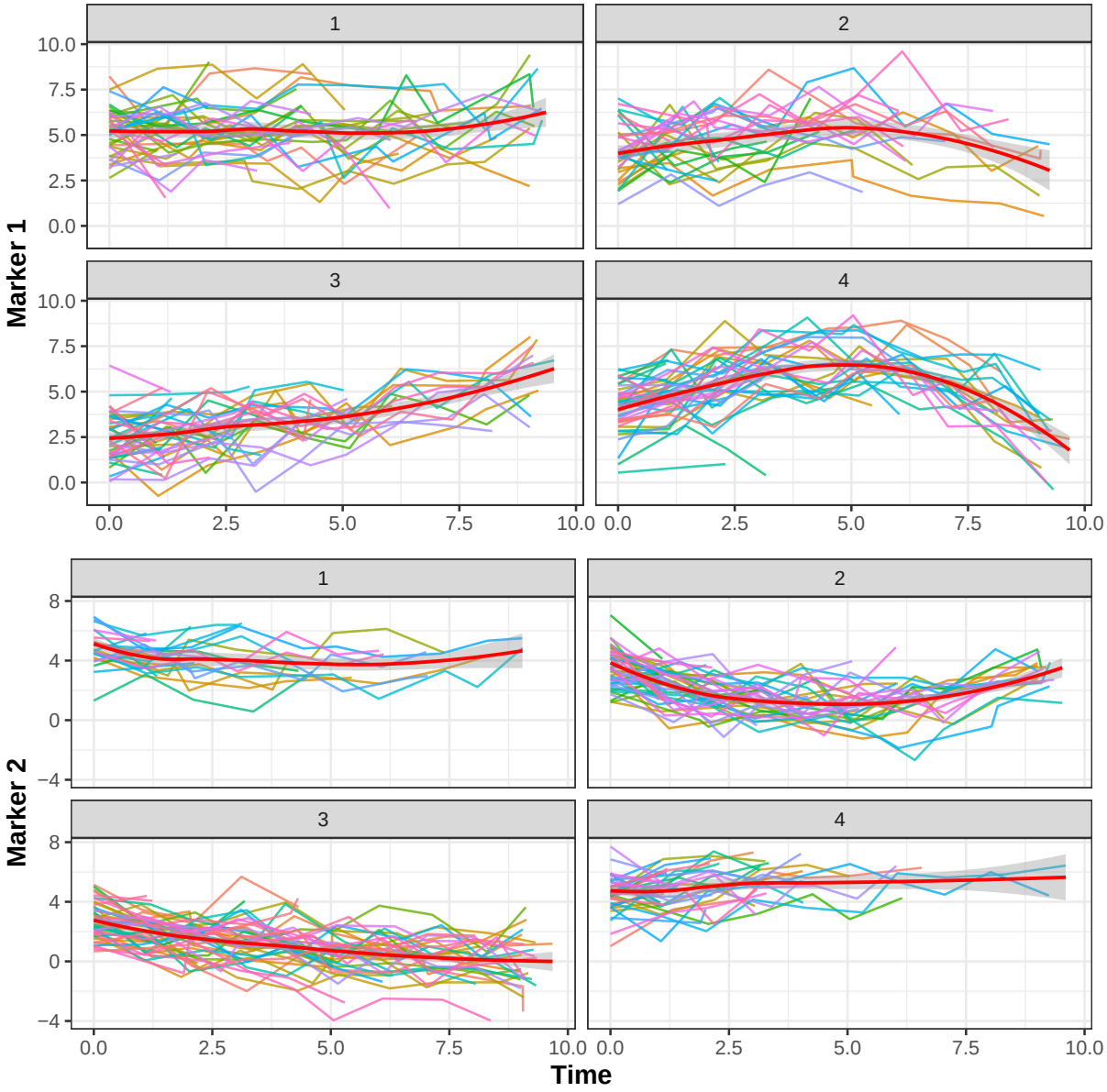
**Individual prediction of dementia** Description and nature of the variables used in the application are detailed from Table IV.3 to Table IV.8. Several groups are also built using these variables (detailed in the same tables as indicated before).

Longitudinal trajectories are displayed in figure IV.8 for normalized clinical and neuropsychological predictors, in figure IV.9 for neuro-degenerative brain-MRI predictors and in figure IV.10 for normalized vascular brain-MRI predictors.

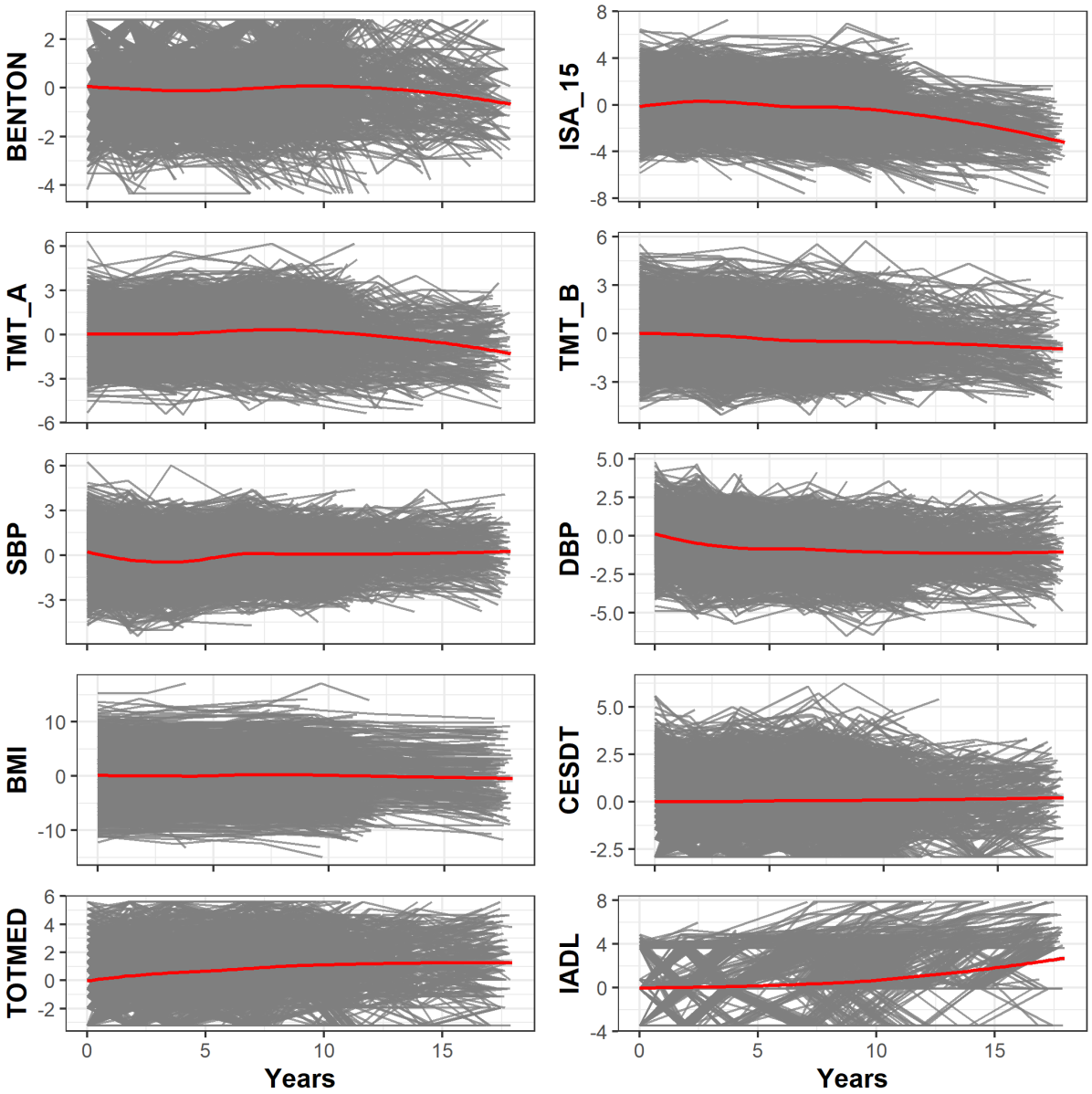
Cumulative incidence functions of dementia and death are displayed in Figure IV.11 over the 20 years of follow-up.

To find the optimal $mtry$ value, we minimized the IBS criteria (see figure IV.12). We found that the lowest IBS value was for $mtry = 21$.

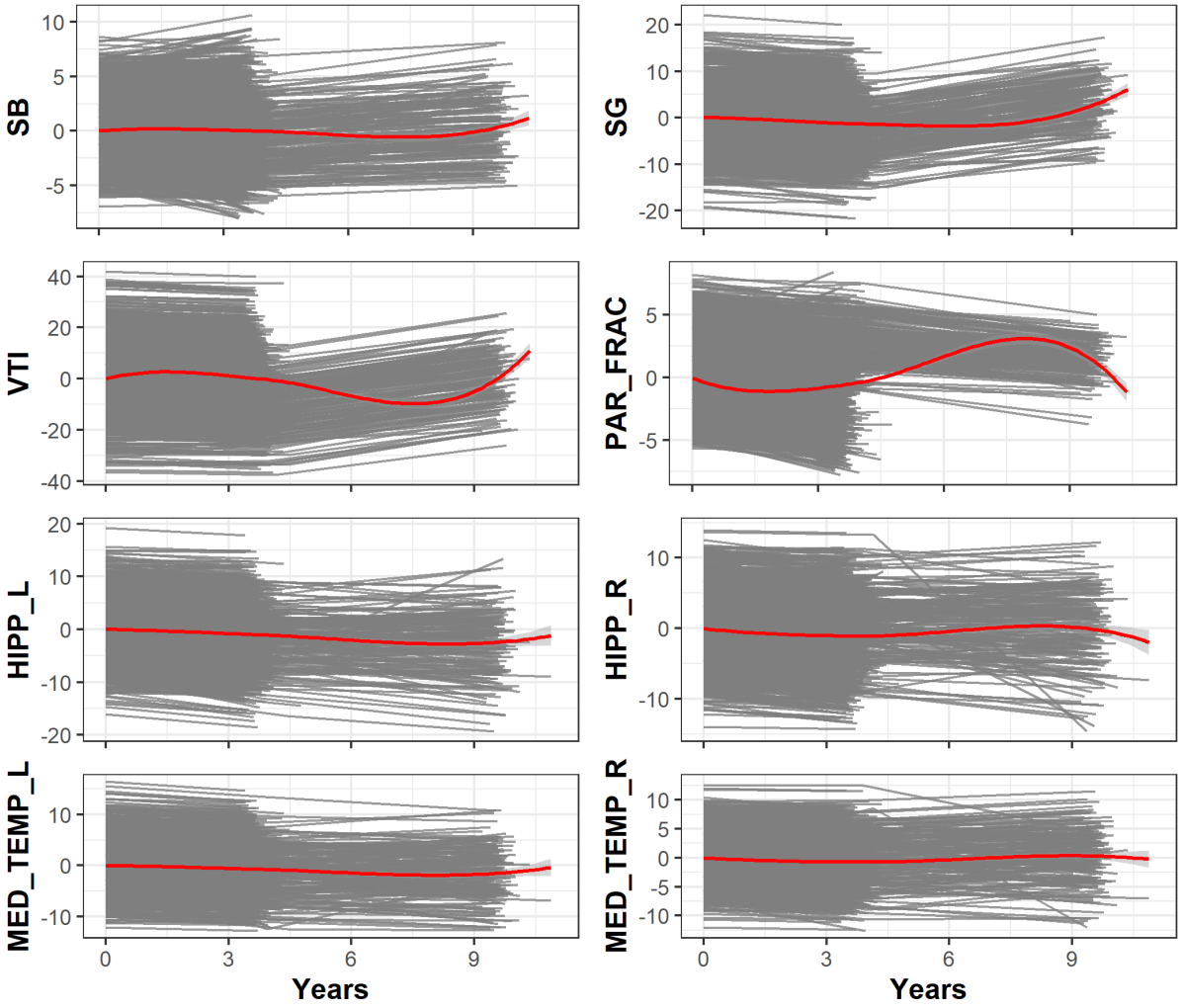
### IV.1.6.3 Figures



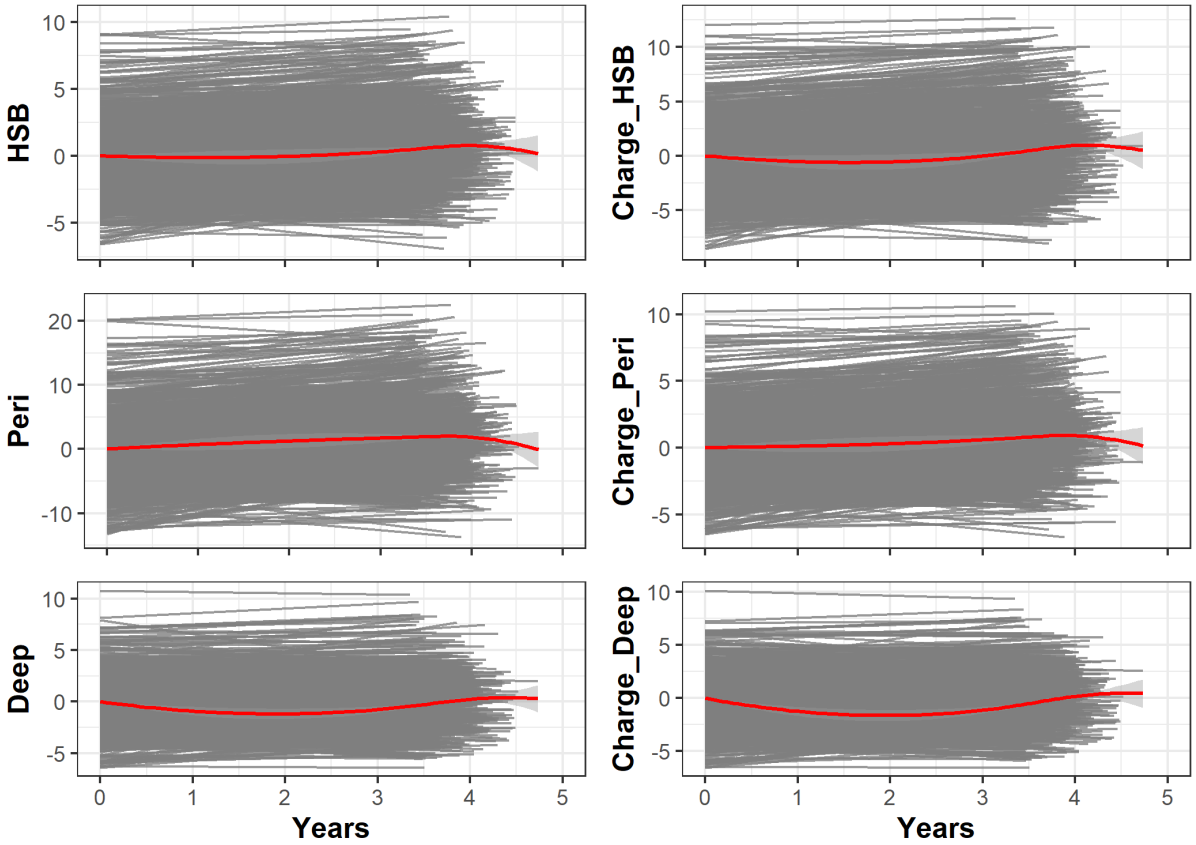

WEB FIGURE IV.7 – Illustration of 200 randomly selected individual trajectories chosen randomly for the two markers up to $t = 10$ in the first simulation study. Individual trajectories are displayed according to the four latent class specific to each marker. Bold red line indicates the mean trajectory given by smoothing method.



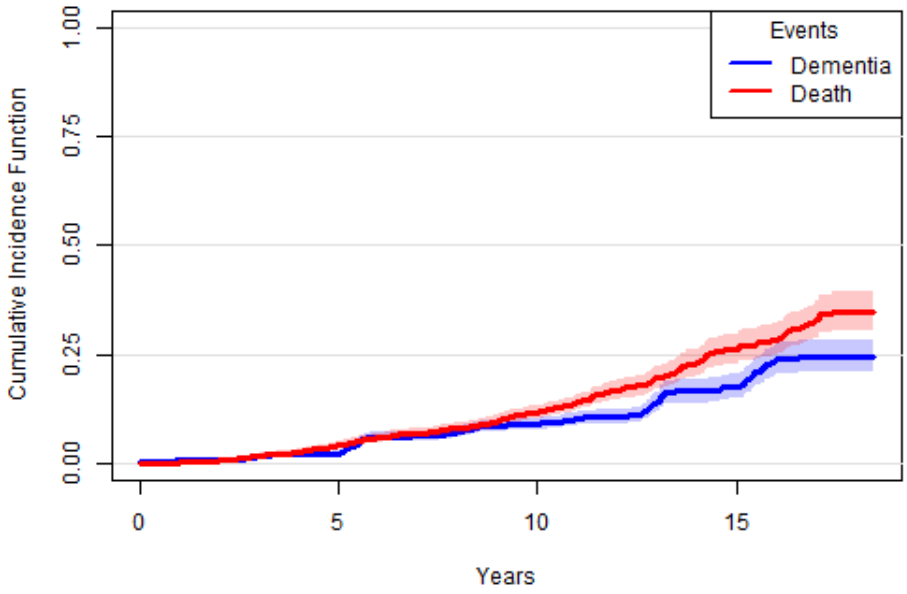

Web Figure IV.8 – Individual trajectories from the normalized clinical and neuropsychological predictors in the 3C study. From left to right, we display the trajectories of Visual retention test of Benton, Isaac Set Test, Trail Making Test A, Trail Making Test B, Systolic Blood Pressure, Distolic Blood Pressure, Body-Mass Index, Depression symptoms, Number of drug consumed by day and Instrumental Activities of Daily Living.



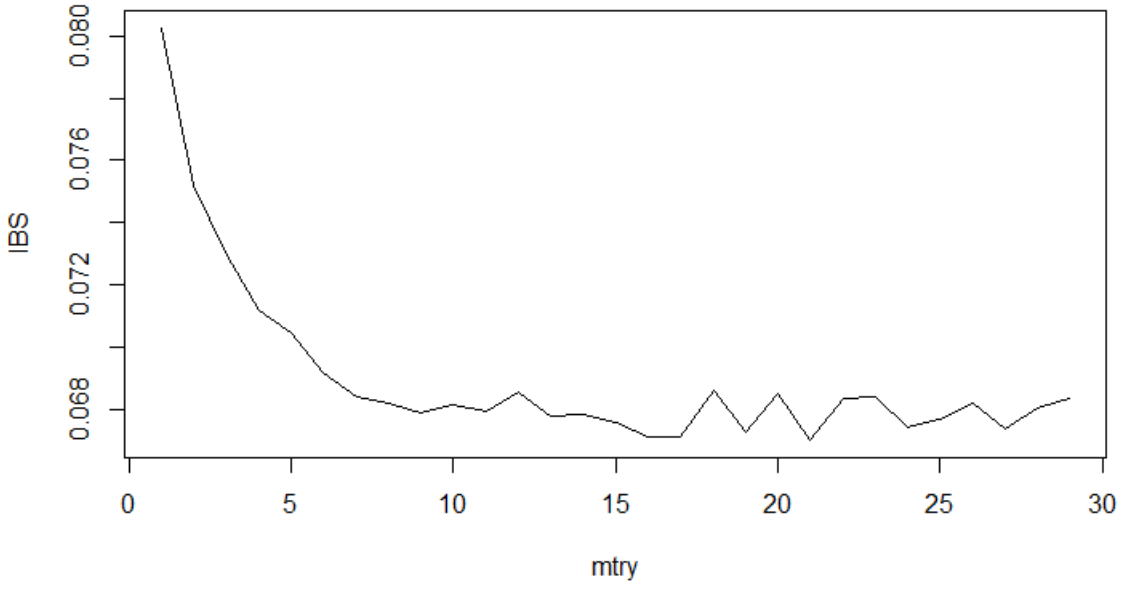

WEB FIGURE IV.9 – Individual trajectories from the normalized neuro-degenerative brain-MRI predictors in the 3C study. From left to right, we display the trajectories of white matter volume, grey matter volume, intracranial volume, parenchymal fraction, left hippocampal volume, right hippocampal volume, left mediotemporal lobe volume and right mediotemporal lobe volume.



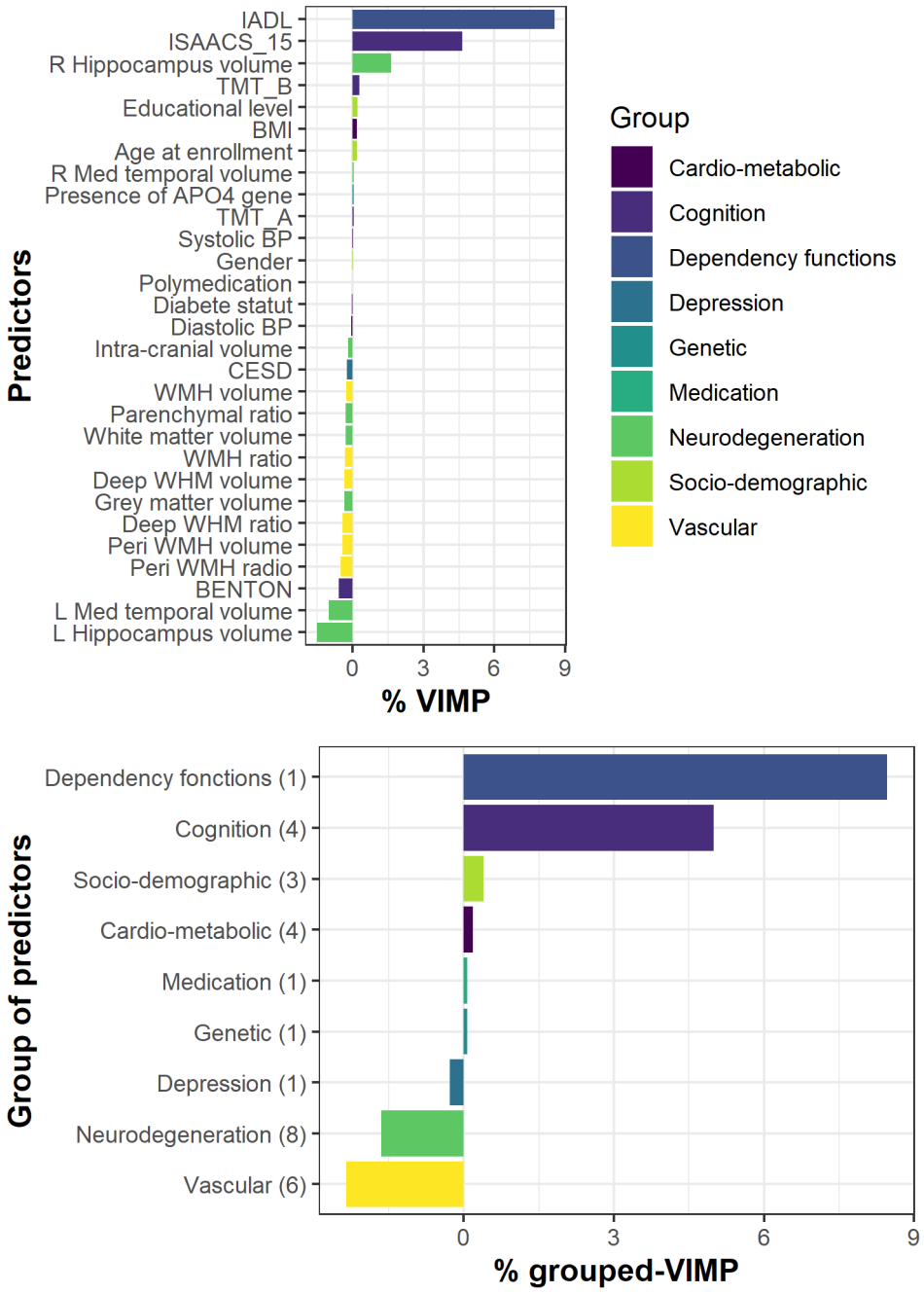

Web Figure IV.10 – Individual trajectories from the normalized vascular brain-MRI predictors in the 3C study. From left to right, we display the trajectories of hypersignals volume in the white matter, proportion of hypersignals in the white matter, hypersignals volume in the periventricular white matter, proportion of hypersignals volume in the periventricular white matter, hypersignals volume in the deep white matter and proportion of hypersignals volume in the deep white matter.




WEB FIGURE IV.11 – Cumulative incidence function for dementia (in blue) and death (in red) event over 17 years of follow-up.




WEB FIGURE IV.12 – Integrated Brier Score (IBS) according to the range of *mtry* parameter values from 1 to 29. Minimal IBS value was found for $mtry = 21$.




Web Figure IV.13 – (A) Importance variable (VIMP) and (B) grouped importance variable (gVIMP) averaged over 10 permutation procedures for each dementia predictor or group of dementia predictors with OOB error computed between 5 and 10 years.

#### IV.1.6.4 Tables

Web Table IV.1 – Association parameters between the markers and the time to event for the small dimension scenarios. As a reminder, $b_{02}$ indicates the baseline random-effect from the marker 2.

| **Scenario** | **Predictor** | $\beta$ **parameter** |
|---|---|---|
| Small1 | $b_{11}$ | 3 |
| | $b_{02}$ | -2 |
| | $b_{11} * b_{02}$ | -3.5 |
| Small2 | $I(c_1 = 1)$ | -2.5 |
| | $I(c_1 = 2)$ | -1 |
| | $I(c_1 = 3)$ | 1.5 |
| | $I(c_1 = 4)$ | 3 |

Web Table IV.2 – Association parameters between the markers and the time to event for the large dimension scenarios. As a reminder, $b_{02}$ indicates the baseline random-effect from the marker 2.

| **Scenario** | **Predictor** | $\beta$ **parameter** |
|---|---|---|
| Large1 | $b_{01}$ | 3 |
| | $b_{12}$ | 1 |
| | $b_{13}$ | -2 |
| | $b_{15}$ | -2 |
| | $b_{01} * b_{12}$ | 2 |
| | $b_{13} * b_{15}$ | -2 |
| Large2 | $I(c_1 = 1)$ | -2 |
| | $I(c_1 = 2)$ | 2 |
| | $I(c_3 = 1)$ | -1 |
| | $I(c_3 = 3)$ | 3 |

Web Table IV.3 – List of variables for the cognition group used in 3C application.

| Abbreviation | Description | Type |
|---|---|---|
| BENTON | Visual retention test of Benton | Continuous time-dependent |
| ISA_15 | Isaac Set Test | Continuous time-dependent |
| TMT_A | Trail Making Test A | Continuous time-dependent |
| TMT_B | Trail Making Test B | Continuous time-dependent |

Web Table IV.4 – List of variables for the cardio-metabolic group used in 3C application.

| Abbreviation | Description | Type |
|---|---|---|
| DIABBIS | Diebete status | Binary covariate |
| SBP | Systolic Blood Pressure | Continuous time-dependent |
| DBP | Diastolic Blood Pressure | Continuous time-dependent |
| BMI | Body-Mass Index | Continuous time-dependent |

Web Table IV.5 – List of variables for the socio-demographic group used in 3C application.

| Abbreviation | Description | Type |
|---|---|---|
| GENDER | Gender | Binary covariate |
| DIPNIV | Educational level | 5-factor covariate |
| AGE0 | Age at enrollment | Continuous covariate |

Web Table IV.6 – List of variables for the neuro-degenerative group used in 3C application.

| **Abbreviation** | **Description** | **Type** |
|---|---|---|
| SB | White matter volume | Continuous time-dependent |
| SG | Gray matter volume | Continuous time-dependent |
| VTI | Intracranial volume | Continuous time-dependent |
| PAR_FRAC | Parenchymal fraction | Continuous time-dependent |
| HIPP_L | Left hippocampal volume | Continuous time-dependent |
| HIPP_R | Right hippocampal volume | Continuous time-dependent |
| MED_TEMP_L | Left mediotemporal lobe volume | Continuous time-dependent |
| MED_TEMP_R | Right mediotemporal lobe volume | Continuous time-dependent |

Web Table IV.7 – List of variables for the vascular charge group used in 3C application.

| **Abbreviation** | **Description** | **Type** |
|---|---|---|
| HSB | Hypersignals volume in the white matter | Continuous time-dependent |
| Charge_HSB | Proportion of hypersignals in the white matter | Continuous time-dependent |
| Peri | Hypersignals volume in the periventricular white matter | Continuous time-dependent |
| Charge_Peri | Proportion of hypersignals volume in the periventricular white matter | Continuous time-dependent |
| Deep | Hypersignals volume in the deep white matter | Continuous time-dependent |
| Charge_Deep | Proportion of hypersignals volume in the deep white matter | Continuous time-dependent |

Web Table IV.8 – List of other variables used in 3C application.

| **Abbreviation** | **Group** | **Description** | **Type** |
|---|---|---|---|
| CESDT | Depression | Depressive symtoms | Continuous time-dependent |
| TOT_MED | Medication | Number of drug consumed by day | Continuous time-dependent |
| IADL | Functional dependency | Instrumental Activities of Daily Living | Continuous time-dependent |
| APOE4 | Genetic | Presence of apolipoprotein e4 allel | Continuous covariate |

# IV.2 Random Forest with longitudinal irregularly measured predictors : The `DynForest` R package

## IV.2.1 Introduction

Random forests are a non-parametric powerful method for prediction purpose. Introduced by Breiman [Breiman, 2001] for classification (categorical outcome) and regression (continuous outcome) framework, random forests are particularly designed to tackle modeling issues in high-dimension context ($n << p$). They can also easily take into account complex association between the outcome and the predictors without any pre-specification where regression models are rapidly limited.

More recently, this methodology was extended to survival data [Ishwaran et al., 2008] and competing events [Ishwaran et al., 2014]. Random forests were implemented in several R [R Core Team, 2019] packages such as `randomForestSRC` [Ishwaran and Kogalur, 2022], `ranger` [Wright and Ziegler, 2017] or `xgboost` [Chen and Guestrin, 2016] among others. However, these packages are all limited to time-fixed predictors. Yet, in many applications, in particular in public health, it may be relevant to include longitudinal predictors, collected at regular or irregular times with measurement errors, in order to improve the predictive performance of the method.

We developed an original methodology based on random forest to incorporate longitudinal predictors that may be prone-to-error and possibly intermittently measured [Devaux et al., 2022b]. This paper aims to describe the `DynForest` R package associated to this methodology, allowing to predict a continuous, categorical or survival outcome using multivariate longitudinal predictors.

In section 2, we briefly present `DynForest` methodology through its algorithm. In section 3, we present the different functions of `DynForest` and we illustrate them in section 4 for a survival outcome, in section 5 for a categorical outcome and in section 6 for a continuous outcome. To conclude, we discuss in section 7 the limitations and future

improvements.

### IV.2.2 `DynForest` principle

`DynForest` methodology is a random forest methodology which can include both time-fixed predictors of any nature and time-dependent predictors possibly measured at irregular times. The purpose of `DynForest` is to predict an outcome which can be categorical, continuous or survival (with possibly competing events).

The random forest should be first build on a learning dataset of $N$ subjects including : $Y$ the outcome ; $\mathcal{M}_x$ an ensemble of $P$ time-fixed predictors ; $\mathcal{M}_y$ an ensemble of $Q$ time-dependent predictors. Through the random forest, an ensemble of $B$ trees are grown as detailed below.

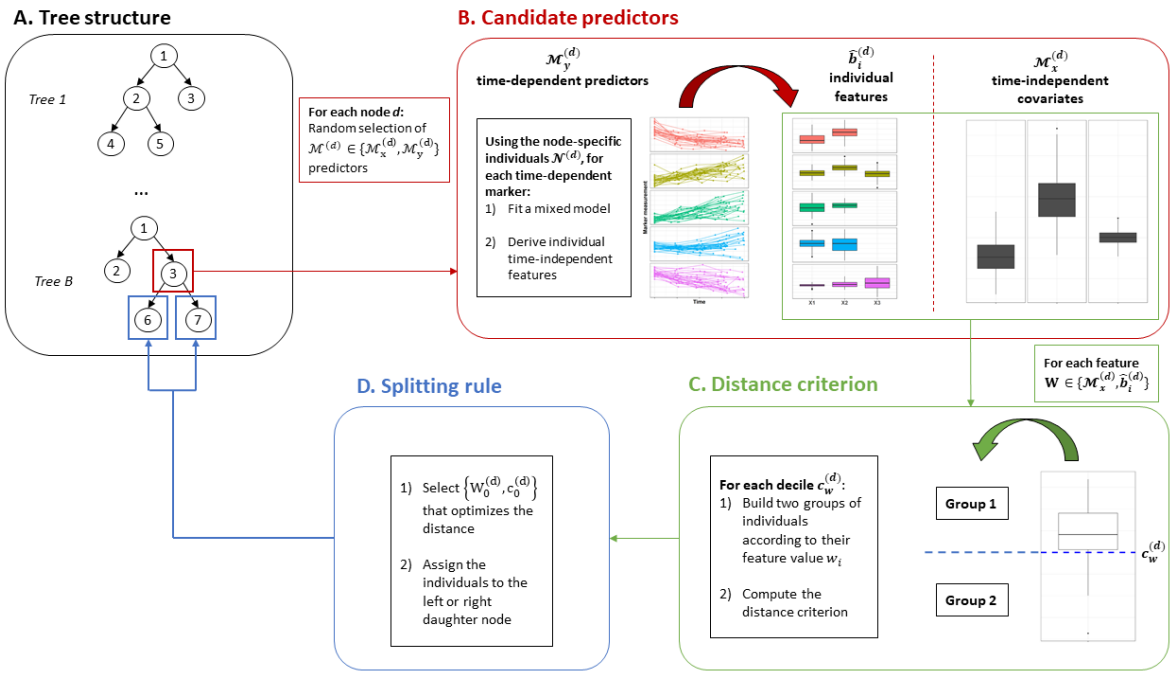



FIGURE IV.14 – Overall scheme of the tree building in `DynForest` with (A) the tree structure, (B) the node-specific treatment of time-dependent predictors to obtain time-fixed features, (C) the dichotomization of the time-fixed features, (D) the splitting rule.

#### IV.2.2.1 The tree building

For each tree $b$ ($b = 1, ..., B$), we aim to partition the subjects into groups that are homogeneous regarding the outcome $Y$. The tree building process is summarized in figure IV.14.

We first draw a bootstrap sample from the original dataset of $N$ subjects. The subjects excluded by the bootstrap constitute the out-of-bag (OOB) sample, noted $OOB^b$ for the tree $b$. Then, at each node $d \in \mathcal{D}$, we recursively repeat the following steps using the $N^{(d)}$ subjects located at node $d$ :

1. An ensemble of $\mathcal{M}^{(d)} = \{\mathcal{M}_x^{(d)}, \mathcal{M}_y^{(d)}\}$ candidate predictors are randomly selected among $\{\mathcal{M}_x, \mathcal{M}_y\}$ (see figure IV.14B). The size of $\mathcal{M}^{(d)}$ is defined by the hyperparameter $mtry$ ;
2. For each longitudinal predictor in $\mathcal{M}_y^{(d)}$ :
   (a) We independently model the trajectory using a flexible linear mixed model [Laird and Ware, 1982] according to time (specified by the user)
   (b) We derive an ensemble $\mathcal{M}_{y\star}^{(d)}$ of individual time-independent features. These features are the individual random-effects $\widehat{b}_i^{(d)}$ of the linear mixed model predicted from the repeated data for individual $i$
   (c) We define $\mathcal{M}_\star^{(d)} = \{\mathcal{M}_x^{(d)}, \mathcal{M}_{y\star}^{(d)}\}$ our new ensemble of candidate features
3. For each candidate feature $W \in \mathcal{M}_\star^{(d)}$ :
   (a) We build a series of splits $c_W^{(d)}$ according to the feature values if continuous, or modalities otherwise (see figure IV.14C), leading to two groups.
   (b) We quantify the distance between the two groups according to the nature of $Y$ :
      - With $Y$ continuous, we compute the weighted within-group variance with the proportion of subjects in each group as weights
      - With $Y$ categorical, we compute the weighted within-group Shannon entropy [Shannon, 1948] (i.e. the amount of uncertainty) with the proportion of subjects in each group as weights

- With $Y$ survival without competing events, we compute the log-rank statistic test [Peto and Peto, 1972]
- With $Y$ survival with competing events, we compute the Fine & Gray statistic test [Gray, 1988]

4. We split the subjects into the two groups that minimize (for continuous and categorical outcome) or maximize (for survival outcome) the distance defined previously. We denote $\{W_0^d, c_0^d\}$ the optimal couple used to split the subjects where they are assigned to the left and right daughter nodes $2d$ and $2d+1$, respectively (see figure IV.14D and A).

To end the recursive procedure described previously, a stopping criterion should be met. We define two stopping criteria : `nodesize` the minimal number of subjects in the nodes and `minsplit` the minimal number of events required to split the node. `minsplit` is only defined with survival outcome. In the following, we call leaves the nodes that cannot be split.

In each leaf $h \in \mathcal{H}$, a summary $\pi^{h^b}$ is computed using the individual belonging to the leaf $h$. The leaf summary is defined according to the outcome. We return :

- the mean, with $Y$ continuous
- the modality with the highest probability, with $Y$ categorical
- the cumulative incidence function over time estimated by Nelson-Aalen [Nelson, 1969, Aalen, 1976], with $Y$ survival without competing events
- the cumulative incidence function over time estimated by Aalen-Johansen [Aalen and Johansen, 1978], with $Y$ survival with competing events

#### IV.2.2.2 Individual prediction of the outcome

**Out-Of-Bag individual prediction** For a subject $\star$, its individual prediction on a single tree $\hat{\pi}^{h^b_\star}$ is obtained by dropping down the subject along the tree. At each node $d$, the subject $\star$ is assigned to the left or right node according to its data and the optimal couple $\{W_0^d, c_0^d\}$. Random-effects for $\star$ are predicted from the individual repeated measures using

the estimated parameters from the linear mixed model if $W_0^d$ is a random-effect feature.

The overall OOB prediction $\hat{\pi}_\star$ can be computed by averaging the tree-based prediction over the random forest as follows :

$$\hat{\pi}_\star = \frac{1}{|\mathcal{O}_\star|} \sum_{b \in \mathcal{O}_\star} \hat{\pi}^{h_\star^b} \tag{IV.10}$$

where $\mathcal{O}_\star$ is the ensemble of trees where $\star$ is $OOB$ and $|\mathcal{O}_\star|$ denotes its length.

**Individual dynamic prediction from a landmark time** With a survival outcome, the OOB prediction described in the previous paragraph can be extended to compute dynamic prediction at landmark $s$ where the longitudinal data are collected until this time. For a new subject $\star$, we thus define the individual prediction $\hat{\pi}_\star(s)$ at landmark time $s$ with :

$$\hat{\pi}_\star(s) = \frac{1}{B} \sum_{b=1}^{B} \hat{\pi}^{h_\star^b}(s) \tag{IV.11}$$

where $\hat{\pi}^{h_\star^b}(s)$ is the tree-based prediction computed by dropping down $\star$ along the tree by considering longitudinal predictors collected until $s$ and time-fixed predictors.

#### IV.2.2.3 Out-Of-Bag prediction error

Using the OOB individual predictions, an OOB prediction error can be internally assessed, in particular to tune the hyperparameters of the random forest. The OOB prediction error estimates the difference between observed and predicted values and is defined according to the nature $Y$ by :

- the mean square error (MSE), with $Y$ continuous
- the missclassification error, with $Y$ categorical
- the Integrated Brier Score (IBS) [Sène et al., 2016], with $Y$ survival

We want to minimize as much as possible the OOB prediction error by tuning `mtry`, `nodesize` and `minsplit` hyperparameters.

#### IV.2.2.4 Explore the most predictive variables

**Variable importance** The variable importance (VIMP) measures the loss of OOB error of prediction [Ishwaran et al., 2008] when removing the link between a predictor and the outcome by using permutation of the predictor values at the subject level (for time-fixed predictors) or at observation level (for time-dependent predictors). In other words, large VIMP value indicates good prediction ability for the predictor.

However, the possible correlation between the predictors may lead to incorrect VIMP statistic [Gregorutti et al., 2017]. The grouped variable importance (gVIMP) can be computed on a group of correlated predictors (defined by the user) in the same fashion as the VIMP statistic, except the permutation is performed simultaneously on all the predictors of the group. Large gVIMP value indicates good prediction ability for the group of predictors.

**Minimal depth** The minimal depth is another statistic to quantify the distance between the root node and the first node for which the predictor is used to split the subjects (1 for first level, 2 for second level, 3 for third level, ...). This statistic can be computed at the predictor level or at the feature level, allowing to fully understand the tree building process.

We strongly advice to compute the minimal depth with `mtry` hyperparameter chosen at its maximum to ensure that all predictors are systematically among candidate predictors for splitting the subjects.

### IV.2.3 `DynForest` R package

`DynForest` methodology was implemented into the R package `DynForest` [Devaux, 2022] freely available on The Comprehensive R Archive Network (CRAN) to users.

The package includes two main functions : `DynForest()` and `predict()` for learning and prediction steps. These functions are fully described in section IV.2.3.1 and IV.2.3.2. Other functions are also available and briefly described in Table IV.9. These functions are

illustrated in an example for survival, categorical and continuous outcome.

TABLE IV.9 – Brief description of the functions available in `DynForest`.

| Function | Description |
|---|---|
| *Learning and prediction steps* | |
| `DynForest()` | Build the random forest with possibly longitudinal predictors with association pre-specified by the user |
| `predict()` | Predict the outcome on new subjects using the individual-specific information that possibly includes longitudinal predictors collected until a landmark time |
| *Assessment function* | |
| `compute_OOBerror()` | Compute the Out-Of-Bag error to be minimized to tune the random forest |
| *Exploring functions* | |
| `compute_VIMP()` | Compute the importance of variables |
| `compute_gVIMP()` | Compute the importance of a group of variables |
| `var_depth()` | Extract information about the tree building process |
| *Plot functions* | |
| `plot_VIMP()` | Plot the importance of variables by value or pourcentage |
| `plot_gVIMP()` | Plot the importance of a group of variables by value or pourcentage |
| `plot_mindepth()` | Plot minimal depth by predictors or features |
| *Other function* | |
| `summary()` | Display information about the type of random forest, predictors included, parameters used, Out-Of-Bag error (if computed) and brief summaries about the leaves |

#### IV.2.3.1 `DynForest()` function

`DynForest()` is the function to build the random forest. The call of this function is :

```
DynForest(timeData = NULL, fixedData = NULL, idVar = NULL,
  timeVar = NULL, timeVarModel = NULL, Y = NULL,
  ntree = 200, mtry = NULL, nodesize = 1, minsplit = 2, cause = 1,
  nsplit_option = "quantile", ncores = NULL,
  seed = round(runif(1, 0, 10000)), verbose = TRUE)
```

**Arguments** Argument `timeData` contains the dataframe in longitudinal format (i.e. one observation per row) for the time-dependent predictors. In addition to time-dependent

predictors, this dataframe should also include an unique identifier and the measurement times. This argument could be set to `NULL` if no time-dependent predictor is included. Argument `fixedData` contains the dataframe in wide format (i.e. one subject per row) for the time-fixed predictors. In addition to time-fixed predictors, this dataframe should also include the same identifier as used in `fixedData`. This argument could be set to `NULL` if no time-fixed predictor is included. Argument `idVar` provides the name of identifier variable included in `timeVar` and `fixedData` dataframes. Argument `timeVar` provides the name of time variable included in `timeData` dataframe. Argument `timeVarModel` contains a list of list for each longitudinal predictor included in `timeData` dataframe. Each list should contain `fixed` and `random` arguments to define the association at population and individual level for the linear mixed model to be estimated ; fixed and random follow the formula from `lcmm` R package [Proust-Lima et al., 2017]. Argument `Y` contains a list of two elements : `type` to define the nature of the outcome (`surv` for survival outcome with possibly competing causes, `scalar` for continuous outcome and `factor` for categorical outcome) and $Y$ the dataframe including the identifier (same as in `timeData` and `fixedData` dataframes) and outcome variables.

Arguments `ntree`, `mtry`, `nodesize` and `minsplit` are the default hyperparameters of the random forest. Argument `ntree` controls the number of trees in the random forest. Argument `mtry` indicates the number of variables randomly drawn at each node. Argument `nodesize` indicates the minimal number of subjects allowed in the leaves. Argument `minsplit` controls the minimal number of events required to split the node (only for survival outcome).

Argument `cause` indicates the event the interest (only for survival outcome). Argument `nsplit_option` indicates the method to build the two groups of individual at each node. By default, we build the groups according to deciles (`quantile` option) but they could be built according to random values (`sample` option).

Argument `ncores` indicates the number of cores used to grow in parallel the trees. Argument `seed` can be fixed to replicate the results. Argument `verbose` allows to display

progression bar during the execution of the function.

**Values** `DynForest()` function returns an object of class `DynForest` containing several elements :

- `data` a list with longitudinal predictors (`Curve` element), continuous predictors (`Scalar` element) and categorical predictors (`Factor` element)
- `rf` a list with several element about the tree building process for each tree in column, which includes :
  - `feuilles` the leaf identifier for the tree subjects
  - `idY` the identifiers for the tree subjects
  - `Y_pred` the estimated outcome in each leaf
  - `model_param` the estimated parameters from the mixed model of the longitudinal predictors used to split the subject for each node
  - `Ytype`, `hist_nodes`, `Y`, `boot` and `Ylevels` internal information used in other functions
- `type` the nature of the outcome
- `times` the event times (only for survival outcome)
- `cause` the cause of interest (only for survival outcome)
- `causes` the causes indicator (only for survival outcome)
- `Inputs` a list of predictors names for Curve (longitudinal predictor), Scalar (continuous predictor) and Factor (categorical predictor)
- `Curve.model` the mixed model specification for each longitudinal predictor
- `param` a list of hyperparameters used to grow the random forest
- `comput.time` the computation time

The main information returned by `rf` is `V_split` element. `V_split` returns a table sorted by the node/leaf identifier (`num_noeud` column) with each row representing a node/leaf. Each column provides information about the splits :

- `type` : the nature of the predictor (`Curve` for longitudinal predictor, `Scalar` for

continuous predictor or `Factor` for categorical predictor) if the node was split, `Leaf` otherwise ;

- `var_split` : the predictor number used for the split ;
- `var_summary` : the predictor summary number used for the split ;
- `threshold` : the threshold used for the split (only with `Curve` and `Scalar`). No information is returned for `Factor` ;
- `N` : the number of subjects in the node/leaf ;
- `Nevent` : the number of events of interest in the node/leaf (only with survival outcome) ;
- `depth` : the depth level of the node/leaf.

**Additional information about the dependencies** `DynForest()` function internally calls other functions from related packages to build the random forest :

- `hlme()` function (from `lcmm` package [Proust-Lima et al., 2017]) to model the time-dependent predictors defined with `timeData` and `timeVarModel` arguments
- `Entropy()` function (from `base` package) to compute the Shannon entropy
- `survdiff()` function (from `survival` package [Therneau, 2022]) to compute the log-rank statistic test
- `crr()` function (from `cmprsk` package [Gray, 2020]) to compute the Fine & Gray statistic test

#### IV.2.3.2 `predict()` function

`predict()` is the function to predict the outcome on new subjects. Landmark time can be specified to consider only longitudinal data collected up to this time to compute the prediction. The call of this function is :

```
predict(object, timeData = NULL, fixedData = NULL,
  idVar, timeVar, t0 = NULL)
```

**Arguments** Argument `object` contains a `DynForest` object from `DynForest()` function. Argument `timeData` contains the dataframe in longitudinal format (i.e. one observation per row) for the time-dependent predictors for new subjects. In addition to time-dependent predictors, this dataframe should also include an unique identifier and the time measurements. This argument can be set to `NULL` if no time-dependent predictor is included. Argument `fixedData` contains the dataframe in wide format (i.e. one subject per row) for the time-fixed predictors for new subjects. In addition to time-fixed predictors, this dataframe should also include an unique identifier. This argument can be set to `NULL` if no time-fixed predictor is included. Argument `idVar` provides the name of identifier variable included in `timeVar` and `fixedData` dataframes. Argument `timeVar` provides the name of time-measurement variable included in `timeData` dataframe. Argument `t0` defines the landmark time ; only the longitudinal data collected up to this time are to be considered.

**Values** `predict()` function returns several elements :

- `t0` the landmark time defined in argument
- `times` times used to compute the individual predictions (only with survival outcome)
- `pred_indiv` the predicted outcome for the new subject. With survival outcome, predictions are provided for each time defined in `times` element.
- `pred_leaf` a table of predicted leaf per subject in row and per tree in column
- `pred_indiv_proba` the proportion of the trees leading to the modality prediction (only with categorical outcome)

## IV.2.4 How to use `DynForest` R package with survival outcome ?

### IV.2.4.1 Introduction to `pbc2` dataset

We use `DynForest` on the pbc2 dataset [Murtaugh et al., 1994] to illustrate our methodology. Data come from the clinical trial conducted by the Mayo Clinic between 1974 and 1984. For the illustration, we consider a subsample of the original dataset resulting to 312

patients and 7 predictors. Among these predictors, the level of serum bilirubin (serBilir), aspartate aminotransferase (SGOT), albumin and alkaline were measured at inclusion and during the follow-up leading to a total of 1945 observations. Sex, age and the drug treatment were collected at the enrollment. During the follow-up, 140 patients died before transplantation, 29 patients were transplanted and 143 patients were alive. The time of first event (alive or any event) was considered as the event time. We aim to predict in this illustration the death without transplantation on patients suffering from primary billiary cholangitis (PBC) using clinical and socio-demographic predictors, considering the transplantation as a competing event.

#### IV.2.4.2 Managing data

To begin, we load `DynForest` package and `pbc2` data and we split the subjects into two datasets : (i) one dataset to train the random forest using 2/3 of patients ; (ii) one dataset to predict on the other 1/3 of patients.

```
# Load package
library(DynForest)

# Split the data for training and prediction steps
set.seed(1234)
id <- unique(pbc2$id)
id_sample <- sample(id, length(id)*2/3)
id_row <- which(pbc2$id%in%id_sample)
pbc2_train <- pbc2[id_row,]
pbc2_pred <- pbc2[-id_row,]
```

Then, we build the dataframe in the longitudinal format (i.e. one observation per row) for the longitudinal predictors including : `id` the unique patient identifier ; `time` the observed time measurements ; `serBilir`, `SGOT`, `albumin` and `alkaline` the longitudinal predictors. We also build the dataframe with the time-fixed predictors including : `id` the unique patient identifier ; `age`, `drug` and `sex` predictors measured at enrollment. The nature of each predictor needs to be properly defined with `as.factor()` function for categorical predictors (e.g. `drug` and `sex`).

```
# Build data objects
```

```
timeData_train <- pbc2_train[,c("id","time",
                                "serBilir","SGOT",
                                "albumin","alkaline")]
fixedData_train <- unique(pbc2_train[,c("id","age","drug","sex")])
```

The first step aims to build the random forest using the `DynForest()` function. We need to specify the mixed model of each longitudinal predictor through a list containing the fixed and random formula for the fixed effect and random effects of the mixed models, respectively. To allow for a flexible trajectory over time, splines can be used in formula using `splines` package.

```
# Create object with longitudinal association for each predictor
timeVarModel <- list(serBilir = list(fixed = serBilir ~ time,
                                     random = ~ time),
                     SGOT = list(fixed = SGOT ~ time + I(time^2),
                                 random = ~ time + I(time^2)),
                     albumin = list(fixed = albumin ~ time,
                                    random = ~ time),
                     alkaline = list(fixed = alkaline ~ time,
                                     random = ~ time))
```

Here, we assume a linear trajectory for `serBilir`, `albumin` and `alkaline`, and quadratic trajectory for `SGOT`.

For this illustration, we build the outcome object containing a list with `type` set to `surv` (for survival data) and `Y` contains a dataframe in wide format (one subject per row) with : `id` the unique patient identifier ; `years` the time-to-event data ; `event` the event indicator.

```
# Create object with the outcome
Y <- list(type = "surv",
          Y = unique(pbc2[,c("id","years","event")]))
```

### IV.2.4.3 Build the random forest

We build the random forest using `DynForest()` function with the following code :

```
# Build the random forest
res_dyn <- DynForest(timeData = timeData_train,
                     fixedData = fixedData_train,
                     timeVar = "time", idVar = "id",
                     timeVarModel = timeVarModel, Y = Y,
                     ntree = 200, mtry = 3, nodesize = 2, minsplit = 3,
```

```
cause = 2, seed = 1234)
```

In a survival context with multiple events, it is also necessary to specify the event of interest with the argument `cause`. We thus fixed `cause` = 2 to specify the event of interest (i.e. the death event). For the hyperparameters, we arbitrarily chose `mtry` = 3, `nodesize` = 2 and `minsplit` = 3.

Overall information about the random forest can be output with `summary()` function. The summary for the random forest in our illustration is displayed below :

```
# Get summary
summary(res_dyn)

DynForest executed with survival (competing risk) mode
        Splitting rule: Fine & Gray statistic test
        Out-of-bag error type: Integrated Brier Score
        Leaf statistic: Cumulative incidence function
----------------
Input
        Number of subjects: 208
        Curve: 4 predictor(s)
        Scalar: 1 predictor(s)
        Factor: 2 predictor(s)
----------------
Tuning parameters
        mtry: 3
        nodesize: 2
        minsplit: 3
        ntree: 200
----------------
----------------
DynForest summary
        Average depth by tree: 6.61
        Average number of leaves by tree: 28.01
        Average number of subjects by leaf: 4.71
        Average number of events of interest by leaf: 1.91
----------------
Out-of-bag error based on Integrated Brier Score
        Tree-based out-of-bag error: Not computed!
        Individual-based out-of-bag error: Not computed!
----------------
Time to build the random forest
        Time difference of 2.685799 mins
----------------
```

We executed `DynForest()` function in survival mode with competing events. In this mode, we use the Fine & Gray statistic test for the splitting rule and the cumulative incidence function (CIF) for the leaf statistic. To build the random forest, we included 208

subjects with 4 longitudinal (Curve), 1 continuous (Scalar) and 2 categorical (Factor) predictors. The `summary()` function also returns some statistics about the trees. For instance, we have on average 4.7 subjects and 1.9 death events by leaf. The number of subjects by leaf should always be higher than `nodesize` hyperparameter. OOB error should be first computed using `compute_OOBerror()` function (see section IV.2.4.4) to be displayed on summary output.

To further investigate the process in each tree, the split details for the tree 1 can be output with the following code :

```
head(res_dyn$rf[,1]$V_split)
```

```
     type num_noeud var_split var_summary   threshold   N Nevent depth
1   Curve         1         3           1 -0.21993804 129     49     1
2   Curve         2         2           1  5.57866304  26     21     2
51 Scalar         3         1          NA 61.83057715 103     28     2
3   Curve         4         2           3  1.42021938  18     13     3
21 Factor         5         1          NA          NA   8      8     3
4   Curve         6         3           2 -0.01010312  92     22     3
```

```
tail(res_dyn$rf[,1]$V_split)
```

```
    type num_noeud var_split var_summary threshold N Nevent depth
46  Leaf       192        NA          NA        NA 4      2     8
47  Leaf       193        NA          NA        NA 2      2     8
48  Leaf       194        NA          NA        NA 2      1     8
20 Curve       195         4           1 -27.58024 4      3     8
49  Leaf       390        NA          NA        NA 2      1     9
50  Leaf       391        NA          NA        NA 2      2     9
```

For instance for the interpretation of the node split, the subjects were split at node 1 using the first random-effect (`var_summary` = 1) of the third `Curve` predictor (`var_split` = 3) with `threshold` = −0.2199. The predictor name can be found using the predictor number in `timeData` and `fixedData` datasets. Therefore, the subjects at node 1 with `albumin` values below to -0.2199 drop in node 2, otherwise in node 3. Another example with the leaves, 4 subjects are included in the 192, and among them 2 subjects have the event of interest.

Estimated cumulative incidence function (CIF) for a single tree can be displayed using `$Y_pred` element of `$rf`. For instance, the CIF of the cause of interest for leaf 192 can be displayed using the following code :

```
# Display CIF for cause of interest
plot(res_dyn$rf[,1]$Y_pred[[192]]$`2`, type = "l", col = "red",
     xlab = "Years", ylab = "CIF", ylim = c(0,1))
```

However, CIF computed on a single tree is not relevant. It should also be computed over all trees of the random forest. For a subject, estimated CIF over the random forest is obtained by averaging the tree-specific CIF of the leaf where the subject belongs. For instance, for subject 104, we display in figure IV.15 the tree-specific CIF for the 9 first trees where this subject is used to grow the trees. This figure shows how the estimated CIF can be differ across the trees and requires to be averaged as they are obtained on a few subjects from the leaves in which subject 104 is assigned.

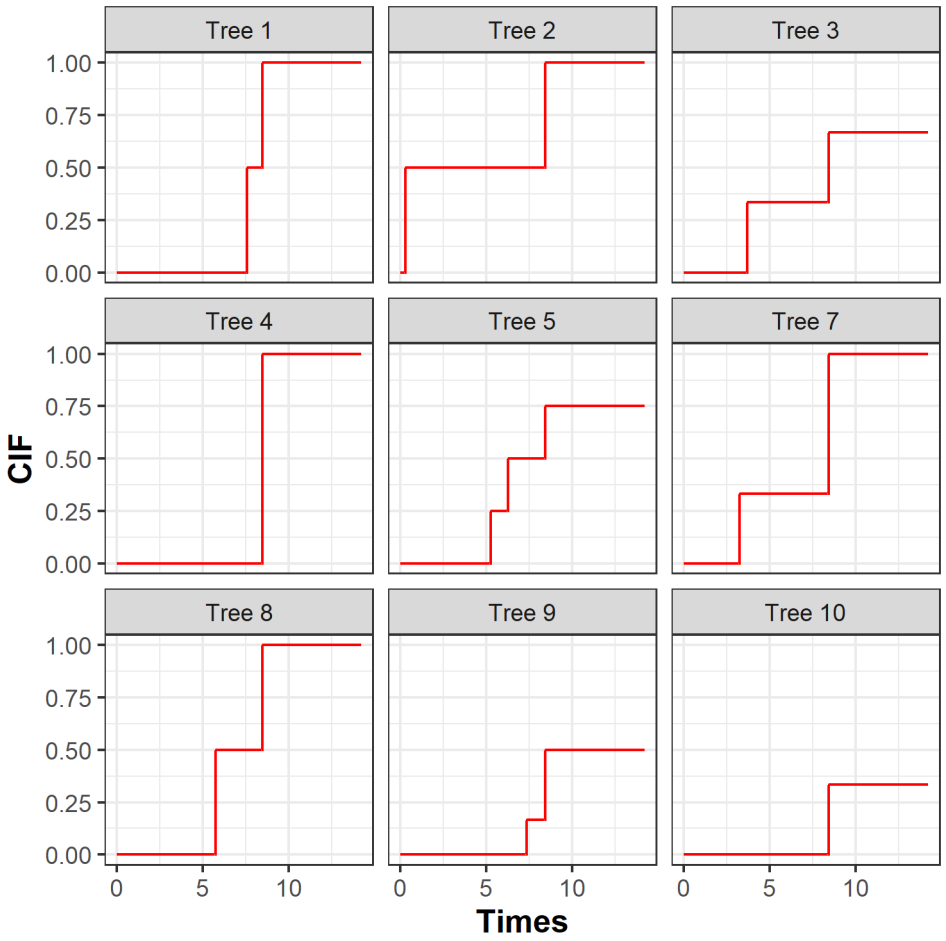



FIGURE IV.15 – Estimated cumulative incidence functions of death before transplantation for subject 104 over 9 trees.

#### IV.2.4.4 Out-Of-Bag error

The Out-Of-Bag error (OOB) aims at assessing the prediction abilities of the random forest. It is computed using `compute_OOBerror()` function with a `DynForest` class object as main argument, such as :

```
# Compute OOB error
res_dyn_OOB <- compute_OOBerror(DynForest_obj = res_dyn)
```

`compute_OOBerror()` returns the OOB error by individual (`$xerror`) or by tree (`$oob.err`). The overall OOB error for the random forest is obtained by averaging the OOB error (by individual or tree). OOB error can also be displayed using `summary()` function applied to the returning object.

In a survival context, the OOB error is evaluated using the Integrated Brier Score (IBS) [Gerds and Schumacher, 2006] from 0 to the maximum event time. The time range can be modified using `IBS.min` and `IBS.max` arguments to define the minimum and maximum, respectively. To maximize the prediction ability of the random forest, we want to minimize the OOB error. This can be done by tuning the hyperparameters (see section IV.2.4.7).

#### IV.2.4.5 Predict the outcome

The software allows to predict the outcome for a new subject using the trained random forest (in section IV.2.4.3). Dynamic predictions can be computed by fixing a prediction time (called landmark time) from which prediction is made using all the history of the individual up to this landmark time (including the longitudinal and time-fixed predictors).

For the illustration, we only select the subjects still at risk at the landmark time of 4 years. We build the dataframe for those subjects and we predict the individual specific CIF using `predict()` function as follows :

```
# Build data for prediction
id_pred <- unique(pbc2_pred$id[which(pbc2_pred$years>4)])
pbc2_pred_tLM <- pbc2_pred[which(pbc2_pred$id%in%id_pred),]
timeData_pred <- pbc2_pred_tLM[,c("id","time",
                                  "serBilir","SGOT",
                                  "albumin","alkaline")]
```

```
fixedData_pred <- unique(pbc2_pred_tLM[,c("id","age","drug","sex")])

# Prediction step
pred_dyn <- predict(object = res_dyn,
                    timeData = timeData_pred,
                    fixedData = fixedData_pred,
                    idVar = "id", timeVar = "time",
                    t0 = 4)
```

`predict()` function provides several elements as described in section IV.2.3.2. To get more graphical results, the `plot_CIF()` function can be used to display the CIF of death before transplantation for given subjects. For instance, we computed the CIF for subjects 102 and 260 with the following code and displayed them on the figure IV.16.

```
# Plot CIF for subjects 102 and 260
plot_CIF(DynForestPred_obj = pred_dyn,
         id = c(102, 260))
```

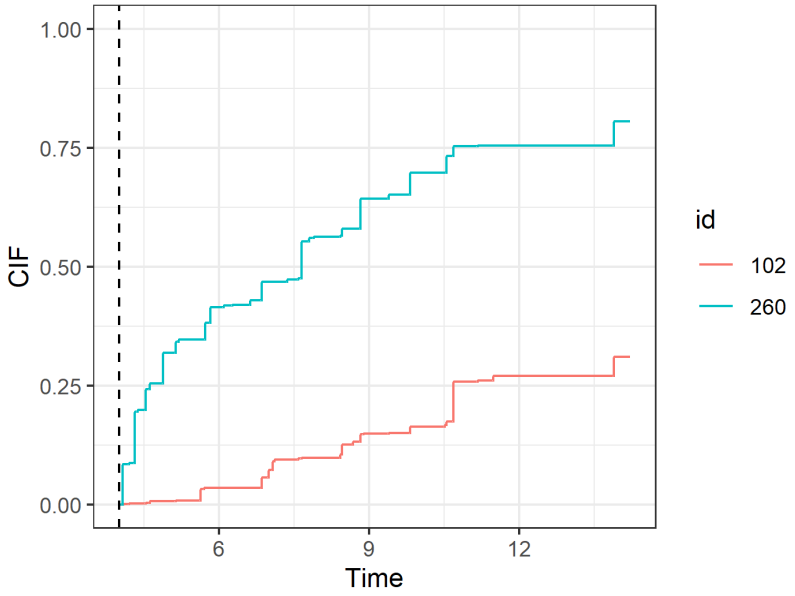



FIGURE IV.16 – Predicted cumulative incidence function for subjects 102 and 260 from landmark time of 4 years (represented by a dashed line).

The first year after the landmark time (at 4 years), we observe a rapid increase of the risk of death for subject 260 compared to subject 102. We also notice that after 10 years from landmark time, subject 260 has a probability of death almost three times higher that the one of subject 102.

### IV.2.4.6 Explore the most predictive variables

**Variable importance** The main objective of the random forest is to predict an outcome. But usually, we are interested to identify which predictors are the most predictive. The VIMP statistic can be computed using `compute_VIMP()` function. This function returns the VIMP statistic for each predictor with `$Importance` element. These results can also be displayed using `plot_VIMP()` function in percentage with `PCT` argument set to TRUE.

```
# Compute VIMP statistic
res_dyn_VIMP <- compute_VIMP(DynForest_obj = res_dyn_OOB)

# Plot VIMP statistic
plot_VIMP(DynForest_obj = res_dyn_VIMP, PCT = TRUE)
```

The VIMP results are displayed in figure IV.17A. The most predictive variables are `serBilir` and `albumin` with the largest VIMP pourcentage. Without `serBilir`, the OOB error of prediction was reduced by 46%.

In the case of correlated predictors, the predictors can be regrouped into dimensions and the VIMP can be computed at the dimension group level with the gVIMP statistic. Permutation is done for each variable of the group simultaneously. The gVIMP is computed with the `compute_gVIMP()` function. This function has the `group` argument to define the group of predictors as a list. For instance, with two groups of predictors (named group1 and group2), the gVIMP statistic is computed using the following code :

```
# Define groups
group <- list(group1 = c("serBilir","SGOT"),
              group2 = c("albumin","alkaline"))

# Compute gVIMP statistic
res_dyn_gVIMP <- compute_gVIMP(DynForest_obj = res_dyn_OOB,
                               group = group)

# Plot gVIMP statistic
plot_gVIMP(DynForest_obj = res_dyn_gVIMP, PCT = TRUE)
```

Similar to VIMP statistic, the gVIMP results can be displayed using `plot_gVIMP()` function. The figure IV.17B shows that group1 has the highest gVIMP percentage with

51%.

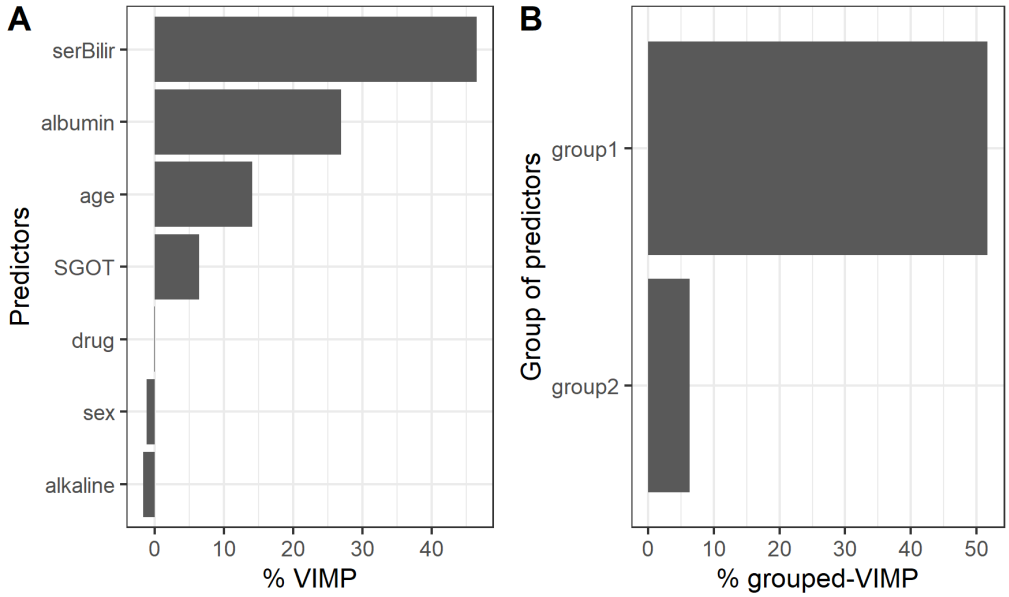



FIGURE IV.17 – Using VIMP statistic (A), we observe that `serBilir` and `albumin` are the most predictive predictors. Using grouped-VIMP statistic (B), group1 (`serBilir` and `SGOT`) has more predictive ability than group2 (`albumin` and `alkaline`).

To compute the gVIMP statistic, the groups can be defined regardless of the number of predictors. However, the comparison between the groups may be harder when group sizes are very different.

**Minimal depth** To go further into the understanding of the tree building process, the `var_depth()` function extracts useful information about the average minimal depth by feature (`$min_depth`), the minimal depth for each feature and each tree (`$var_node_depth`), the number of times that the feature is used for splitting for each feature and each tree (`$var_count`).

From the `var_depth()` object, `plot_mindepth()` function allows to plot the distribution of the average minimal depth across the trees. `plot_level` argument defines how the average minimal depth is plotted, by predictor or feature.

```
# Extract tree building information
depth_dyn <- var_depth(res_dyn)

# Plot average minimal depth by predictor
```

```
plot_mindepth(var_depth_obj = depth_dyn,
              plot_level = "predictor")

# Plot average minimal depth by feature
plot_mindepth(var_depth_obj = depth_dyn,
              plot_level = "feature")
```

The distribution of the minimal depth level is displayed in figure IV.18 by predictor and feature. Note that the minimal depth level should always be interpreted with the number of trees where the predictor/feature is found. Indeed, to accurately appreciate the importance of a variable minimal depth, it has to be part of the candidates at each node. This is why we strongly advice to compute the minimal depth on random forest with `mtry` hyperparameter chosen at its maximum.

In our example, we ran a random forest with mtry hyperparameter set to its maximum (i.e. `mtry` = 7) and we computed the minimal depth on this random forest. We observe that `serBilir`, `albumin` and `age` have the lowest minimal depth, indicating these predictors are used to split the subjects at early stage in 200 out of 200 trees, i.e 100% (figure IV.18A). The minimal depth level by feature (figure IV.18B) provides more advanced details about the tree building process. For instance, we can see that the random-effects for `serBilir` (indicating by bi0 and bi1 on the graph) are the earliest features used on 198 and 194 out of 200 trees, respectively.

#### IV.2.4.7 Guidelines to tune the hyperparameters

The predictive performance of the random forest strongly depends on the hyperparameters `mtry`, `nodesize` and `minsplit`, and should therefore be chosen thoroughly. `nodesize` and `minsplit` hyperparameters control the tree depth. The trees need to be deep enough to ensure that the predictions are accurate. By default in `DynForest()` function, we fixed `nodesize` = 1 and `minsplit` = 2, being the minimum. However, with a large number of individuals, the tree depth could be slighty decreased by increasing these hyperparameters to reduce the computation time.

`mtry` hyperparameter defines the number of predictors randomly drawn at each node.

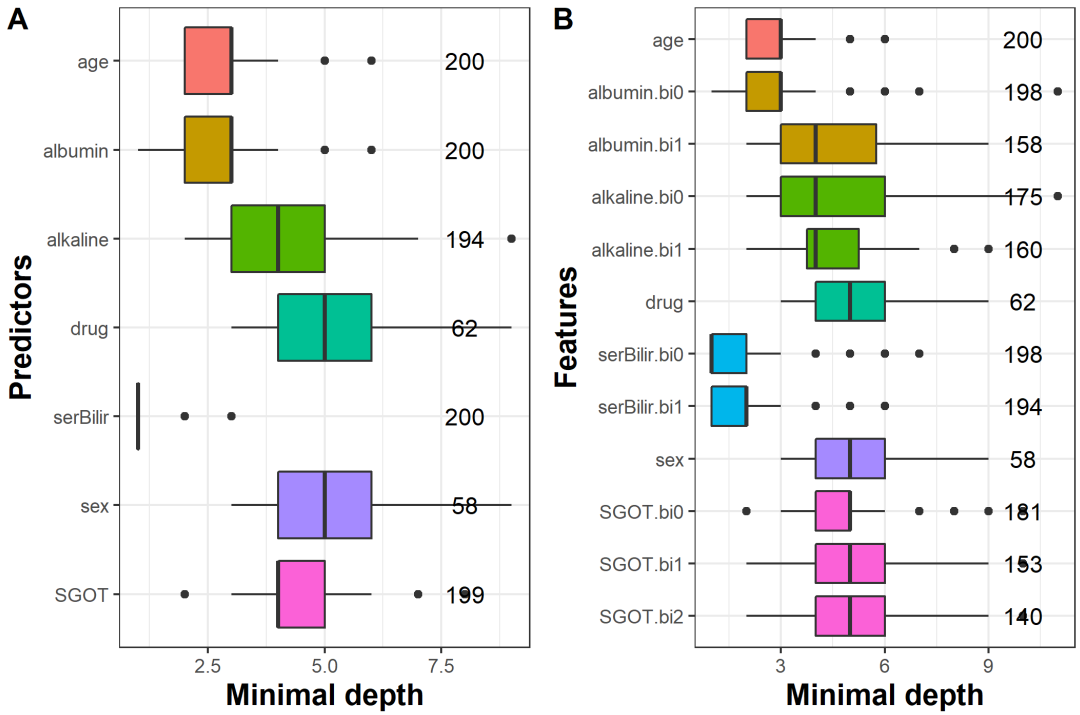



FIGURE IV.18 – Average minimal depth level by predictor (A) and feature (B).

By default, we chose `mtry` equal to the square root of the number of predictors as usually recommanded [Bernard et al., 2009]. However, this hyperparameter should be carefully tuned with possible values between 1 and the number of predictors. Indeed, the predictive performance of the random forest is highly related to this hyperparameter.

In the illustration, we tuned `mtry` for every possible values (1 to 7). The figure IV.19 displays the evolution of the OOB error according to `mtry` hyperparameter.

We can see on this figure large OOB error difference according to `mtry` hyperparameter. In particular, we observe the worst predictive performance for lower values, then similar results with values from 4 to 7. The optimal value (i.e. with the lowest OOB error) was found with `mtry` = 7. This graph reflects how it is crucial to carefully tune this hyperparameter.

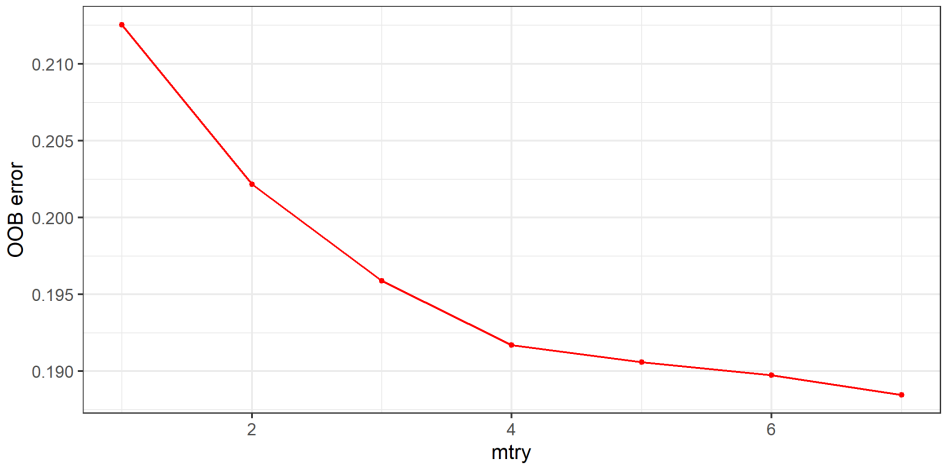



FIGURE IV.19 – OOB error according to `mtry` hyperparameter. The optimal value was found for `mtry` = 7.

### IV.2.5 How to use `DynForest` R package with categorical outcome ?

In this section, we use `DynForest` in a classification perspective using `pbc2` data. For the illustration purpose, we want to predict the death at 10 years on subjects still at risk at 4 years from the repeated data up to 4 years. Note that this is only for illustrative purpose as this technique does not handle censoring correctly.

#### IV.2.5.1 Managing data

For the illustration, we select patients still at risk at 4 years and we recode the `event` variable with `event` = 1 for subjects died during between 4 years and 10 years, `event` = 0 otherwise. We split the subjects into two datasets : (i) one dataset to train the random forest using 2/3 of patients ; (ii) one dataset to predict on the other 1/3 of patients.

```
# Load and manage data
library(DynForest)

# Select subjects alive at 4 years
# from repeated data collected up to 4 years
pbc2 <- pbc2[which(pbc2$years>4&pbc2$time<=4),]

# Death event between 4 years and 10 years (event = 1)
# Otherwise (event = 0)
pbc2$event <- ifelse(pbc2$event==2, 1, 0)
```

```
pbc2$event[which(pbc2$years>10)] <- 0

# Split the data for training and prediction steps
set.seed(1234)
id <- unique(pbc2$id)
id_sample <- sample(id, length(id)*2/3)
id_row <- which(pbc2$id%in%id_sample)
pbc2_train <- pbc2[id_row,]
pbc2_pred <- pbc2[-id_row,]
```

We use the same strategy as in the survival context (section IV.2.4) to build the random forest, with the same predictors and the same association for time-dependent predictors.

```
# Build longitudinal data
timeData_train <- pbc2_train[,c("id","time",
                                "serBilir","SGOT",
                                "albumin","alkaline")]

# Specify modeling for each time-dependent predictors
timeVarModel <- list(serBilir = list(fixed = serBilir ~ time,
                                     random = ~ time),
                     SGOT = list(fixed = SGOT ~ time + I(time^2),
                                 random = ~ time + I(time^2)),
                     albumin = list(fixed = albumin ~ time,
                                    random = ~ time),
                     alkaline = list(fixed = alkaline ~ time,
                                     random = ~ time))

# Build fixed data
fixedData_train <- unique(pbc2_train[,c("id","age","drug","sex")])
```

Using categorical outcome, the definition of the output object is slightly different. We should specify `type`="factor" to define the categorical outcome, and the dataframe in `Y` should contain only 2 columns, the variable identifier `id` and the outcome `death`.

```
# Build outcome object
Y <- list(type = "factor",
          Y = unique(pbc2_train[,c("id","event")]))
```

#### IV.2.5.2 Build the random forest

We executed `DynForest()` function to build the random forest with hyperparameters `mtry` = 7 and `nodesize` = 2 and we displayed the results using `summary()` function as follows :

```
# Run DynForest function
res_dyn <- DynForest(timeData = timeData_train,
                     fixedData = fixedData_train,
                     timeVar = "time", idVar = "id",
                     timeVarModel = timeVarModel,
                     mtry = 7, nodesize = 2,
                     Y = Y, seed = 1234)

# Get summary
summary(res_dyn)

DynForest executed with classification mode
        Splitting rule: Minimize weighted within-group Shannon entropy
        Out-of-bag error type: Missclassification
        Leaf statistic: Majority vote
----------------
Input
        Number of subjects: 150
        Curve: 4 predictor(s)
        Scalar: 1 predictor(s)
        Factor: 2 predictor(s)
----------------
Tuning parameters
        mtry: 7
        nodesize: 2
        ntree: 200
----------------
----------------
DynForest summary
        Average depth by tree: 5.84
        Average number of leaves by tree: 16.68
        Average number of subjects by leaf: 9.3
----------------
Out-of-bag error based on Missclassification
        Tree-based out-of-bag error: Not computed!
        Individual-based out-of-bag error: Not computed!
----------------
Time to build the random forest
        Time difference of 1.738206 mins
----------------
```

In this illustration, we built the random forest using 150 subjects because we only kept the subjects still at risk at landmark time at 4 years. We have on average 9.3 subjects by leaf, and the average depth level by tree is 5.8.

#### IV.2.5.3 Out-Of-Bag error

With categorical outcome, the OOB prediction error is evaluated using a missclassification criterion. This criterion can be computed with `compute_OOBerror()` function as

follows :

```
# Compute OOB error
res_dyn_OOB <- compute_OOBerror(DynForest_obj = res_dyn)

# Get OOB error over the random forest
mean(res_dyn_OOB$oob.err)
[1] 0.23
```

With our random forest, we predicted the wrong outcome for 23% of the subjects. This criterion should be minimized as possible, by tuning `mtry` and `nodesize` hyperparameters.

#### IV.2.5.4 Predict the outcome

We then predict the probability of death on subjects still at risk at landmark time at 4 years. In classification mode, the predictions are performed using majority vote. The prediction over the trees is thus a modality of the categorical outcome along with the proportion of the trees which lead to this modality. Prediction are computed using `predict()` function, then a dataframe can be easily built from returning object to get the prediction and probability for each subject as followed :

```
# Build data for prediction
timeData_pred <- pbc2_pred[,c("id","time",
                              "serBilir","SGOT",
                              "albumin","alkaline")]

fixedData_pred <- unique(pbc2_pred[,c("id","age","drug","sex")])

# Predict the outcome for new subjects
pred_dyn <- predict(object = res_dyn,
                    timeData = timeData_pred,
                    fixedData = fixedData_pred,
                    idVar = "id", timeVar = "time",
                    t0 = 4)

# Table with prediction and probability
head(data.frame(pred = pred_dyn$pred_indiv,
                proba = pred_dyn$pred_indiv_proba))

    pred proba
101    0 0.960
104    0 0.785
106    1 0.575
108    0 0.945
112    1 0.540
114    0 0.605
```

As shown in this example, some predictions are made with varying confidence from 54.0% for subject 112 to 96.0% for subject 101. We predict no event for subject 101 with a probability of 96.0% and an event for subject 106 with a probability of 57.5%.

#### IV.2.5.5 Explore the most predictive variables

**Variable importance** The most predictive variables can be computed using `compute_VIMP()` and displayed using `plot_VIMP()` function as followed :

```
# Compute VIMP statistic
res_dyn_VIMP <- compute_VIMP(DynForest_obj = res_dyn_OOB)

# Plot VIMP statistic
plot_VIMP(DynForest_obj = res_dyn_VIMP, PCT = TRUE)
```

Again, we found that the most predictive variables were `albumin`, `age` and `serBilir` for which the OOB prediction error was reduced by 22%, 7% and 6%, respectively.

**Minimal depth** The minimal depth is computed using `var_depth()` function and is displayed at predictor and feature level using `plot_mindepth()` function. The results are displayed in figure IV.20 using the random forest with maximal `mtry` hyperparameter value (i.e. `mtry` = 7) for better understanding.

```
# Extract tree building information
depth_dyn <- var_depth(res_dyn)

# Plot average minimal depth by predictor
plot_mindepth(var_depth_obj = depth_dyn,
              plot_level = "predictor")

# Plot average minimal depth by feature
plot_mindepth(var_depth_obj = depth_dyn,
              plot_level = "feature")
```

We observe that `serBilir` and `albumin` have the lowest minimal depth : these predictors are used to split the subjects in 199 and 196 out of 200 trees, respectively (figure IV.20A). The figure IV.20B provides further results. In particular, this graph shows the baseline random-effect (indicated by bi0) of `serBilir` and `albumin` are the earliest predictors used to split the subjects with 197 and 192 out of 200 trees, respectively.

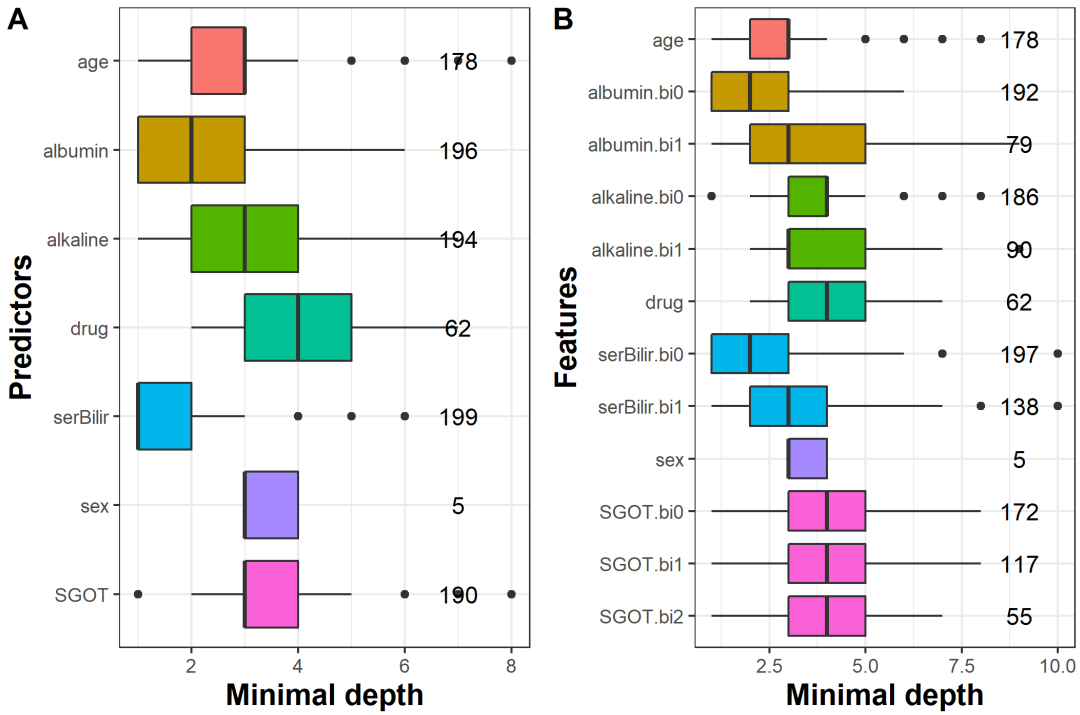



FIGURE IV.20 – Average minimal depth level by predictor (A) and feature (B).

### IV.2.6 How to use `DynForest` R package with continuous outcome ?

In this section, we present an illustration of `DynForest` with a continuous outcome. `DynForest` was used on a simulated dataset with 200 subjects and 10 predictors (6 time-dependent and 4 time-fixed predictors). The 6 longitudinal predictors were generated using a linear mixed model with linear trajectory according to time. We considered 6 measurements by subject (at baseline and then randomly drawn around theoretical annual visits up to 5 years). Then, we generated the continuous outcome using a linear regression with the baseline random-effect of marker 1 and slope random-effect of marker 2 as linear predictors. We generated two datasets, one for each step (training and prediction).

The aim of this illustration is to predict the continuous outcome using time-dependent predictors collected up to landmark time at 5 years and time-fixed predictors.

### IV.2.6.1 Managing data

First of all, we load the data and we build the mandatory objects needed to execute `DynForest()` function. We specify the model for the longitudinal predictors. For the illustration, we consider linear trajectories over time for the 6 longitudinal predictors.

```
# Load data
load(data_simu_cont)

# Build predictors objects
timeData_train <- data_train[,c("id","time",
                                paste0("marker",seq(6)))]

timeVarModel <- lapply(paste0("marker",seq(6)),
                       FUN = function(x){
                         fixed <- reformulate(termlabels = "time",
                                              response = x)
                         random <- ~ time
                         return(list(fixed = fixed, random = random))
                       })

fixedData_train <- unique(data_train[,c("id",
                                        "cont_covar1","cont_covar2",
                                        "bin_covar1","bin_covar2")])
```

To define the object for a continuous outcome, the `type` argument should be chosen to "scalar" to set up the random forest in regression mode. The dataframe `Y` should include two columns with the unique identifier `id` and the continuous outcome `Y_res`.

```
# Build outcome object
Y <- list(type = "scalar",
          Y = unique(data_train[,c("id","Y_res")]))
```

### IV.2.6.2 Build the random forest

To build the random forest, we chose default hyperparameters (i.e. `ntree` = 200 and `nodesize` = 1), except for `mtry` which was fixed to 10. We ran `DynForest()` and we provided overall results with `summary()` function with the following code :

```
# Run DynForest function
res_dyn <- DynForest(timeData = timeData_train,
                     fixedData = fixedData_train,
                     timeVar = "time", idVar = "id",
                     timeVarModel = timeVarModel,
```

```
                    mtry = 10,
                    Y = Y, seed = 1234)

# Get some summaries
summary(res_dyn)

DynForest executed with regression mode
        Splitting rule: Minimize weighted within-group variance
        Out-of-bag error type: Mean square error
        Leaf statistic: Mean
----------------
Input
        Number of subjects: 200
        Curve: 6 predictor(s)
        Scalar: 2 predictor(s)
        Factor: 2 predictor(s)
----------------
Tuning parameters
        mtry: 10
        nodesize: 1
        ntree: 200
----------------
----------------
DynForest summary
        Average depth by tree: 9.06
        Average number of leaves by tree: 126.47
        Average number of subjects by leaf: 3.03
----------------
Out-of-bag error based on Mean square error
        Tree-based out-of-bag error: Not computed!
        Individual-based out-of-bag error: Not computed!
----------------
Time to build the random forest
        Time difference of 4.435966 mins
----------------
```

The random forest was executed in regression mode (for a continuous outcome). The splitting rule aimed to minimize the weighted within-group variance. We built the random forest using 200 subjects and 10 predictors (6 time-dependent and 4 time-fixed predictors) with hyperparameters `ntree` = 200, `mtry` = 10 and `nodesize` = 1. As we can see, `nodesize` = 1 leads to deeper trees (the average depth by tree is 9.1) and few subjects by leaf (3 on average).

#### IV.2.6.3 Out-Of-Bag error

For continuous outcome, the OOB prediction error is evaluated using mean square error (MSE). We used `compute_OOBerror()` function to compute the OOB prediction

error by individual. We then average it to get the overall OOB prediction error over the random forest.

```
# Compute OOB error
res_dyn_OOB <- compute_OOBerror(DynForest_obj = res_dyn)

# Get OOB error over the random forest
mean(res_dyn_OOB$oob.err)
[1] 0.2100969
```

We obtained 0.21 for the MSE. This quantity needs to be minimized using hyperparameters `mtry` and `nodesize`.

#### IV.2.6.4 Predict the outcome

In regression mode, the tree-specific predictions are averaged across the trees to get an unique prediction over the random forest. `predict()` function provides the individual predictions. We first created the objects with the same predictors used to build the random forest. We then predicted the continuous outcome by using the data collected up to landmark time at 5 years.

```
# Build object for prediction
timeData_pred <- data_pred[,c("id","time",
                              paste0("marker",seq(6)))]

fixedData_pred <- unique(data_pred[,c("id","cont_covar1","cont_covar2",
                                      "bin_covar1","bin_covar2")])

# Prediction step
pred_dyn <- predict(object = res_dyn,
                    timeData = timeData_pred,
                    fixedData = fixedData_pred,
                    idVar = "id", timeVar = "time", t0 = 5)
```

`predict()` function provides several results for the new subjects. We can extract from its returning object the individual predictions using the following code :

```
# Get individual predictions
head(pred_dyn$pred_indiv)

         1          2          3          4          5          6
 4.4336042 -1.2757025  0.7662503  1.5558106  5.3174034  7.5175713
```

For instance, we predicted 4.43 for subject 1, -1.28 for subject 2 and 0.77 for subject 3.

#### IV.2.6.5 Explore the most predictive variables

In this illustration, we want to evaluate if `DynForest` can identify the true predictors (i.e. baseline random-effect of marker1 and slope random-effect of marker2). To do this, we used the minimal depth which allows to understand the random forest at feature level.

This information about the minimal depth can be extracted using `var_depth()` function and can be displayed with `plot_mindepth()` function. For the purpose of this illustration, we displayed the minimal depth in figure IV.21 by predictor and by feature.

```
# Extract tree building information
depth_dyn <- var_depth(res_dyn)

# Plot average minimal depth by predictor
plot_mindepth(var_depth_obj = depth_dyn,
              plot_level = "predictor")

# Plot average minimal depth by feature
plot_mindepth(var_depth_obj = depth_dyn,
              plot_level = "feature")
```

We observe in figure IV.21A that marker2 and marker1 have the lowest minimal depth, as expected. To go further, we also looked into the minimal depth computed on features. We perfectly identified the slope random-effect of marker2 (i.e. marker2.bi1) and the baseline random-effect of marker1 (i.e. marker1.bi0) as predictors in our simulation.

### IV.2.7 Discussion

The `DynForest` R package provides an easy-to-use random forest methodology for predictors that may contain longitudinal variables possibly measured irregularly with error. Note that the method can also be used without any longitudinal predictors such as other random forest packages.

We implemented several statistics to identify the predictive ability of each variable with the VIMP, gVIMP and average minimal depth. For survival outcome, compared to

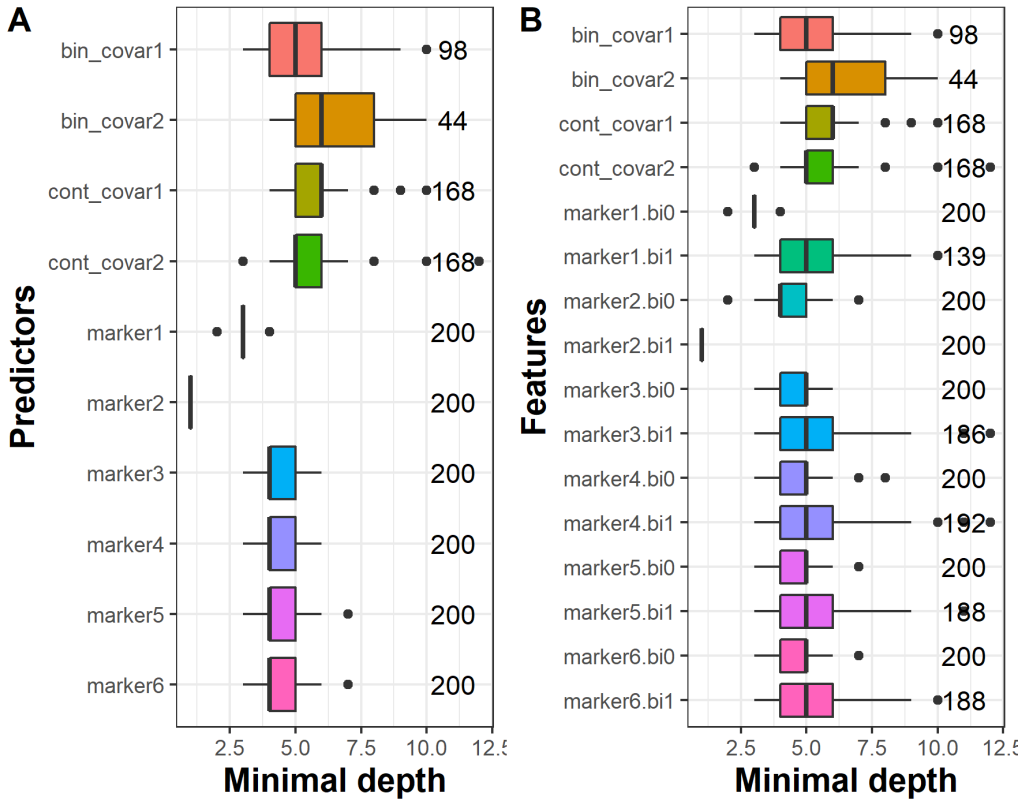



FIGURE IV.21 – Average minimal depth level by predictor (A) and by feature (B).

`randomForestSRC` R package, we considered two different stopping criteria `nodesize` and `minsplit` to favor the deepest forests possible and avoid suboptimal splits. We aimed `DynForest` to be the most user-friendly as possible. To achieve that, we implemented various functions to summarize and display the results. We also included a vignette with a step-by-step procedure on `pbc2` and on a simulated dataset for the different natures of outcomes.

Nevertheless, several improvements could be considered in the future. We used linear mixed models for longitudinal continuous outcomes, but alternative could be considered such as PACE algorithm [Yao et al., 2005] based on functional data analysis. We could also consider different natures of longitudinal predictors (e.g. binary) for which generalized linear mixed models could be used. `DynForest` currently handles continuous, categorical and survival (with possibly competing events) outcomes. But other outcomes could be envisaged such as curves, recurrent events or interval-censored time-to-events. We leave

these perspectives for future.

# IV.3 Prédiction du vasospasme cérébral chez les patients en unité de neuro-réanimation

## IV.3.1 Introduction

L'hémorragie sous-arachnoïdienne (HSA) est un saignement à la surface du cerveau, pouvant être consécutif à un anévrisme ou un traumatisme. Le vasospasme est une complication suite à une HSA, pouvant entraîner une grave détérioration neurologique dans 17% à 40% des patients [Charpentier et al., 1999], et le décès dans les cas les plus graves. Le vasospasme cérébral se définit par le rétrécissement des vaisseaux sanguins du cerveau conduisant à une diminution de l'apport en sang et en oxygène.

Le vasospasme se déclenche en moyenne entre 3 et 14 jours après la survenue de l'HSA [Charpentier et al., 1999], et nécessite une prise en charge rapide. La détection du vasospasme au plus tôt est donc un enjeu majeur pour envisager le meilleur traitement à mettre en place. Pour cela, il est important de comprendre les facteurs de risque consécutifs à la survenue du vasospasme. Dans la littérature, quelques facteurs de risque ont été trouvés dont l'âge, l'échelle de Fisher (quantité de sang au scanner), l'échelle d'évaluation clinique de la *World Federation of Neurological Surgeons* (WFNS) (gravité de l'HSA), l'hyperglycémie ou encore la consommation de tabac [Rabb et al., 1994, Lasner et al., 1997, Harrod et al., 2005, Dupont et al., 2009]. Cependant, ces études ont considéré ces variables uniquement à l'inclusion et à l'aide d'un modèle de régression logistique, et en ne prenant donc pas en compte la temporalité. Dans ce travail, nous souhaitons aller plus loin et tester si les dynamiques de certaines données cliniques peuvent être associées au risque de vasospasme.

L'objectif de ce travail est de : (i) créer un modèle de prédiction de la survenue du vasospasme suite à une HSA, à partir de données cliniques répétées ; (ii) comprendre les mécanismes de survenue du vasospasme.

### IV.3.2 Méthodologie

Pour répondre à ces objectifs, les patients admis dans l'unité de neuro-réanimation du Centre Hospitalier de Bordeaux suite à une HSA entre entre juin 2018 et juin 2019 ont été recrutés, pour un total de 209 patients.

Les patients ont été monitorés durant leur séjour, permettant un recueil automatique des données toutes les heures et jusqu'à un maximum de 14 jours. Les 4 variables longitudinales collectées sont la température corporelle, la pression artérielle moyenne (PAM), la natrémie et la glycémie. En plus de ces variables répétées, 9 variables démographiques et cliniques ont été recueillies à l'inclusion comme l'âge et le sexe du patient, le statut tabagique, les échelles WFNS et de Fisher, la présence d'une hémorragie intra-ventriculaire (HIV), d'un hématome intracérébral (HIP), d'un antécédent d'hypertension artérielle (HTA) et présence d'hydrocéphalie (quantité excessive de liquide céphalo-rachidien).

Les dates et heures exactes de survenue du vasospasme ou du décès ont également été collectées. En revanche, seule la date d'HSA était connue, l'heure a été définie à minuit pour tous les patients. De plus, 4 patients sont décédés avant la survenue d'un vasospasme. Ce nombre n'étant pas suffisant pour considérer le décès sans vasospasme comme un évènement compétitif, nous les avons comptabilisés comme sortie d'étude aux dates de dernières nouvelles car n'étant plus à risque de vasospasme. Le temps de l'évènement d'intérêt a été défini par l'intervalle entre la date de début de l'HSA et la date des dernières nouvelles (sortie d'étude ou vasospasme), jusqu'au temps de censure à droite défini à 20 jours.

Nous avons choisi d'utiliser la méthode `DynForest`, basée sur les forêts aléatoires en survie, pour analyser la survenue du vasospasme à partir de multiples données socio-démographique et cliniques. Cependant, telle que décrite en section IV.1, la méthode modélise les données longitudinales par des modèles mixtes classiques. Bien qu'ils soient très performants, ces modèles ne sont pas les plus adaptés pour prendre en compte un très grand nombre de mesures répétées. En effet, le nombre possible de mesures collectées pour un même patient est de 336 au maximum. Pour cela, les données longitudinales

ont été agrégées toutes les 12 heures à partir de la moyenne des mesures disponibles. La moyenne des temps sur les 12 heures a été considérée comme nouveau temps de mesure donnant 22 temps de mesure en moyenne. De plus, outre le niveau de PAM, sa variabilité (PAM_sd) peut aussi être très pertinente [Faust et al., 2014]. Nous avons créé une variable de variabilité à partir de l'écart des mesures sur les 12 heures consécutives. Les variables longitudinales ont été modélisées à l'aide de *splines* avec un noeud interne situé à 7,2 jours, pour prendre en compte la possible non-linéarité des trajectoires.

La méthodologie `DynForest` nécessite au moins une mesure par sujet pour l'ensemble des variables. Ainsi, à partir des 209 patients de l'étude, 16 patients ont été exclus conduisant à un total de 193 patients avec 14 variables explicatives (5 répétées et 9 à l'inclusion). L'utilisation de `DynForest` nécessite de fixer les différents hyperparamètres $mtry$, $nodesize$ et $minsplit$. Ainsi, nous avons fixé $nodesize = 1$ et $minsplit = 2$ pour avoir des arbres très profonds. L'hyperparamètre $mtry$ a été optimisé suivant toutes les valeurs possibles, de 1 à 14 (nombre de prédicteurs), à partir du critère de l'*Integrated Brier Score* (IBS) entre 4 et 10 jours sur l'ensemble des patients. La valeur optimale a été trouvée pour $mtry = 1$, soit un tirage complètement aléatoire des variables à chaque noeud de l'arbre.

Les probabilités de survenue du vasospasme cérébrale ont été estimées sur un horizon $w = 1, 2, 4$ jours pour les patients encore à risque aux temps *landmark* $s$ à 4, 6 et 8 jours. Ces prédictions ont été réalisées à l'aide d'une validation croisée 10-blocs, pour éviter un sur-apprentissage des données. Enfin, le pouvoir prédictif a été évalué en utilisant le Brier Score (BS) et l'Area Under the ROC Curve (AUC), et répété 50 fois pour estimer la variabilité des résultats.

### IV.3.3 Résultats

Parmi les 193 patients, le vasospasme est apparu pour 50 d'entre eux en moyenne à 8 jours après l'HSA. Les caractéristiques des variables à l'inclusion sont présentées dans le tableau IV.10 pour les 193 patients. En comparant ces caractéristiques en fonction de la survenue du vasospasme, nous observons que la proportion de patients ayant eu un

vasospasme durant le suivi de l'étude sont plutôt des fumeurs (40% contre 29%), ont un score WFNS plus élevé (40% contre 23% pour les scores 4 et 5), ont un score de Fisher plus important (86% contre 60% pour le grade 4), sont atteints plus fréquemment d'hémorragie intra-ventriculaire (72% contre 46%) et d'hydrocéphalie (68% contre 26%). En revanche, les patients sont équitablement répartis selon l'âge (56 ans en moyenne dans les deux groupes), la proportion d'homme (42% contre 46%), la présence d'hématome intracérébral (28% contre 24%) et l'antécédent d'hypertension artérielle (26% contre 31%).

TABLE IV.10 – Description des variables mesurées à l'inclusion en fonction de la survenue du vasospasme pour les 193 patients. Les variables qualitatives sont renseignées par l'effectif et le pourcentage alors que les variables quantitatives sont indiquées par la moyenne et l'écart-type.

| **Variables à l'inclusion** | **Vasospasme** (n = 50) | **Pas de vasospasme** (n = 143) |
|---|---|---|
| Âge en année (moyenne, écart-type) | 56 (10) | 56 (15) |
| Hommes | 21 (42) | 66 (46) |
| Fumeur | 20 (40) | 41 (29) |
| Échelle WFNS | | |
| 1 | 13 (26) | 91 (64) |
| 2 | 11 (22) | 10 (7) |
| 3 | 2 (4) | 10 (7) |
| 4 | 10 (20) | 11 (8) |
| 5 | 14 (28) | 21 (15) |
| Échelle de Fisher | | |
| 1 | 0 (0) | 7 (5) |
| 2 | 3 (6) | 22 (15) |
| 3 | 4 (8) | 28 (20) |
| 4 | 43 (86) | 86 (60) |
| Hémorragie intra-ventriculaire | 36 (72) | 66 (46) |
| Hématome intracérébral | 14 (28) | 35 (24) |
| Antécédent d'hypertension artérielle | 13 (26) | 44 (31) |
| Hydrocéphalie | 34 (68) | 37 (26) |

Durant le suivi de l'étude, 22 mesures ont été collectées en moyenne par patient pour les variables de PAM, PAM_sd et température corporelle, contre 14 mesures en moyenne pour la glycémie et la natrémie. Les trajectoires individuelles des variables longitudinales sont représentées sur la figure IV.22, montrant une légère augmentation de la PAM et de la température pendant quelques jours après la survenue de l'HSA.

Les moyennes et écarts-types du BS et de l'AUC sont présentés dans le tableau IV.11 après 50 réplications. La méthode `DynForest` a été comparée à un modèle de Cox [Cox,

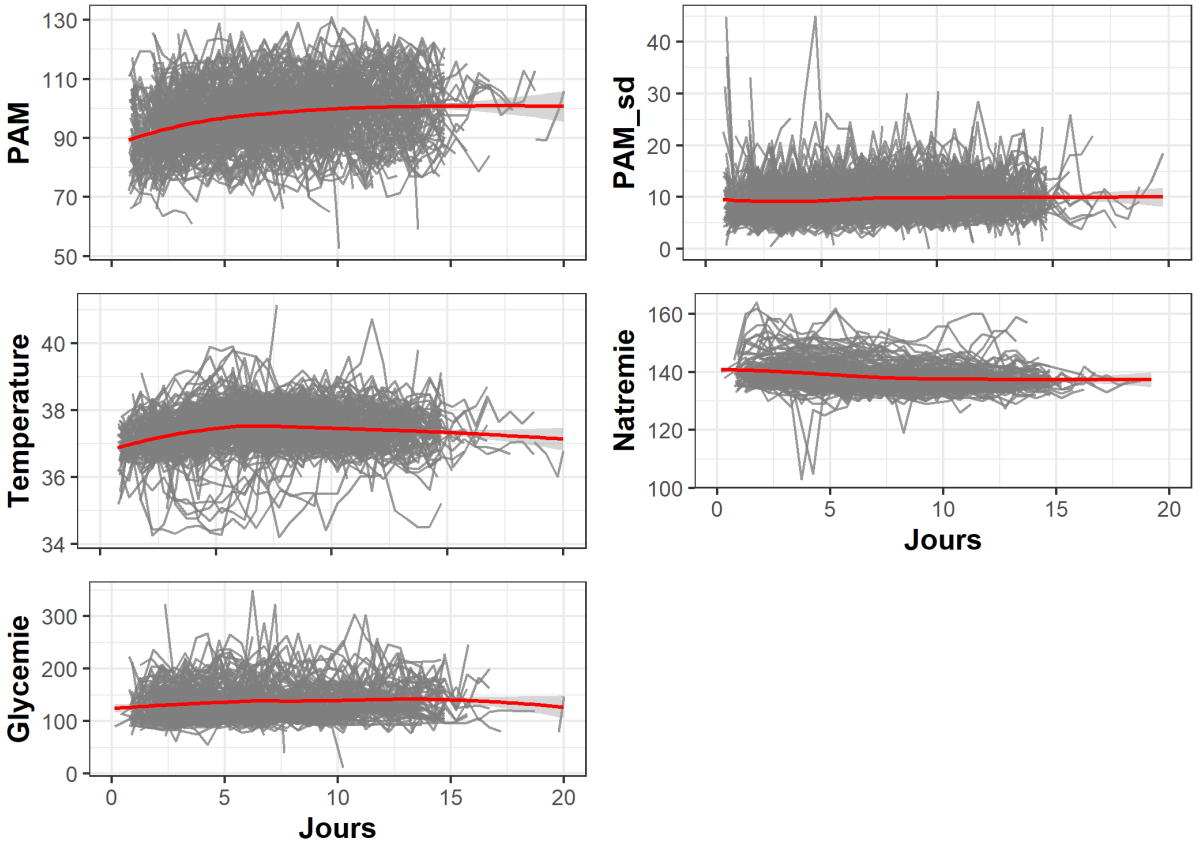



FIGURE IV.22 – Trajectoires individuelles des variables longitudinales à partir des 193 patients. La ligne rouge indique la trajectoire moyenne lissée par la méthode *loess*.

1972] et à un modèle conjoint [Rizopoulos, 2016] (`JMbayes`) pour comparer le gain à prendre en compte les données répétées et à modéliser plus de prédicteurs. Dans les modèles conjoints, seules la PAM et sa variabilité ont été considérées comme variables longitudinales, en plus des autres variables mesurées à l'inclusion. Les variables longitudinales ont été modélisées avec la même spécification que dans `DynForest`. La valeur courante et la pente courante de la PAM et de sa variabilité ont été choisies pour modéliser le risque de survenue du vasospasme. Dans le modèle de Cox, toutes les variables ont été mesurées à l'inclusion et ajoutées additivement pour constituer l'association avec le risque de survenue du vasospasme.

Les résultats montrent des performances équivalentes en terme de BS pour les différents temps *landmark*. En revanche, les résultats sont plus hétérogènes concernant l'AUC. Après 4 jours, la méthode `DynForest` permet de mieux discriminer les patients que les modèles

conjoints ou Cox, avec des AUC de 0,707 et 0,740 à respectivement $w = 1, 2$. Cependant, après 6 et 8 jours, les performances prédictives de `DynForest` sont très faibles avec des AUC entre 0,533 et 0,630, alors que les autres méthodes sont bien plus performantes, avec des scores entre 0,623 et 0,804.

TABLE IV.11 – Performance prédictives des méthodes `DynForest`, `JMbayes` et Cox. Les performances ont été évaluées 50 fois à l'aide du Brier Score (BS) et de l'aire sous la courbe ROC (AUC) après validation croisée 10-blocs pour un temps d'horizon $w = 1, 2$ à partir des données jusqu'au temps landmark $s = 4, 6, 8$.

| | **BS** | | **AUC** | |
|---|---|---|---|---|
| **Landmark time** $s$ | $w = 1$ | $w = 2$ | $w = 1$ | $w = 2$ |
| $s = 4$ | | | | |
| **DynForest** | 0,022 (0,000) | 0,050 (0,001) | 0,707 (0,093) | 0,740 (0,031) |
| **JMbayes** | 0,023 (0,000) | 0,050 (0,001) | 0,623 (0,052) | 0,682 (0,022) |
| **Cox** | 0,022 (0,000) | 0,049 (0,001) | 0,591 (0,071) | 0,634 (0,033) |
| $s = 6$ | | | | |
| **DynForest** | 0,023 (0,000) | 0,049 (0,001) | 0,534 (0,113) | 0,533 (0,049) |
| **JMbayes** | 0,023 (0,000) | 0,048 (0,001) | 0,623 (0,037) | 0,673 (0,024) |
| **Cox** | 0,022 (0,000) | 0,048 (0,001) | 0,630 (0,041) | 0,672 (0,032) |
| $s = 8$ | | | | |
| **DynForest** | 0,032 (0,001) | 0,054 (0,001) | 0,591 (0,085) | 0,630 (0,045) |
| **JMbayes** | 0,029 (0,001) | 0,047 (0,001) | 0,801 (0,030) | 0,812 (0,021) |
| **Cox** | 0,030 (0,001) | 0,049 (0,001) | 0,804 (0,032) | 0,784 (0,029) |

En plus des performances, l'importance des variables (VIMP) a été calculée pour déterminer les variables les plus prédictives à partir de `DynForest`. Le calcul de la VIMP nécessite d'effectuer des permutations dans les données. Pour diminuer la variabilité des résultats, la VIMP moyenne a été calculée à l'aide de 10 réplications. Les résultats en figure IV.23 indiquent sur les variables les plus prédictives sont l'hypertension artérielle, l'âge, le score WFNS et l'hydrocéphalie. Dans une moindre mesure, la glycémie, la consommation de tabac et la variabilité de la PAM sont également prédictrices de la survenue du vasospasme.

### IV.3.4 Discussion

Dans ce travail, nous avons analysé la survenue du vasospasme en tant que données de survie avec censure à droite. A l'aide de la méthode `DynForest`, nous avons pu modéliser le risque de vasospasme à partir d'un ensemble de variables cliniques répétées, ce qui n'a

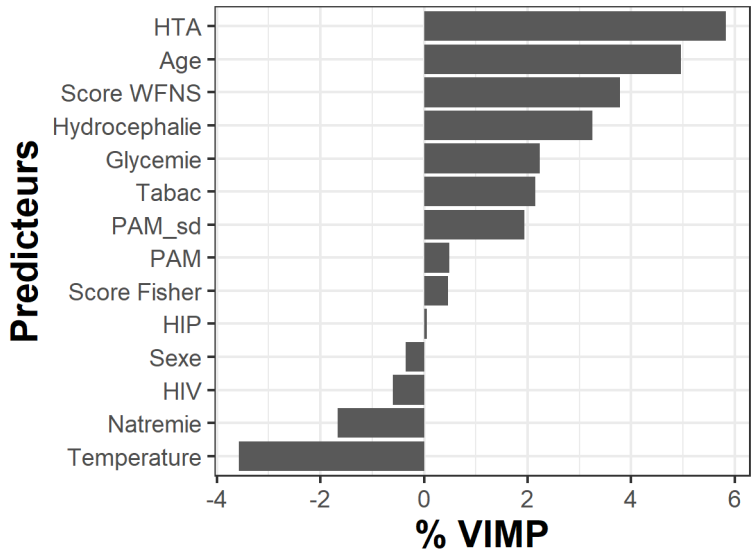



FIGURE IV.23 – Importance des variables (VIMP) moyennées sur 10 réplications. Plus la VIMP est élevé, plus la variable est prédictrice du vasospasme. A l'inverse, les variables avec une VIMP nulle ou négative ne permettent pas d'améliorer les prédictions.

jamais été réalisé auparavant. La qualité de prédiction obtenue par cette méthode est très variable selon le temps *landmark* considéré. Cette qualité est notamment bonne quand les données sont considérées jusqu'à un temps *landmark* à 4 jours, mais deviennent plus faible pour des temps de 6 et 8 jours.

La performance de `DynForest` a été comparée aux modèles conjoints et de Cox. Les résultats montrent que l'approche `DynForest` est plus performante pour un temps *landmark* à 4 jours, mais moins performante pour des temps *landmark* à 6 et 8 jours. De plus, les résultats suggèrent très peu de différence entre les modèles conjoints et le modèle de Cox, pouvant sous-entendre que l'apport de données longitudinales est limité dans ce travail.

A partir des VIMP de la méthode `DynForest`, plusieurs variables prédictrices attendues ont été retrouvées dont l'âge, le score WFNS, la glycémie et la consommation de tabac, confirmant les autres études de la littérature. Néanmoins, la VIMP ne permet pas de confirmer le sens de l'association.

Après optimisation de la méthode `DynForest`, l'hyperparamètre *mtry* a été fixé à 1. Ce paramétrage est plutôt atypique puisque les variables sont choisies totalement aléa-

toirement pour construire la forêt. Pour cela, nous avons appliqué `DynForest` en fixant l'hyperparamètre *mtry* à 7 et 14, correspondant respectivement à un niveau d'aléatoire modéré et quasi-nul. Les résultats ont montré des performances légèrement différentes selon le temps *landmark*, suggérant que l'optimisation du *mtry* est difficile lorsque de nombreux temps de prédiction sont considérés.

Pour aller plus loin dans ce travail, la modélisation des données de survie peut être améliorée. En effet, les patients n'entrent pas dans l'étude dès la survenue de l'HSA, et cette entrée retardée des patients en unité de neuro-réanimation n'a pas été prise en compte dans le test de partitionnement des individus dans les forêts aléatoires. Malgré beaucoup de données répétées, nous avons limité la forme de l'évolution à des *splines* à un noeud interne. Il serait intéressant d'aller explorer des formes d'évolutions plus souples ou des techniques autre que les modèles mixtes, comme l'analyse en données fonctionnelles [Yao et al., 2005].

Enfin, pour envisager un jour de diffuser en routine une application pour la prédiction du vasospasme après HSA, il est important d'avoir de meilleurs performances prédictives aux différents temps *landmark*. Une des pistes possibles est d'avoir un plus grand nombre de patients à inclure dans l'étude, avec un recueil de variables plus important, et en particulier les variables longitudinales.

---

## CONCLUSION

Dans ce chapitre, nous avons proposé une nouvelle approche pour modéliser des évènements de santé, possiblement en compétition, en prenant en compte de multiples variables longitudinales au travers de forêts aléatoires en survie. Cette méthodologie a été implémentée dans le package R `DynForest`, et diffusée sur le site du CRAN à l'aide d'une vignette pour faciliter son utilisation.

A travers différentes applications, nous avons pu voir comment cette approche peut être utilisée pour fournir des prédictions individuelles à différents temps *landmark* à partir d'une unique estimation de la méthode. De plus, à l'aide de plusieurs statistiques comme l'importance des variables et la profondeur minimum, nous pouvons explorer les variables les plus prédictives.

En plus des données de survie, notre approche a déjà été étendue

pour prédire également des variables de natures catégorielles et continues. Néanmoins, de nombreuses perspectives sont encore possibles et sont présentées dans le prochain chapitre.

# Chapitre V

# Discussion

## Sommaire

## V.1 Résumé des travaux de thèse

Dans ce travail de thèse, nous nous sommes intéressés à la prédiction individuelle d'évènements de santé à partir de multiples prédicteurs longitudinaux. Nous avons décidé d'axer notre travail de recherche autour des méthodes d'apprentissage pour l'analyse des données de survie. En particulier, comment ces méthodes peuvent être combinées avec des modèles biostatistiques pour données longitudinales. Pour répondre à cet objectif, deux nouvelles méthodes ont été présentées dans cette thèse :

1. dans une approche *landmark*, les variables longitudinales sont modélisées par des modèles mixtes pour calculer des résumés expliquant au mieux les dynamiques de ces variables. Ces résumés sont ensuite utilisés en tant que prédicteurs dans diverses méthodes de survie adaptées à la grande dimension, pour *in fine* estimer le risque de survenue de l'évènement. Les méthodes de survie peuvent également être combinées pour former un *superlearner* qui permet d'adapter le poids de chaque méthodologie en fonction de ses performances à prédire l'évènement.
2. par une forêt aléatoire pour données de survie, dans laquelle les données longitudinales de prédicteurs sont modélisées tout au long de la construction des différents arbres composant la forêt aléatoire pour tenir compte à la fois des mesures à temps irréguliers et avec erreur des prédicteurs ainsi que la troncature des données répétées induite par la survenue de l'événement.

Ces deux nouvelles méthodes ont été validées dans le cadre de simulations en moyenne dimension, mais également en petite dimension pour s'assurer que ces méthodes ne soient pas moins performantes que celles de référence. Nous avons également appliqué ces méthodes dans diverses illustrations pour prédire notamment la survenue de démence, de vasospasme cérébral après une hémorragie sous-arachnoïdienne ou encore le décès en population âgée générale.

Bien que ces méthodes soient différentes, elles partagent le même objectif à savoir prédire pour de nouveaux individus le risque de survenue d'un évènement à partir de multiples

données longitudinales. Ces méthodes peuvent être utilisées dans plusieurs contextes. Par exemple, lors de l'étude de la survenue d'un évènement à partir d'un temps précis comme dans la prédiction à partir de 85 ans, l'approche *landmark* est particulièrement adaptée. En effet, l'introduction du temps *landmark* permet de se placer dans le cadre idéal par l'utilisation des données collectées jusqu'à ce temps pour les individus encore à risque à ce temps. L'approche *landmark* permet aussi de construire des outils de prédiction différents à chaque temps de prédiction, ce qui s'est révélé pertinent dans le cadre de l'application de la survenue du décès en population générale à l'aide des prédicteurs principaux qui évoluent avec le vieillissement. A l'inverse, l'approche par forêt aléatoire en survie fournit une approche plus générale pour prédire un événement en fonction de prédicteurs répétés. Les atouts majeurs de cette méthode sont : (i) elle nécessite d'être estimée une seule fois pour prédire le risque d'événement à tout temps ; (ii) elle permet de prendre en compte l'intégralité de l'information disponible. L'approche par forêt aléatoire en survie s'applique donc facilement lorsque la prédiction est à effectuer à partir de temps de prédiction non identifiés à l'avance, par exemple après une nouvelle mesure de marqueur dans le suivi d'une maladie chronique.

Dans un but de partage et de reproductibilité de la recherche, les méthodes proposées dans cette thèse ont été développées dans des packages R et diffusées librement. Ainsi, l'approche *landmark* est disponible sur GitHub au nom `hdlandmark`. La version stable de l'approche par forêt aléatoire en survie est téléchargeable depuis le CRAN sous le nom `DynForest` et sa version de développement est disponible sur GitHub.

## V.2 Perspectives

Les méthodes présentées dans cette thèse fonctionnent dans le cas général mais de nombreuses extensions sont encore possible pour traiter certaines particularités des données longitudinales et de survie.

### V.2.1 Extension en survie

#### V.2.1.1 Censure par intervalle

Lors de la collecte des données de survie, le temps de survenue exact de l'évènement n'est pas toujours connu. En pratique, la seule information disponible peut être qu'il soit survenu entre deux temps (par exemple entre deux visites lors d'une étude de cohorte). Dans ce cas, les temps d'évènements sont dits censurés par intervalle. Par soucis de simplicité, le temps d'évènement est souvent remplacé soit par le temps de visite suivant (la visite de diagnostic par exemple), soit par le temps précédent (la dernière visite connue pour une sortie d'étude) ou encore le milieu de l'intervalle (pour une estimation grossière du temps de survenue de l'évènement). Ainsi, les données peuvent être analysées en les considérant comme censurées à droite. Cependant, lorsque nous cherchons à modéliser correctement le risque, négliger la censure par intervalle avec l'utilisation de modèles pour données censurées à droite peut potentiellement entraîner un biais dans l'estimation des paramètres et une sous-estimation de leurs variances [Lindsey and Ryan, 1998]. Cette censure par intervalle pourrait être prise en compte dans nos approches utilisant les forêts aléatoires en survie. En effet, dans le cas simple d'un seul évènement d'intérêt, la statistique du test du *log-rank* est utilisée pour trouver le partitionnement optimal des individus en deux sous-groupes. Ce test pourrait ainsi être remplacé par un test adapté aux données censurées par intervalle [Sun, 1996, Fay, 1999] notamment disponibles dans le package R `interval` [Fay and Shaw, 2010].

#### V.2.1.2 Modèles multi-états

La figure II.2 introduit les risques compétitifs, définit par la survenue du premier évènement. Néanmoins, l'analyse de plusieurs évènements au cours du temps est possible. Par exemple, dans un modèle *illness-death* (voir figure V.1), les individus peuvent évoluer des stades sain, malade ou mort, et changer de statut (également appelé état) au cours du temps. L'intensité de la transition entre chaque état est définie par une fonction de

risque $\lambda(t)$. Ce type de modèle est appelé modèle multi-états [Andersen et al., 2012].

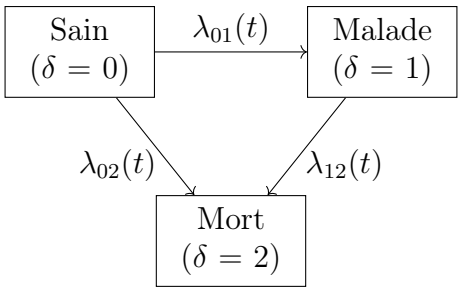



FIGURE V.1 – Illustration des modèles multi-états avec un modèle *illness-death*. Tous les sujets sont considérés comme sain ($\delta = 0$) au début de l'étude et peuvent changer d'état au cours du temps. $\lambda(t)$ défini l'intensité de transition entre les différents états.

En présence de censure à droite, un modèle multi-états peut être obtenu en combinant des modèles de Cox cause-spécifique pour estimer chacune des transitions. Ce principe peut alors être appliqué à notre approche *landmark* pour multiples données longitudinales. Dans ce cas, les modèles de Cox cause-spécifique peuvent être estimés par vraisemblance pénalisée pour prendre en compte le grand nombre de predicteurs et la possible corrélation entre eux. En revanche, cette extension ne peut être appliquée en présence de censure par intervalle. Il serait alors nécessaire d'utiliser par exemple, un modèle multi-états régularisé pour données censurées par intervalle.

#### V.2.1.3 Analyse d'évènements récurrents

Dans certains cas, l'évènement d'intérêt peut se produire plusieurs fois au cours du temps. Plus généralement appelés évènements récurrents, ils se retrouvent fréquemment en santé, par exemple avec la réhospitalisation ou la rechute de cancer. Les évènements récurrents peuvent être analysés par des modèles à fragilité [Rondeau et al., 2007]. Les modèles à fragilité sont des modèles de survie où un effet aléatoire est introduit pour prendre en compte la répétition des évènements lors de la modélisation du risque.

Ces modèles ont également été étendus pour prendre en compte un évènement terminal (comme le décès) en plus des évènements récurrents (voir figure V.2). Le modèle conjoint pour les fonctions de risque des évènements récurrents $r_i$ et de l'évènement terminal $\lambda_i$

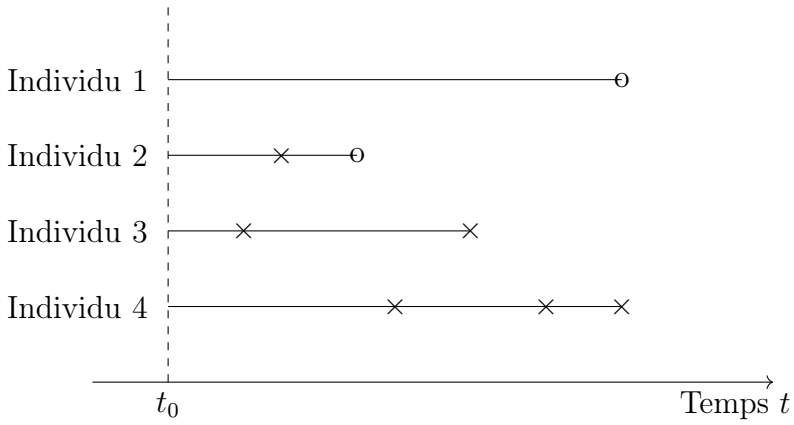



FIGURE V.2 – Illustration des évènements récurrents. Chaque ligne horizontale représente la durée de suivi pour un individu $i$. Les croix indiquent la récurrence des évènements jusqu'à la survenue d'un évènement terminal (avec un rond blanc) ou une censure.

[Rondeau et al., 2007] se défini par :

$$\begin{cases} r_i(t|\omega_i) = \omega_i r_0(t) \exp\left(\beta_1 Z_i(t)\right) \\ \lambda_i(t|\omega_i) = \omega_i^{\alpha} \lambda_0(t) \exp\left(\beta_2 Z_i(t)\right) \end{cases} \tag{V.1}$$

où $r_0(t)$ et $\lambda_0(t)$ sont les risques de base pour les évènements récurrents et l'évènement terminal, respectivement. $\beta_1$ et $\beta_2$ sont les coefficients associés aux variables explicatives $Z_i(t)$. $\omega_i$ est un effet aléatoire partagé pour prendre en compte la corrélation individuelle entre les évènements récurrents et l'évènement terminal. Le paramètre $\alpha$ permet de définir la force de l'association entre les deux risques, avec $\alpha = 0$ pour des risques indépendants. Ce modèle peut être estimé par vraisemblance pénalisée dans le package `frailtypack` [Rondeau et al., 2012].

Dans plusieurs champs de la santé, il est important de prendre en compte l'évolution d'un marqueur pour estimer les risques d'évènements récurrents et d'évènement terminal, par exemple en cancérologie où l'évolution du volume tumorale peut expliquer le risque de rechute de cancer ou de décès. Pour cela, Król *et al.* [Król et al., 2016] ont proposé un modèle conjoint où une variable longitudinale est modélisée en plus des risques définis dans l'équation V.1. La variable longitudinale et les risques d'évènements sont associés au travers d'un effet aléatoire partagé comme présenté en section II.3.3.

Pour prendre en compte de multiples données longitudinales dans ce contexte, l'approche `DynForest` par forêt aléatoire peut potentiellement être utilisée. Nous pourrions tout d'abord modéliser les variables longitudinales à chaque noeud lors de la construction des arbres, de la même façon que présenté dans le chapitre IV. Ensuite, nous pourrions envisager de modéliser conjointement les risques suivant l'équation V.1 en utilisant les effets aléatoires des variables longitudinales comme variables explicatives. La principale difficulté réside dans la façon de trouver les groupes optimaux pour répartir les individus. En effet, le partitionnement optimal des individus est obtenu à partir de la maximisation d'une statistique de test. Dans un futur travail, il serait donc intéressant de définir quel test est le plus adapté pour comparer ces risques à partir des variables candidates pour partitionner les individus.

## V.2.2 Modélisation des données longitudinales

### V.2.2.1 Approche par analyse en données fonctionnelles pour variables continues

Dans les méthodes que nous avons développées, nous nous sommes focalisés sur les variables longitudinales continues, en particulier lorsqu'elles sont distribuées selon une loi normale. Pour cela, nous avons décidé de les modéliser par des modèles linéaires mixtes. Cependant, il existe d'autres techniques pour modéliser ce type de données. L'analyse en composante principale fonctionnelle (en anglais FPCA pour *Functional Principal Component Analysis*) [Dauxois et al., 1982] est une méthode alternative pour l'analyse des données longitudinales continues. Cette méthode traite les données sous un nouvel angle où elles sont analysées comme des fonctions stochastiques au cours du temps. La fonction du temps $Y_{ik}(t)$ pour un individu $i$ et une variable $k$ peut se définir, suivant la transformation de Karhunen-Loève [Karhunen, 1946, Loève, 1946], par :

$$Y_{ik}(t) = \mu_k(t) + \sum_{l=1}^{\infty} \xi_{ikl}\phi_{kl}(t) \tag{V.2}$$

En d'autres termes, la fonction $Y_{ik}(t)$ se divise comme la somme d'une fonction moyenne $\mu_k(t)$ et d'une somme infinie de fonctions temporelles propres orthonormales (appelées aussi composantes) $\phi_{kl}(t)$ décrivant les changements de trajectoires au cours du temps $t$. $\xi_{ikl}$ est un score individuel associé à chaque fonction $\phi_{kl}(t)$ indiquant pour chaque individu l'intensité de changement de trajectoire. Un pourcentage de la variance est expliqué par chacune des fonctions $\phi_{kl}(t)$, ainsi le nombre de fonctions est choisi à partir du pourcentage de variance expliquée par celles-ci. Cette technique permet de réduire la dimension suivant le même principe qu'en analyse en composantes principales [Jolliffe, 2005].

Le processus défini en équation V.2 peut être estimé par l'algorithme PACE (pour *Principal Analysis by Conditional Estimation*) [Yao et al., 2005]. Cet algorithme permet de prendre en compte les erreurs de mesure. Néanmoins, PACE fonctionne uniquement lorsque les données sont mesurées aux même temps pour tous les sujets, et donc, ne peut être utilisé dans les applications où les données sont collectées à des temps spécifiques pour chacun des sujets. En revanche, les données manquantes ne sont pas prises en compte dans cette méthode et nécessitent d'être imputées au préalable.

Dans certains cas où les données sont mesurées très fréquemment, comme les données issues de la collecte électronique, cet algorithme est également parfaitement adapté. Il pourrait être utilisé dans nos méthodes à la place des modèles linéaires mixtes. En effet, les résumés calculés issues des modèles mixtes peuvent être facilement remplacés par les scores individuels associés à chaque fonction propre. De plus, la modélisation par PACE est très rapide et est, par conséquent, une alternative intéressante à mettre en place dans la méthode par forêt aléatoire pour diminuer les temps de calculs.

#### V.2.2.2 Nature des variables

Dans l'approche par forêt aléatoire pour données de survie, nous nous sommes concentrés sur la modélisation des données longitudinales continues. En effet, ces données sont les plus présentes dans le domaine de la santé (en particulier avec les données cliniques ou de laboratoire). Cependant, dans la suite de ce travail, il sera également nécessaire de

pouvoir prendre en compte d'autres types de données longitudinales comme les variables catégorielles (comme les échelles de gravité) ou de comptage (comme la consommation de médicaments). Pour cela, nous pourrions utiliser les modèles mixtes généralisés (comme décrit en section II.3.2) et choisir une fonction de lien compte tenu de la nature de la variable. Les effets aléatoires pourront être calculés suivant l'équation II.33 pour être utilisés en tant que résumés.

Dans cette approche, nous pouvons prendre en compte une variable réponse de nature continue, catégorielle ou de survie. Pour aller plus loin, nous pourrions également étendre cette méthodologie pour prédire des trajectoires. Pour cela, nous pourrions utiliser une trajectoire moyenne en tant que résumé dans les feuilles, et la distance de Fréchet pour quantifier la distance entre les groupes lors de la séparation des individus à chaque noeud [Capitaine et al., 2019].

### V.2.3 Développement des packages R

Dans cette thèse, nous avons développé deux packages R qui sont d'ores et déjà disponibles auprès de la communauté scientifique.

#### V.2.3.1 `hdlandmark`

Le package R `hdlandmark` pour l'approche *landmark* est disponible sur GitHub. Cependant, une mise à jour du package est nécessaire pour faciliter son utilisation. En effet, une unique fonction est actuellement disponible pour réaliser à la fois les étapes d'estimation et de prédiction. De plus, un travail de documentation est nécessaire comme l'écriture d'une vignette R pour faciliter la prise en main du package. Ces modifications sont essentielles avant une diffusion sur le site du CRAN.

#### V.2.3.2 `DynForest`

Le package R `DynForest` pour l'approche par forêt aléatoire est notre méthode la plus aboutie en terme de développement puisqu'il est disponible sur le CRAN. Bien que cette

méthode soit avancée dans son développement, il serait intéressant de réfléchir aux moyens d'accélérer sa vitesse d'exécution. En effet, la répétition des modèles à chaque noeud et pour chaque arbre de la forêt entraîne un fort impact sur les temps de calcul même si nous avons déjà accéléré la procédure en initialisant les algorithmes d'estimation des modèles mixtes en fonction des estimations précédentes.

Dans notre méthode, nous proposons d'optimiser son pouvoir prédictif en utilisant l'erreur OOB basée sur l'IBS. L'IBS permet d'évaluer à la fois la calibration et la discrimination du modèle, mais dans certaines applications, seule la discrimination est évaluée par l'AUC. Pour prendre en compte ce besoin dans `DynForest`, nous pourrions modifier le critère d'évaluation pour le calcul de l'erreur OOB en utilisant l'AUC intégrée (iAUC) sur les temps de prédictions ou, pour aller plus loin, une combinaison entre IBS et iAUC avec une pondération définie par un nouvel hyper-paramètre.

### V.2.4 Interface dynamique

A partir des prédictions individuelles que nous calculons avec nos méthodes, l'objectif final est de fournir un outil pour les cliniciens pour exploiter ces prédictions. Dans le cas de notre application sur la prédiction du vasospasme cérébrale, l'évaluation de notre modèle a montré de bonnes performances lorsque nous utilisons les données longitudinales jusqu'à 4 jours.

Pour fournir les prédictions individuelles auprès des cliniciens, nous pourrions envisager le développement d'une application (par l'interface R shiny par exemple) dans laquelle il serait possible de saisir les mesures des différentes variables longitudinales (et représenter leurs évolutions par exemple) et obtenir la probabilité individuelle de survenue de vasospasme cérébrale prédite à partir de ces données.

## V.3 Conclusion générale

Au cours de cette thèse, nous avons pu voir que l'inclusion de multiples données longitudinales pour prédire des évènements de santé est un problème statistique et numérique complexe. Cependant, en utilisant des méthodes d'apprentissage automatique couplées à des modèles pour données longitudinales, nous avons proposé deux solutions à notre problème, l'une en utilisant une approche *landmark* et l'autre avec une approche par forêt aléatoire en survie. Ces deux approches ont été implémentées dans des packages R disponibles pour les utilisateurs.

Ces approches ont mis en évidence la façon dont les méthodes d'apprentissage automatique peuvent être utilisées pour résoudre le problème d'analyse de multiples données longitudinales. Cela répond à la demande de plus en plus forte de la communauté de santé avec des données longitudinales désormais disponibles en plus grande quantité grâce à l'informatisation des systèmes de santé. Ces données peuvent provenir de différentes sources, par exemple à l'échelle nationale comme le système national de données de santé (SNDS) ou à une échelle locale avec le monitorage des patients dans les établissements de santé par exemple.

A travers cette thèse, nous offrons des premières solutions pour traiter ce flux de données et ouvrons la voie à une utilisation plus massive des données longitudinales pour prédire individuellement des évènements de santé.

# Bibliographie

[3C Study Group, 2003] 3C Study Group (2003). Vascular factors and risk of dementia : design of the Three-City Study and baseline characteristics of the study population. *Neuroepidemiology*, 22(6) :316–325.

[Aalen, 1976] Aalen, O. (1976). Nonparametric Inference in Connection with Multiple Decrement Models. *Scandinavian Journal of Statistics*, 3(1) :15–27.

[Aalen and Johansen, 1978] Aalen, O. O. and Johansen, S. (1978). An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. *Scandinavian Journal of Statistics*, 5(3) :141–150.

[Agüero-Torres et al., 1998] Agüero-Torres, H., Fratiglioni, L., Guo, Z., Viitanen, M., von Strauss, E., and Winblad, B. (1998). Dementia is the major cause of functional dependence in the elderly : 3-year follow-up data from a population-based study. *American journal of public health*, 88(10) :1452–1456.

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

[American Heart Association, 2013] American Heart Association (2013). Guideline on the assessment of cardiovascular risk : A report of the american college of cardiology. *Circulation*, 32 :313–320.

[Andersen et al., 2012] Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (2012). *Statistical models based on counting processes*. Springer Science & Business Media.

[Andersen and Keiding, 2002] Andersen, P. K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11(2) :91–115.

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

[Bates et al., 2015] Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1) :1–48.

[Bender et al., 2005] Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11) :1713–1723.

[Benton, 1945] Benton, A. L. (1945). A visual retention test for clinical use. *Archives of Neurology & Psychiatry*, 54(3) :212–216.

[Bernard et al., 2009] Bernard, S., Heutte, L., and Adam, S. (2009). Influence of hyper-parameters on random forest accuracy. In *International workshop on multiple classifier systems*, pages 171–180. Springer.

[Birkhead et al., 2015] Birkhead, G. S., Klompas, M., and Shah, N. R. (2015). Uses of electronic health records for public health surveillance to advance public health. *Annual review of public health*, 36 :345–359.

[Blanche et al., 2013] Blanche, P., Dartigues, J.-F., and Jacqmin-Gadda, H. (2013). 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.

[Blanche et al., 2015] Blanche, P., Proust-Lima, C., Loubère, L., Berr, C., Dartigues, J.-F., 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 : Comparing Dynamic Predictive Accuracy of Joint Models. *Biometrics*, 71(1) :102–113.

[Breiman, 2001] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1) :5–32.

[Breslow, 1974] Breslow, N. (1974). Covariance Analysis of Censored Survival Data. *Biometrics*, 30(1) :89–99.

[Brier et al., 1950] Brier, G. W. et al. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1) :1–3.

[Capitaine et al., 2019] Capitaine, L., Bigot, J., Thiébaut, R., and Genuer, R. (2019). Fréchet random forests for metric space valued regression with non euclidean predictors. *arXiv preprint arXiv :1906.01741*.

[Charpentier et al., 1999] Charpentier, C., Audibert, G., Guillemin, F., Civit, T., Ducrocq, X., Bracard, S., Hepner, H., Picard, L., and Laxenaire, M. C. (1999). Multivariate analysis of predictors of cerebral vasospasm occurrence after aneurysmal subarachnoid hemorrhage. *Stroke*, 30(7) :1402–1408.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

[Chun and Keles, 2010] Chun, H. and Keles, S. (2010). 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.

[Cox, 1972] Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society : Series B (Methodological)*, 34(2) :187–202.

[Dauxois et al., 1982] Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function : some applications to statistical inference. *Journal of multivariate analysis*, 12(1) :136–154.

[Devaux, 2022] Devaux, A. (2022). *DynForest : Random Forest with Multivariate Longitudinal Predictors*. R package version : 1.0.0.

[Devaux et al., 2022a] Devaux, A., Genuer, R., Peres, K., and Proust-Lima, C. (2022a). Individual dynamic prediction of clinical endpoint from large dimensional longitudinal biomarker history : a landmark approach. *BMC Medical Research Methodology*, 22(1) :1–14.

[Devaux et al., 2022b] Devaux, A., Helmer, C., Dufouil, C., Genuer, R., and Proust-Lima, C. (2022b). Random survival forests for competing risks with multivariate longitudinal endogenous covariates. *arXiv preprint arXiv :2208.05801*.

[Dupont et al., 2009] Dupont, S. A., Wijdicks, E. F., Manno, E. M., Lanzino, G., and Rabinstein, A. A. (2009). Prediction of angiographic vasospasm after aneurysmal subarachnoid hemorrhage : value of the hijdra sum scoring system. *Neurocritical Care*, 11(2) :172–176.

[Fan and Li, 2001] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456) :1348–1360.

[Faust et al., 2014] Faust, K., Horn, P., Schneider, U. C., and Vajkoczy, P. (2014). Blood pressure changes after aneurysmal subarachnoid hemorrhage and their relationship to cerebral vasospasm and clinical outcome. *Clinical neurology and neurosurgery*, 125 :36–40.

[Fay, 1999] Fay, M. P. (1999). Comparing several score tests for interval censored data. *Statistics in Medicine*, 18(3) :273–285.

[Fay and Shaw, 2010] Fay, M. P. and Shaw, P. A. (2010). Exact and asymptotic weighted logrank tests for interval censored data : the interval r package. *Journal of statistical software*, 36(2).

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

[Fine and Gray, 1999] Fine, J. P. and Gray, R. J. (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, 94(446) :496–509.

[Folstein et al., 1983] Folstein, M. F., Robins, L. N., and Helzer, J. E. (1983). The mini-mental state examination. *Archives of general psychiatry*, 40(7) :812–812.

[Fournier et al., 2019] Fournier, M.-C., Foucher, Y., Blanche, P., Legendre, C., Girerd, S., Ladrière, M., Morelon, E., Buron, F., Rostaing, L., Kamar, N., Mourad, G., Garrigue, V., Couvrat-Desvergnes, G., Giral, M., Dantan, E., and DIVAT Consortium (2019). Dynamic predictions of long-term kidney graft failure : an information tool promoting patient-centred care. *Nephrology Dialysis Transplantation*, 34(11) :1961–1969.

[Fu et al., 2017] Fu, Z., Parikh, C. R., and Zhou, B. (2017). Penalized variable selection in competing risks regression. *Lifetime data analysis*, 23(3) :353–376.

[Fuhrer et al., 2003] Fuhrer, R., Dufouil, C., and Dartigues, J. F. (2003). Exploring sex differences in the relationship between depressive symptoms and dementia incidence : prospective results from the paquid study. *Journal of the American Geriatrics Society*, 51(8) :1055–1063.

[Gerds and Schumacher, 2006] Gerds, T. A. and Schumacher, M. (2006). Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. *Biometrical Journal*, 48(6) :1029–1040.

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

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

[Golmakani and Polley, 2020] Golmakani, M. K. and Polley, E. C. (2020). Super Learner for Survival Data Prediction. *The International Journal of Biostatistics*, 16(2) :20190065. Place : Berlin, Boston Publisher : De Gruyter.

[Gray, 2020] Gray, B. (2020). *cmprsk : Subdistribution Analysis of Competing Risks*. R package version 2.2-10.

[Gray, 1988] Gray, R. J. (1988). A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. *The Annals of Statistics*, 16(3) :1141–1154.

[Gregorutti et al., 2015] Gregorutti, B., Michel, B., and Saint-Pierre, P. (2015). Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis*, 90 :15–35.

[Gregorutti et al., 2017] Gregorutti, B., Michel, B., and Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, 27(3) :659–678.

[Harrigton and Fleming, 1982] Harrigton, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, 69(3) :553–566.

[Harrod et al., 2005] Harrod, C. G., Bendok, B. R., and Batjer, H. H. (2005). Prediction of cerebral vasospasm in patients presenting with aneurysmal subarachnoid hemorrhage : a review. *Neurosurgery*, 56(4) :633–654.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction*, volume 2. Springer, New-York.

[Heagerty and Zheng, 2005] Heagerty, P. J. and Zheng, Y. (2005). Survival Model Predictive Accuracy and ROC Curves. *Biometrics*, 61(1) :92–105.

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

[Hickey et al., 2016] Hickey, G. L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2016). Joint modelling of time-to-event and multivariate longitudinal outcomes : recent developments and issues. *BMC medical research methodology*, 16(1) :1–15.

[Hickey et al., 2018] Hickey, G. L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2018). joineRML : a joint model and software package for time-to-event and multivariate longitudinal outcomes. *BMC medical research methodology*, 18(1) :1–14.

[Hippisley-Cox et al., 2017] Hippisley-Cox, J., Coupland, C., and Brindle, P. (2017). Development and validation of qrisk3 risk prediction algorithms to estimate future risk of cardiovascular disease : prospective cohort study. *bmj*, 357.

[Houwelingen and Putter, 2012] Houwelingen, J. C. v. and Putter, H. (2012). *Dynamic prediction in clinical survival analysis*. Number 123 in Monographs on statistics and applied probability. CRC Press, Boca Raton.

[Isaacs and Kennie, 1973] Isaacs, B. and Kennie, A. T. (1973). The set test as an aid to the detection of dementia in old people. *The British Journal of Psychiatry*, 123(575) :467–470.

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

[Ishwaran and Kogalur, 2022] Ishwaran, H. and Kogalur, U. (2022). *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*. R package version 3.1.1.

[Ishwaran et al., 2008] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3) :841–860.

[Ishwaran et al., 2011] Ishwaran, H., Kogalur, U. B., Chen, X., and Minn, A. J. (2011). Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining : The ASA Data Science Journal*, 4(1) :115–132.

[Ishwaran et al., 2010] Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J., and Lauer, M. S. (2010). High-Dimensional Variable Selection for Survival Data. *Journal of the American Statistical Association*, 105(489) :205–217.

[Jiang et al., 2021] Jiang, S., Xie, Y., and Colditz, G. A. (2021). 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)*, 70(1) :66–79.

[Jolliffe, 2005] Jolliffe, I. (2005). Principal component analysis. *Encyclopedia of statistics in behavioral science*.

[Kalbfleisch and Prentice, 2011] Kalbfleisch, J. D. and Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.

[Kaplan and Meier, 1958] Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282) :457–481.

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

[Karhunen, 1946] Karhunen, K. (1946). Zur spektraltheorie stochastischer prozesse. *Ann. Acad. Sci. Fennicae, AI*, 34.

[Katz, 1983] Katz, S. (1983). Assessing self-maintenance : activities of daily living, mobility, and instrumental activities of daily living. *Journal of the American Geriatrics Society*.

[Katzman et al., 2018] Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). DeepSurv : personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1) :1–12.

[Keogh et al., 2019] Keogh, R. H., Seaman, S. R., Barrett, J. K., Taylor-Robinson, D., and Szczesniak, R. (2019). Dynamic Prediction of Survival in Cystic Fibrosis : A Landmarking Analysis Using UK Patient Registry Data. *Epidemiology*, 30(1) :29–37.

[Król et al., 2016] Król, A., Ferrer, L., Pignon, J.-P., Proust-Lima, C., Ducreux, M., Bouché, O., Michiels, S., and Rondeau, V. (2016). Joint model for left-censored longitudinal data, recurrent events and terminal event : Predictive abilities of tumor burden for cancer evolution with application to the ffcd 2000–05 trial. *Biometrics*, 72(3) :907–916.

[Kuller et al., 2003] Kuller, L. H., Lopez, O. L., Newman, A., Beauchamp, N. J., Burke, G., Dulberg, C., Fitzpatrick, A., Fried, L., and Haan, M. N. (2003). Risk factors for dementia in the cardiovascular health cognition study. *Neuroepidemiology*, 22(1) :13–22.

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

[Lasner et al., 1997] Lasner, T. M., Weil, R. J., Riina, H. A., King, J. T., Zager, E. L., Raps, E. C., and Flamm, E. S. (1997). Cigarette smoking—induced increase in the risk of symptomatic vasospasm after aneurysmal subarachnoid hemorrhage. *Journal of neurosurgery*, 87(3) :381–384.

[Lawton and Brody, 1969] Lawton, M. P. and Brody, E. M. (1969). Assessment of older people : self-maintaining and instrumental activities of daily living. *The gerontologist*, 9(3_Part_1) :179–186.

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

[Lechevallier-Michel et al., 2004] Lechevallier-Michel, N., Fabrigoule, C., Lafont, S., Letenneur, L., and Dartigues, J.-F. (2004). Normes pour le mmse, le test de rétention visuelle de benton, le set test d'isaacs, le sous-test des codes de la wais et le test de barrage de zazzo chez des sujets âgés de 70 ans et plus : données de la cohorte paquid. *Revue Neurologique*, 160(11) :1059–1070.

[Li and Luo, 2019] Li, K. and Luo, S. (2019). Dynamic prediction of Alzheimer's disease progression using features of multiple longitudinal outcomes and time-to-event data. *Statistics in Medicine*, 38(24) :4804–4818.

[Lin et al., 2002] Lin, H., McCulloch, C. E., and Mayne, S. T. (2002). Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Statistics in Medicine*, 21(16) :2369–2382.

[Lin et al., 2021] Lin, J., Li, K., and Luo, S. (2021). Functional survival forests for multivariate longitudinal outcomes : Dynamic prediction of Alzheimer's disease progression. *Statistical methods in medical research*, 30(1) :99–111.

[Lindsey and Ryan, 1998] Lindsey, J. C. and Ryan, L. M. (1998). Methods for interval-censored data. *Statistics in medicine*, 17(2) :219–238.

[Loève, 1946] Loève, M. (1946). Fonctions aléatoires à décomposition orthogonale exponentielle. *La Revue Scientifique*, 84 :159–162.

[Marquardt, 1963] Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2) :431–441.

[Maziarz et al., 2017] Maziarz, M., Heagerty, P., Cai, T., and Zheng, Y. (2017). 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.

[Mogensen et al., 2012] Mogensen, U. B., Ishwaran, H., and Gerds, T. A. (2012). Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. *Journal of Statistical Software*, 50(11).

[Mortamais et al., 2013] Mortamais, M., Artero, S., and Ritchie, K. (2013). Cerebral white matter hyperintensities in the prediction of cognitive decline and incident dementia. *International Review of Psychiatry*, 25(6) :686–698.

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

[Nelson, 1969] Nelson, W. (1969). Hazard Plotting for Incomplete Failure Data. *Journal of Quality Technology*, 1(1) :27–52.

[Paige et al., 2018] Paige, E., Barrett, J., Stevens, D., Keogh, R. H., Sweeting, M. J., Nazareth, I., Petersen, I., and Wood, A. M. (2018). 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.

[Parast et al., 2019] Parast, L., Mathews, M., and Friedberg, M. W. (2019). Dynamic risk prediction for diabetes using biomarker change measurements. *BMC medical research methodology*, 19(1) :1–12.

[Perel et al., 2012] Perel, P., Prieto-Merino, D., Shakur, H., Clayton, T., Lecky, F., Bouamra, O., Russell, R., Faulkner, M., Steyerberg, E. W., and Roberts, I. (2012). Predicting early death in patients with traumatic bleeding : development and validation of prognostic model. *BMJ*, 345(aug15 1) :e5166–e5166.

[Perperoglou et al., 2019] Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., and Schmid, M. (2019). A review of spline function procedures in R. *BMC medical research methodology*, 19(1) :1–16.

[Peto and Peto, 1972] Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society : Series A (General)*, 135(2) :185–198.

[Philipps et al., 2014] Philipps, V., Amieva, H., Andrieu, S., Dufouil, C., Berr, C., Dartigues, J.-F., Jacqmin-Gadda, H., and Proust-Lima, C. (2014). Normalized Mini-Mental State Examination for Assessing Cognitive Change in Population-Based Brain Aging Studies. *Neuroepidemiology*, 43(1) :15–25.

[Philipson et al., 2020] Philipson, P., Hickey, G. L., Crowther, M. J., and Kolamunnage-Dona, R. (2020). Faster monte carlo estimation of joint models for time-to-event and multivariate longitudinal data. *Computational Statistics & Data Analysis*, 151 :107010.

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

[Prentice et al., 1978] Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The Analysis of Failure Times in the Presence of Competing Risks. *Biometrics*, 34(4) :541–554.

[Probst and Boulesteix, 2017] Probst, P. and Boulesteix, A.-L. (2017). To tune or not to tune the number of trees in random forest. *The Journal of Machine Learning Research*, 18(1) :6673–6690.

[Probst et al., 2019] Probst, P., Wright, M., and Boulesteix, A.-L. (2019). Hyperparameters and Tuning Strategies for Random Forest. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 9(3). arXiv : 1804.03515.

[Proust-Lima et al., 2019] Proust-Lima, C., Philipps, V., Dartigues, J.-F., Bennett, D. A., Glymour, M. M., Jacqmin-Gadda, H., and Samieri, C. (2019). Are latent variable models preferable to composite score approaches when assessing risk factors of change ? Evaluation of type-I error and statistical power in longitudinal cognitive studies. *Statistical Methods in Medical Research*, 28(7) :1942–1957.

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

[Proust-Lima et al., 2014] Proust-Lima, C., Séne, M., Taylor, J. M., and Jacqmin-Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data : A review. *Statistical Methods in Medical Research*, 23(1) :74–90.

[Proust-Lima and Taylor, 2009] Proust-Lima, C. and Taylor, J. M. G. (2009). 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.

[R Core Team, 2019] R Core Team (2019). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Rabb et al., 1994] Rabb, C., Tang, G., Chin, L., and Giannotta, S. (1994). A statistical analysis of factors related to symptomatic cerebral vasospasm. *Acta neurochirurgica*, 127(1) :27–31.

[Radloff, 1977] Radloff, L. S. (1977). The ces-d scale : A self-report depression scale for research in the general population. *Applied psychological measurement*, 1(3) :385–401.

[Reitan and Wolfson, 1985] Reitan, R. M. and Wolfson, D. (1985). *The Halstead-Reitan neuropsychological test battery : Theory and clinical interpretation*, volume 4. Reitan Neuropsychology.

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

[Rizopoulos, 2012] Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data*. Chapman and Hall/CRC, New-York.

[Rizopoulos, 2016] 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).

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

[Rizopoulos et al., 2017] Rizopoulos, D., Molenberghs, G., and Lesaffre, E. M. (2017). Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal*, 59(6) :1261–1276.

[Rizopoulos et al., 2009] Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 71(3) :637–654.

[Rondeau et al., 2012] Rondeau, V., Marzroui, Y., and Gonzalez, J. R. (2012). frailtypack : an r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software*, 47 :1–28.

[Rondeau et al., 2007] Rondeau, V., Mathoulin-Pelissier, S., Jacqmin-Gadda, H., Brouste, V., and Soubeyran, P. (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation : application on cancer events. *Biostatistics*, 8(4) :708–721.

[Rustand et al., 2022] Rustand, D., van Niekerk, J., Krainski, E. T., Rue, H., and Proust-Lima, C. (2022). Fast and flexible inference approach for joint models of multivariate longitudinal and survival data using integrated nested laplace approximations. *arXiv preprint arXiv :2203.06256*.

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

[Sène et al., 2016] Sène, M., Taylor, J. M., Dignam, J. J., Jacqmin-Gadda, H., and Proust-Lima, C. (2016). Individualized dynamic prediction of prostate cancer recurrence with and without the initiation of a second treatment : Development and validation. *Statistical methods in medical research*, 25(6) :2972–2991.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3) :379–423.

[Signorelli et al., 2021] Signorelli, M., Spitali, P., Szigyarto, C. A.-K., Consortium, T. M.-M., and Tsonaka, R. (2021). Penalized regression calibration : A method for the prediction of survival outcomes using complex longitudinal and high-dimensional data. *Statistics in Medicine*, 40(27) :6178–6196.

[Simon et al., 2011] Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, 39(5).

[Stephan et al., 2015] Stephan, B. C., Tzourio, C., Auriacombe, S., Amieva, H., Dufouil, C., Alpérovitch, A., and Kurth, T. (2015). Usefulness of data from magnetic resonance imaging to improve prediction of dementia : population based cohort study. *bmj*, 350.

[Steyerberg et al., 2010] Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010). Assessing the Performance of Prediction Models : A Framework for Traditional and Novel Measures. *Epidemiology*, 21(1) :128–138.

[Sun, 1996] Sun, J. (1996). A non-parametric test for interval-censored failure time data with application to aids studies. *Statistics in medicine*, 15(13) :1387–1395.

[Suresh et al., 2017] Suresh, K., Taylor, J. M., Spratt, D. E., Daignault, S., and Tsodikov, A. (2017). Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model. *Biometrical Journal*, 59(6) :1277–1300.

[Sweeting et al., 2017] Sweeting, M. J., Barrett, J. K., Thompson, S. G., and Wood, A. M. (2017). The use of repeated blood pressure measures for cardiovascular risk predic-

tion : a comparison of statistical models in the ARIC study. *Statistics in Medicine*, 36(28) :4514–4528.

[Sène et al., 2014] Sène, M., Bellera, C. A., and Proust-Lima, C. (2014). 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.

[Tangri et al., 2011] Tangri, N., Stevens, L. A., Griffith, J., Tighiouart, H., Djurdjev, O., Naimark, D., Levin, A., and Levey, A. S. (2011). A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure. *JAMA*, 305(15) :1553–1559.

[Tanner et al., 2021] Tanner, K. T., Sharples, L. D., Daniel, R. M., and Keogh, R. H. (2021). 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)*, 184(1) :3–30.

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

[Terry M. Therneau and Patricia M. Grambsch, 2000] Terry M. Therneau and Patricia M. Grambsch (2000). *Modeling Survival Data : Extending the Cox Model.* Springer, New York.

[Therneau, 2022] Therneau, T. M. (2022). *A Package for Survival Analysis in R.* R package version 3.4-0.

[Tibshirani, 1997] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4) :385–395.

[Tierney et al., 2010] Tierney, M. C., Moineddin, R., and McDowell, I. (2010). Prediction of all-cause dementia using neuropsychological tests within 10 and 5 years of diagnosis in a community-based sample. *Journal of Alzheimer's Disease*, 22(4) :1231–1240.

[Tierney et al., 2005] Tierney, M. C., Yao, C., Kiss, A., and McDowell, I. (2005). Neuropsychological tests accurately predict incident alzheimer disease after 5 and 10 years. *Neurology*, 64(11) :1853–1859.

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

[Tsiatis et al., 1995] Tsiatis, A. A., Degruttola, V., and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American statistical association*, 90(429) :27–37.

[van der Laan et al., 2007] van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

[Van Houwelingen, 2007] Van Houwelingen, H. C. (2007). Dynamic Prediction by Landmarking in Event History Analysis. *Scandinavian Journal of Statistics*, 34(1) :70–85.

[Viswanathan et al., 2009] Viswanathan, A., Rocca, W. A., and Tzourio, C. (2009). Vascular risk factors and dementia : how to move forward ? *Neurology*, 72(4) :368–374.

[Wahba, 1977] Wahba, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy. *SIAM journal on numerical analysis*, 14(4) :651–667.

[Wilson et al., 1998] Wilson, P. W. F., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., and Kannel, W. B. (1998). Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation*, 97(18) :1837–1847.

[Wright and Ziegler, 2017] Wright, M. N. and Ziegler, A. (2017). ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1). arXiv :1508.04409 [stat].

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

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

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

[Yu et al., 2017] Yu, K.-H., Berry, G. J., Rubin, D. L., Ré, C., Altman, R. B., and Snyder, M. (2017). Association of omics features with histopathology patterns in lung adenocarcinoma. *Cell Systems*, 5(6) :620–627.e3.

[Zhang, 2010] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2).

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

[Zheng and Heagerty, 2007] Zheng, Y. and Heagerty, P. J. (2007). Prospective Accuracy for Longitudinal Markers. *Biometrics*, 63(2) :332–341.

[Zimmerman et al., 2006] Zimmerman, J. E., Kramer, A. A., McNair, D. S., and Malila, F. M. (2006). Acute Physiology and Chronic Health Evaluation (APACHE) IV : Hos-

pital mortality assessment for today's critically ill patients. *Critical Care Medicine*, 34(5) :1297–1310.

[Zou, 2006] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476) :1418–1429.

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(2) :301–320.

**Modélisation et prédiction dynamique individuelle d'événements de santé à partir de données longitudinales multivariées**

**Résumé :** En santé publique, la prédiction d'un évènement de santé est un enjeu crucial pour le devenir du patient. A partir de méthodes statistiques, cette prédiction peut être estimée de manière individuelle en utilisant les données propres à chaque patient. Cependant, la plupart des modèles actuels ne permettent pas de prendre en compte un grand nombre d'informations répétées. L'objectif de ce travail de thèse est de développer de nouvelles méthodes statistiques pouvant intégrer un ensemble de prédicteurs collectés au cours du temps pour prédire au mieux un évènement de santé. Dans la première partie, nous proposons une approche *landmark* où des résumés de données longitudinales, calculés au temps *landmark*, sont utilisés pour estimer le risque de survenue de l'évènement à travers plusieurs méthodes adaptées à la grande dimension. Cette méthode a également été étendue dans le cadre de risques compétitifs pour prédire la survenue de la démence pour les individus de la cohorte des trois-cités. Dans la deuxième partie, nous proposons d'intégrer les données répétées de variables dans les forêts aléatoires en survie pour prendre en compte la possible sortie d'étude informative des patients. Cette nouvelle méthodologie a été développée dans un package R `DynForest` disponible pour les utilisateurs. Elle a été appliquée pour (i) prédire la probabilité de survenue de démence à partir des trajectoires de multiples variables mesurant notamment la dépendance fonctionnelle, la cognition, l'atrophie cérébrale et les lésions vasculaires cérébrales (ii) prédire la survenue du vasospasme cérébral chez les patients ayant subi une hémorragie sous-arachnoïdienne. Par ces travaux, nous ouvrons la voie à l'intégration d'un grand nombre de données longitudinales pour prédire le risque de survenue d'évènements.

**Mots clés :** Prédictions dynamiques ; Données longitudinales ; Données de survie ; Grande dimension ; Forêts aléatoires.

**Dynamic modelling and prediction of health events from multivariate longitudinal data**

**Abstract :** In public health, the prediction of health events is a crucial issue for the patient's future. Using statistical methods, predictions can be individually estimated using patient-specific data. However, most of existing models are not able to take into account large number of repeated information. The objective of this thesis is to develop new statistical methods that can include many predictors collected over the time to improve the ability to predict a health event. In the first part, we propose a landmark approach where features of longitudinal data, computed at landmark time, are included as predictors through various methods adapted to high dimension to predict the risk of event. This method was also extended to competing risk to predict the risk of dementia on patients in the three-city cohort. In the second part, we include the longitudinal information through random survival forests to consider the possible dropout information of patients. This novel methodology has been developed in the `DynForest` R package available to users. It was applied to (i) predict the risk of dementia from multiple longitudinal data measuring functional dependency, cognition, cerebral atrophy and cerebrovascular lesions (ii) predict the risk of cerebral vasospasm in patients suffering from subarachnoid hemorrhage. With this work, we pave the way for the integration of a large number of longitudinal information to predict the risk of various health events.

**Keywords :** Dynamic predictions ; Longitudinal data ; Survival data ; High dimension ; Random forest.

**Unité de recherche**
Inserm U1219, *Bordeaux Population Health*, Université de Bordeaux
146, rue Léo Saignat
33000 Bordeaux, France