Journal de la Société Française de Statistique
Vol. 157 No. 2 (2016)

Évaluation de prédictions dynamiques : quelques méthodes et applications au pronostic de la démence

Title: Evaluating dynamic predictions: some methods and applications to dementia prognosis

Paul Blanche[1]

Résumé : L'utilisation de données longitudinales pour calculer des prédictions individuelles de risque dites "dynamiques" gagne actuellement en popularité. Les prédictions sont dites dynamiques car leur calcul est actualisé au fur et à mesure que l'information sur le profil de santé des sujets évolue au cours de leur suivi. Cet article papier présente des méthodes pour quantifier et comparer des capacités pronostiques pour ce type de prédictions. Une évaluation basée sur les deux concepts de discrimination et de calibration est suggérée et une approche non paramétrique de pondération par l'inverse de la probabilité de censure est présentée pour l'inférence. Cette approche permet de s'adapter naturellement à la présence de données censurées et de risques concurrents, deux situations fréquentes en recherche médicale. Quelques résultats asymptotiques sont présentés. Des tests et des régions de confiance en sont dérivés. Une application sur des prédictions du risque de démence chez les personnes âgées est discutée. Les prédictions sont basées sur les mesures répétées de deux tests psychométriques et sont issues de deux modèles précédemment estimés sur les données de la cohorte Paquid. Leurs capacités pronostiques sont quant à elles évaluées et comparées avec les données externes de la cohorte des Trois-Cités.

Abstract: The computation of dynamic predictions, using longitudinal data, has recently become popular. The term dynamic emphasizes that the predictions can be updated when the information on the subjects' health profiles increases with follow-up time. This paper presents methods to quantify and compare prognostic abilities of such dynamic predictions. The methods aim to evaluate the calibration and discrimination performances of the dynamic predictions. We focus on methods which handle censoring and which can be applied in the competing risks setting. Inverse probability of censoring weighting estimators are suggested. Tests and both pointwise and simultaneous confidence intervals are derived from asymptotic results. We illustrate the methods by comparing predictions of dementia in the elderly, accounting for the competing risk of death. Predictions are computed from two models which model the risk of dementia given repeated measures of two psychometric tests. The models are estimated on the Paquid cohort and prediction abilities are evaluated and compared using external data from the Three-City cohort.

Mots-clés : calibration, censure, discrimination, données répétées, modèle conjoint, risques concurrents
Keywords: calibration, censoring, discrimination, repeated measurements, joint-model, competing risks
Classification AMS 2000 : 62P10, 62N02

1. Introduction

De nombreux espoirs de la recherche médicale reposent actuellement sur le développement de traitements préventifs et l'initiation de prises en charge précoces des patients. C'est le cas depuis longtemps déjà en cancérologie et en cardiologie, et c'est maintenant aussi le cas en neurologie, dans le cadre de la lutte contre la maladie d'Alzheimer (Aisen et al., 2011).

[1] LMBA, Université de Bretagne Sud, Campus de Tohannic, 56017 Vannes.
E-mail : paul.blanche@univ-ubs.fr

 ISSN: 2102-6238

Dans ce contexte, la statistique joue un rôle important en contribuant au développement de modèles pronostiques. Les modèles pronostiques permettent d'aider les cliniciens à prendre des décisions cliniques et leur utilisation est souvent recommandée par les sociétés médicales influentes (Camm et al, 2010; Freedman et al., 2005). Pourtant, dans le cadre de la lutte contre la maladie d'Alzheimer, les modèles pronostiques sont encore largement absents, bien qu'ils apparaissent aujourd'hui d'un intérêt certain. À court terme, ils pourraient aider à la sélection de sujets à haut risque de démence pour augmenter la puissance d'essais cliniques randomisés (Aisen et al., 2011). À plus long terme, ils pourraient aider à améliorer la prise en charge de personnes âgées suspectées de décliner vers une démence.

Avant de pouvoir utiliser des modèles pronostiques en pratique clinique, il est essentiel de d'abord évaluer les performances des prédictions qu'ils fournissent. Pour cela, les évaluations des performances en termes de discrimination et de calibration sont parmi les plus importantes (Gail and Pfeiffer, 2005; Vickers and Cronin, 2010; Pepe and Janes, 2013). Brièvement, on dit qu'un modèle est calibré pour le risque de démence si, pour tout pourcentage x, on peut s'attendre à ce que x sujets sur 100 deviennent déments parmi les sujets qui ont un risque prédit par le modèle de $x\%$. On dit aussi qu'un modèle pronostique de la démence a de bonnes capacités de discrimination si (i) l'étendue des valeurs prédites par le modèle est large et (ii) les sujets ayant un risque prédit élevé (respectivement bas) sont effectivement plus à risque de démence (respectivement moins à risque). On considère souvent que la calibration d'un modèle est particulièrement importante pour pouvoir apporter des conseils personnalisés aux patients, tandis que la discrimination est particulièrement importante pour aider au dépistage de sujets ayant des profils de risque élevé.

Dans cet article, on s'intéresse à l'évaluation de la discrimination et de la calibration dans le cadre spécifique de modèles pronostiques dits dynamiques, en présence de données censurées et de risques concurrents. Notre intérêt pour ces trois spécificités est motivé par la nature des données des cohortes Paquid (Dartigues et al., 1992) et Trois-Cités (The 3C Study Group, 2003) que l'on souhaite analyser et qui incluent ces particularités. Ces cohortes sont des cohortes de personnes âgées avec lesquelles on souhaite construire, évaluer et comparer des modèles pronostiques de démence.

Le cadre spécifique dit "dynamique" vient du fait que l'on observe des données répétées de tests cognitifs. On s'intéresse alors à des modèles pronostiques permettant d'actualiser les prédictions au fur et à mesure que l'information sur l'évolution du profil cognitif d'un sujet s'accroît. Par exemple, chaque fois qu'une personne âgée revoit un psychologue et que ses fonctions cognitives sont réévaluées, une nouvelle prédiction de survenue de démence à 5 ans peut être calculée, en tenant compte de toute l'information disponible à ce moment. Dans l'application présentée en Section 5, on s'intéressera particulièrement à des prédictions issues de modèles conjoints pour données longitudinales et de survie (Tsiatis and Davidian, 2004; Rizopoulos, 2012; Taylor et al., 2013; Proust-Lima et al., 2014), bien que d'autres approches soient également envisageables (Zheng and Heagerty, 2005; van Houwelingen, 2007; van Houwelingen and Putter, 2012; Nicolaie et al., 2013).

Au-delà du fait que les données censurées soient fréquentes en biostatistiques (Andersen et al., 1993), on s'intéresse ici spécifiquement à celles-ci car nos données contiennent les observations de sujets perdus de vue, ce qui engendre des problématiques étroitement liées à celles des données manquantes (Tsiatis, 2006). Par exemple, lorsqu'un sujet sain est perdu de vue après trois ans de suivi, on ne peut pas savoir s'il décédera ou s'il développera une démence dans les 2 années

 ISSN: 2102-6238

suivantes. Cette donnée est manquante et cela complexifie l'évaluation de prédictions de démence à cinq ans.

Les risques concurrents sont aussi fréquemment rencontrés en biostatistiques (Andersen et al., 1993). Comme on souhaite évaluer une prédiction de démence, on s'y intéresse ici car on doit tenir compte du risque (non négligeable) de décès sans avoir précédemment connu de démence. Ignorer le risque concurrent de décès, comme c'est parfois fait en considérant le décès sans démence comme une censure, serait une erreur : cela reviendrait à analyser les données, et à interpréter les résultats dans un monde hypothétique et non réaliste où tout le monde devient dément avant de décéder.

Le reste de l'article est organisé comme suit. La section 2 présente les données et les questions de recherche qu'elles soulèvent, qui illustreront nos propos. La section 3 présente les critères avec lesquels les performances des prédictions dynamiques sont évaluées. La section 4 présente les méthodes d'inférence associées et la section 5 présente une application à nos données de cohortes de personnes âgées. La section 6, quant à elle, conclut le manuscrit par quelques remarques.

Ce papier donne suite à la présentation orale de la session du prix du Dr Norbert Marx, à l'occasion des *47èmes Journées de Statistique* organisées par la Société Française de Statistique à Lille, en juin 2015. Ce papier reprend donc de nombreux éléments du papier récompensé (Blanche et al., 2015), bien que le but soit ici de les replacer dans un contexte plus large et didactique.

2. Les cohortes Paquid et Trois-Cités

Paquid (pour "*Personnes âgées Quid ?*") est l'une des premières cohortes européennes visant à étudier le vieillissement cognitif normal et pathologique des personnes âgées. C'est une étude de cohorte prospective qui a débuté en 1988 et qui inclut 3 777 personnes âgées de plus de 65 ans tirées au sort sur des listes électorales de la région Aquitaine (Dartigues et al., 1992).

L'étude des Trois-Cités est une autre étude de cohorte prospective visant à étudier la maladie d'Alzheimer (`www.three-city-study.com`). Lancée en 1999, cette étude inclut 9 294 personnes âgées de plus de 65 ans également recrutées par tirage au sort, mais cette fois sur les listes électorales des villes de Bordeaux, Dijon et Montpellier, d'où le nom Trois-Cités (The 3C Study Group, 2003).

Dans les deux études, les sujets ont été revus approximativement tous les deux ans par des psychologues, après une visite initiale. Une multitude d'informations ont été recueillies au cours de ces visites, dont les scores des sujets à une batterie de tests cognitifs. Ceci a déjà permis d'estimer des dynamiques d'évolution cognitive et de montrer que l'évolution de scores cognitifs pourrait être utilisée pour le dépistage précoce de la maladie d'Alzheimer (Amieva et al., 2008). Les informations relatives aux diagnostics de démence et aux décès ont aussi été recueillies au cours du suivi (qui est toujours en cours).

Les échantillons de Paquid et Trois-Cités ici considérés incluent respectivement les données issues de $n = 2\,970$ et $n = 3\,880$ sujets suivis pendant 20 ans et 10 ans.

2.1. Contexte, motivations et objectifs

Motivé par les potentielles applications à la recherche clinique (Aisen et al., 2011), on s'intéresse ici à utiliser ces données pour calculer, évaluer et comparer des risques d'apparition de démence à

 ISSN: 2102-6238

t= 5 ans. Les calculs des risques sont principalement basés sur l'évolution du score d'une personne âgée à des tests cognitifs, mais sont aussi ajustés sur l'âge, le sexe et le niveau d'éducation.

Les deux tests psychométriques considérés sont le *Mini-Mental State Examination* (MMSE) et le test de fluence verbal d'Isaac (IST). Le MMSE est un test global de cognition évaluant la mémoire, le calcul, l'orientation dans le temps et l'espace et le langage. C'est probablement l'un des tests psychométriques parmi les plus renommés et c'est le plus couramment utilisé par les neurologues. L'IST, quant à lui, évalue la fluidité et la vitesse de production verbale. Ce test consiste essentiellement à demander à un sujet d'énoncer le plus de mots possibles appartenant à des catégories sémantiques spécifiques en un temps limité (e.g. noms de fruits, d'animaux ou de villes en une minute). Curieusement, de précédents travaux suggèrent que (i) l'IST serait relativement sensible aux petites variations de la cognition (Proust-Lima et al., 2007) et que (ii) ce serait l'un des tests cognitifs pour lesquels l'évolution des scores des sujets évoluant progressivement vers une démence se différencierait le plus précocement d'une évolution normale (Amieva et al., 2008).

Dans ce contexte, quelques questions naturelles auxquelles les analyses de la section 5 tenteront d'apporter des éléments de réponses sont : Quelles sont les capacités pronostiques d'un outil pronostique basé sur une évolution du MMSE ou de l'IST ? Un outil pronostique basé sur le prometteur IST apparaît-il meilleur que celui basé sur le populaire MMSE ?

3. Critères d'évaluation

Pour évaluer des prédictions statistiquement, l'idée principale repose simplement sur la comparaison d'observations et de prédictions. Les principales différences entre les différentes méthodes d'évaluation sont essentiellement héritées des différentes possibilités de les comparer.

3.1. Notations

Dans la suite, on notera (T,η) le temps d'événement et le type d'événement étudiés. Pour nos données, T représente le temps entre le début du suivi d'un sujet et la survenue d'une démence ou d'un décès non consécutif à une démence. Autrement dit, T représente le temps entre le début du suivi et la survenue d'une démence pour un sujet qui devient dément, ou le décès, sinon. La variable $\eta \in \{1,2\}$ indique si T indique le temps jusqu'à la survenue d'une démence $(\eta = 1)$ ou d'un décès $(\eta = 2)$.

Pour un temps de prédiction $s \in \mathscr{S}$ et un horizon de prédiction t (e.g. $t = 5$ ans), on note $D(s,t) = \mathbb{1}\{s < T \leq s+t, \eta = 1\}$ la variable aléatoire indicatrice de l'occurrence de l'événement d'intérêt (e.g. démence) entre les temps s et $s+t$. C'est en raison de l'aspect dynamique des données et des prédictions qu'on s'intéresse à un ensemble de temps de prédiction $\mathscr{S}$, par exemple $\mathscr{S} = \{1,2,3,4\}$ ans. On notera aussi par $\pi(s,t)$ une prédiction de l'occurrence de l'événement entre s et $s+t$, sachant que le sujet est à risque en s et l'information disponible sur sa santé en s, notée $\mathscr{Y}(s)$. Autrement dit, $\pi(s,t)$ est une estimation de $\mathbb{P}\big(D(s,t) = 1 | \mathscr{Y}(s), T > s\big)$. Pour nos données, $\mathscr{Y}(s)$ inclut l'âge du sujet, son niveau d'étude, son sexe et toutes les mesures répétées d'un test psychométrique faites avant le temps s, qui mesurent l'évolution cognitive du sujet durant les s premières années de son suivi. Sans perte de généralité, on supposera la prédiction $\pi(s,t)$ à valeur dans $[0,1]$ pour tout s, t.

 ISSN: 2102-6238

3.2. *Concepts*

Le concept de calibration repose sur la comparaison entre la fréquence d'événement parmi les sujets ayant un même risque prédit et la valeur de ces risques prédits. Pour un temps de prédiction s et un horizon de prédiction t, on dira que l'outil (ou modèle) pronostique qui fournit une prédiction du risque à t années $\pi(s,t)$, à partir de l'information $\{\mathscr{Y}(s), T > s\}$, est calibré, si pour tout risque prédit $\pi(s,t) = r$, ce risque correspond réellement au risque d'événement des sujets, i.e.,

$$\forall r \in [0,1] \quad \mathbb{P}\big\{D(s,t) = 1 | \pi(s,t) = r, T > s\big\} = r.$$

Dans ce cas, on dira aussi que les prédictions (issues de cet outil pronostique) sont calibrées en s pour l'horizon t.

Bien que la bonne calibration des prédictions soit nécessaire à leur utilité en pratique, elle n'est cependant souvent pas suffisante. Un modèle peut être à la fois parfaitement calibré et cliniquement complètement inutile. Par exemple, c'est le cas d'un modèle "nul" qui associe le même risque prédit $r_0(s,t) = \mathbb{P}\big\{D(s,t) = 1 | T > s\big\}$ à tout sujet, quelles que soient ses caractéristiques propres. En attribuant ce même risque marginal à tout le monde, ce modèle n'a aucune capacité de discrimination et ne permet pas d'aider à personnaliser des décisions cliniques. Ce modèle est pourtant parfaitement calibré.

Le concept de discrimination repose essentiellement sur la comparaison entre les prédictions des sujets pour lesquels on observe un événement, qu'on appellera des *cas*, et celles des sujets pour lesquels on n'observe pas d'événement, qu'on appellera des *contrôles*. Brièvement, on dit souvent que les prédictions ont de bonnes capacités de discrimination si les risques prédits des *cas* sont généralement plus grands que ceux des *contrôles*.

Les concepts de calibration et de discrimination sont donc complémentaires. En pratique, il est malheureusement souvent difficile d'améliorer un modèle préexistant en termes de discrimination sans en dégrader la calibration, et vice-versa. En quelque sorte, on retrouve ici le dilemme biais-variance intrinsèque à l'estimation statistique. Ce phénomène est, entre autres, à l'origine du principe dénommé "KISS" de façon amusante, pour "*Keep It Simple Stupid*" (Vickers and Cronin, 2010). Très répandu en biostatistique, il est souvent respectable pour guider les choix de modélisation en pratique, au vu de la richesse des données typiquement disponibles.

3.3. *Critères d'évaluation*

Une évaluation de type graphique de la calibration est souvent recommandée (Pepe and Janes, 2013; Gail and Pfeiffer, 2005). Brièvement, l'idée consiste à diviser un échantillon en différents groupes de sujets ayant des risques prédits similaires et à comparer les risques moyens dans chaque groupe aux risques d'événement observés, calculés par estimation non paramétrique. Des exemples sont présentés aux Figures 1a et 1b et commentés en Section 5.2.

Pour quantifier les capacités de discrimination des prédictions, l'approche la plus populaire est celle des courbes ROC ("*Receiver Operating Characteristic*") (Gail and Pfeiffer, 2005; Gerds et al., 2008; Pepe and Janes, 2013). Elle est basée sur la comparaison des distributions des prédictions des *cas* et des *contrôles*. Pour tout temps de prédiction $s > 0$, horizon de prédiction $t > 0$ et seuil de risque $r \in [0,1]$, on définit la sensibilité (ou proportion de vrais positifs) de l'outil

 ISSN: 2102-6238

pronostique comme

$$\mathrm{Se}_\pi(r;s,t) = \mathbb{P}\big\{\pi(s,t) > r \,|\, D(s,t) = 1\big\},$$

la spécificité (ou proportion de vrais négatifs) comme

$$\mathrm{Sp}_\pi(r;s,t) = \mathbb{P}\big\{\pi(s,t) \leq r \,|\, D(s,t) = 0\big\}.$$

La courbe ROC, définie comme l'ensemble des points

$$\mathrm{ROC}_\pi(s,t) = \Big\{\Big(\mathrm{Se}_\pi(r;s,t), 1-\mathrm{Sp}_\pi(r;s,t)\Big), r \in [0,1]\Big\},$$

représente simultanément les capacités à discriminer pour tous les seuils r possibles. Plus cette courbe croit rapidement du point $(0,0)$ au point $(1,1)$ et meilleures sont les capacités de discrimination des prédictions. L'interprétation plus fine de la courbe et le choix d'un seuil r spécifique au-delà duquel une décision médicale peut être conseillée dépend généralement fortement du contexte (Pepe, 2003). Cependant, un indice résumé toujours intéressant est l'aire sous la courbe ROC, notée AUC ("*Area Under the Curve*"), que l'on peut montrer être égale à la probabilité de concordance

$$\mathrm{AUC}_\pi(s,t) = \mathbb{P}\Big(\pi_i(s,t) < \pi_j(s,t) \,\Big|\, D_i(s,t) = 0, D_j(s,t) = 1, T_i > s, T_j > s\Big),$$

où i et j représentent les indices de deux sujets indépendants (Pepe, 2003). L'AUC s'interprète donc simplement comme la probabilité que le risque prédit d'un cas soit supérieur à celui d'un contrôle.

Pour évaluer simultanément la calibration et la discrimination, on peut estimer l'erreur quadratique de prédiction, aussi appelée le "*Brier Score*" (Brier, 1950; Gail and Pfeiffer, 2005; Gerds et al., 2008; Steyerberg, 2009; Pepe and Janes, 2013). Pour $s > 0, t > 0$, il est défini comme

$$\mathrm{BS}_\pi(s,t) = \mathbb{E}\Big[\big\{D(s,t) - \pi(s,t)\big\}^2 \big| T > s\Big]$$

et se décompose naturellement comme

$$\begin{aligned}\mathrm{BS}_\pi(s,t) =& \mathbb{E}\Big[\Big\{\pi(s,t) - \mathbb{E}\Big[D(s,t)\big|\pi(s,t), T > s\Big]\Big\}^2 \,\Big|\, T > s\Big] \\ &+ \mathrm{Var}\Big[D(s,t) \,\Big|\, T > s\Big] - \mathrm{Var}\Big[\mathbb{E}\big\{D(s,t)\big|\pi(s,t), T > s\big\} \,\Big|\, T > s\Big], \end{aligned} \quad (1)$$

comme toute erreur quadratique. Le premier terme (le terme de biais) est d'autant plus faible que la calibration du modèle est bonne, et nul si le modèle est calibré. Le troisième terme (le terme de variance négatif) est quant à lui d'autant plus grand, en valeur absolue, que le risque d'événement conditionnellement à la valeur du risque prédit, i.e. $\mathbb{E}\big[D(s,t)\big|\pi(s,t), T > s\big]$, varie. Ainsi, plus les prédictions ont des valeurs variées, et indiquent des risques d'événements variés (c'est-à-dire plus les capacités de discrimination sont importantes) et plus l'erreur quadratique est petite.

Bien que l'erreur quadratique soit très populaire dans de nombreux domaines de la statistique, par exemple en apprentissage automatique ("*machine learning*") (Hastie et al., 2009), ce critère d'évaluation a aussi souvent été critiqué pour son manque d'interprétation simple et concrète en

 ISSN: 2102-6238

biostatistique (Pepe and Janes, 2013). Pour faciliter son interprétation, il a donc été proposé de le normaliser, afin de créer un critère de type R^2 (Graf et al., 1999) :

$$R^2_\pi(s,t) = 1 - \frac{BS_\pi(s,t)}{BS_0(s,t)},$$

où $BS_0(s,t)$ est définie comme l'erreur de prédiction du modèle pronostique "nul" (de référence), qui prédit le même risque marginal $r_0(s,t) = \mathbb{P}\{D(s,t) = 1 | T > s\}$ à tous les sujets et vérifie $BS_0(s,t) = \mathrm{Var}\big[D(s,t) \,\big|\, T > s\big]$. Rappelons que ce modèle "nul" est parfaitement calibré mais, par définition, n'a aucun pouvoir discriminant. Cette normalisation a plusieurs avantages. Notamment, sous l'hypothèse que le modèle pronostique est calibré, on montre que

$$\begin{aligned} R^2_\pi(s,t) &= \frac{\mathrm{Var}\{\pi(s,t) \,|\, T > s\}}{\mathrm{Var}\{D(s,t) \,|\, T > s\}} \\ &= \mathrm{Corr}^2\big\{D(s,t), \pi(s,t) \,|\, T > s\big\} \\ &= \mathbb{E}\big\{\pi(s,t) \,|\, D(s,t) = 1, T > s\big\} - \mathbb{E}\big\{\pi(s,t) \,|\, D(s,t) = 0, T > s\big\}. \end{aligned}$$

Ces relations permettent des interprétations relativement simples et intéressantes en termes de variance expliquée par les prédictions, de corrélation entre prédictions et indicateurs d'occurrence d'événement et en termes d'une différence entre le risque moyen d'un cas et le risque moyen d'un contrôle (Pepe et al., 2008a; Tjur, 2009; Fournier et al., 2016). Dans tous les cas, même si le modèle n'est pas calibré, l'interprétation de ce $R^2(s,t)$ comme une mesure comparant les performances de l'outil pronostique $\pi(s,t)$ à celles de l'outil pronostique "nul" (de référence) est intéressante. La normalisation facilite aussi l'interprétation de l'évolution de $R^2(s,t)$ pour différents temps de prédiction s, par rapport à celle de l'erreur quadratique $BS(s,t)$, en rapportant les valeurs des erreurs quadratiques sur une échelle intelligible et commune à tous les temps de prédiction s.

4. Inférence

Cette section présente brièvement quelques méthodes de statistique inférentielle pour les critères d'évaluation pronostiques présentés. On s'intéresse ici uniquement à des méthodes non paramétriques, car en pratique on souhaite souvent disposer de méthodes simples et flexibles qui s'adaptent à tout type de modèle sans restrictions particulières, et permettent ainsi la comparaison de modèles pronostiques potentiellement très différents. Ces méthodes non paramétriques exigent cependant des tailles d'échantillons relativement importantes pour être efficaces.

Dans la suite, on suppose que l'on observe le n-échantillon i.i.d.

$$\Big\{\Big(\widetilde{T}_i, \Delta_i, \widetilde{\eta}_i, \pi_i(\cdot,t)\Big), i = 1, \ldots, n\Big\}$$

où $t > 0$ est un horizon de prédiction donné, C_i est un temps de censure, $\Delta_i = \mathbb{1}\{T_i \leq C_i\}$, $\widetilde{T}_i = \min(T_i, C_i)$, $\widetilde{\eta}_i = \Delta_i \eta_i$ et $\pi_i(\cdot,t)$ dénote l'ensemble des prédictions $\{\pi_i(s,t), s \in \mathscr{S}\}$ du sujet i pour un ensemble de temps de prédiction $\mathscr{S}$ donné.

La principale difficulté pour construire des procédures d'inférence vient de l'observation de données censurées. Pour un temps de prédictions s et un horizon t, la difficulté vient du fait

 ISSN: 2102-6238

que l'indicateur d'événement $D_i(s,t) = \mathbb{1}\{s \leq T_i < s+t, \eta_i = 1\}$ n'est pas observé pour tous les sujets. Seul l'indicateur $\widetilde{D}_i(s,t) = \mathbb{1}\{s < \widetilde{T}_i \leq s+t, \widetilde{\eta} = 1\} = \{1 - R_i(s,t)\}D_i(s,t)$ est toujours observé, avec $R_i(s,t) = \mathbb{1}\{s < \widetilde{T}_i < s+t, \Delta_i = 0\}$. En d'autres termes, l'observation de $D_i(s,t)$ est *manquante* dès que $R_i(s,t) = 1$, c'est-à-dire dès que le sujet est perdu de vue sain (i.e., non dément) entre le temps de prédiction s et la fin de la fenêtre de prédiction $s+t$. Dans ce cas, on ne sait pas si la valeur observée de $\widetilde{D}_i(s,t)$ correspond à celle de $D_i(s,t)$. Ce lien étroit entre la problématique des données censurées et celle des données manquantes est à l'origine de la vaste littérature sur les estimateurs basés sur la méthode de la pondération par l'inverse de la probabilité de censure (Tsiatis, 2006). Cette approche permet d'adapter simplement le calcul de proportions empiriques pour obtenir des estimateurs consistents. L'idée consiste en quelque sorte à utiliser l'information disponible sur le mécanisme de génération des données manquantes pour pondérer l'influence des observations non manquantes dans le calcul des estimateurs.

4.1. Estimation ponctuelle

Pour estimer $\mathrm{BS}_\pi(s,t)$, par exemple, l'approche par la méthode de pondération par l'inverse de la probabilité de censure tend à suggérer l'estimateur

$$\widehat{\mathrm{BS}}_\pi(s,t) = \frac{1}{n(s)} \sum_{i=1}^{n} \widehat{W}_i(s,t) \left\{ \widetilde{D}_i(s,t) - \pi_i(s,t) \right\}^2, \tag{2}$$

où $n(s) = \sum_{i=1}^{n} \mathbb{1}\{\widetilde{T}_i > s\}$ est le nombre de sujets à risque en s et où le poids $\widehat{W}_i(s,t)$ corrige pour le fait que $\widetilde{D}_i(s,t)$ est plus fréquemment observé égal à zéro que $D_i(s,t)$, à cause des données censurées (i.e., dès que $R_i(s,t) = 1$). Formellement, les poids sont définis comme

$$\widehat{W}_i(s,t) = \frac{\mathbb{1}\{\widetilde{T}_i > s+t\}}{\widehat{G}(s+t|s)} + \frac{\mathbb{1}\{s < \widetilde{T}_i \leq s+t\}\Delta_i}{\widehat{G}(\widetilde{T}_i|s)},$$

où $\widehat{G}(u)$ représente l'estimateur de Kaplan-Meier (Kaplan and Meier, 1958) de la fonction de survie du temps de censure en u, i.e. $\mathbb{P}(C > u)$, et où $\widehat{G}(u|s) = \widehat{G}(u)/\widehat{G}(s)$ estime la probabilité de ne pas être censuré en u conditionnellement au fait de ne pas l'être en s.

Des poids similaires sont souvent utilisés pour les calculs d'intégrales de Kaplan-Meier (Gill, 1994), ou même simplement pour réécrire l'estimateur de Kaplan-Meier de la fonction de survie $S(\cdot)$ du temps T sous la forme de sommes pondérées (Satten and Datta, 2001) :

$$\widehat{S}(t) = \frac{1}{n} \sum_{i=1}^{n} \widehat{W}_i(0,t)\, \mathbb{1}\{\widetilde{T}_i > t\} = 1 - \frac{1}{n} \sum_{i=1}^{n} \widehat{W}_i(0,t)\, \mathbb{1}\{\widetilde{T}_i \leq t\}.$$

Des formules similaires permettent aussi de réécrire l'estimateur non-paramétrique de Aalen-Johansen (Aalen and Johansen, 1978) du risque absolu $F_1(t) = P(T \leq t, \eta = 1)$, en présence de risques concurrents (Geskus, 2011). Des poids différents de ceux basés sur l'estimateur de Kaplan-Meier peuvent aussi être utilisés, notamment pour prendre en compte la particularité d'une censure dite "dépendante" (Tsiatis, 2006; Datta et al., 2010; Lopez, 2011).

Les formules des estimateurs pondérés pour $\mathrm{Se}_\pi(r;s,t)$, $\mathrm{Sp}_\pi(r;s,t)$, $\mathrm{ROC}_\pi(s,t)$, $\mathrm{AUC}_\pi(s,t)$ et $\mathrm{R}^2_\pi(s,t)$ sont ici omises car elles sont conceptuellement similaires à celle de $\widehat{\mathrm{BS}}_\pi(s,t)$ et déjà détaillées ailleurs (Blanche et al., 2015; Fournier et al., 2016).

 ISSN: 2102-6238

4.2. *Propriétés asymptotiques*

On peut montrer que les estimateurs obtenus par l'approche de pondération par l'inverse de la probabilité de censure, du type (2), ont généralement les bonnes propriétés asymptotiques usuelles, sous des hypothèses faibles. On peut notamment montrer des résultats de convergence et de linéarité asymptotique du type

$$\sup_{s\in\mathscr{S}}\left|\sqrt{n}\left(\widehat{\theta}_\pi(s,t)-\theta_\pi(s,t)\right)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{IF}_{\theta_\pi}\left(\widetilde{T}_i,\widetilde{\eta}_i,\pi_i(s,t),s,t\right)\right|=o_p(1)\,, \tag{3}$$

où $\theta_\pi(s,t)$ dénote l'une des quantités $\mathrm{AUC}_\pi(s,t)$, $\mathrm{BS}_\pi(s,t)$ ou $\mathrm{R}^2_\pi(s,t)$ et $\widehat{\theta}_\pi(s,t)$ l'estimateur correspondant, et où les termes i.i.d. $\mathrm{IF}_{\theta_\pi}(\widetilde{T}_i,\widetilde{\eta}_i,\pi_i(s,t),s,t)$, $i=1,\ldots,n$ sont d'espérance nulle et de variance finie. Ce sont ces résultats qui sont à la base de toutes les procédures d'inférence décrites ci-après.

Brièvement, pour montrer des résultats du type (3), il suffit de supposer les hypothèses suffisantes à la convergence en loi de l'estimateur de Kaplan-Meier, auxquelles s'ajoutent des hypothèses d'identifiabilité naturelles assurant que les probabilités d'observer un cas et un contrôle sont non nulles pour tous $s\in\mathscr{S}$. Cela revient donc, essentiellement, à supposer une censure indépendante (Andersen et al., 1993).

Des éléments de preuves ainsi que les détails des formules des fonctions d'influence $\mathrm{IF}_\theta(\cdot)$ et de leur estimateur $\widehat{\mathrm{IF}}_\theta(\cdot)$ sont présentées dans Blanche et al. (2015). Des résultats similaires ont été obtenus, entre autres, par Hung and Chiang (2010); Datta et al. (2010); Parast et al. (2012). Des résultats très généraux, issus de l'élégante théorie de la géométrie des fonctions d'influence (Tsiatis, 2006), auraient aussi pu être utilisés pour obtenir des preuves et des formules des quantités $\mathrm{IF}_\theta(\cdot)$.

4.3. *Régions de confiance et tests*

En s'appuyant sur la représentation i.i.d. (3), on estime simplement la variance de $\widehat{\theta}_\pi(s,t)$ à partir des estimations des termes i.i.d., par

$$\widehat{\sigma}^2_{(s,t)}=\frac{1}{n}\sum_{i}^{n}\left\{\widehat{\mathrm{IF}}_{\theta_\pi}(\widetilde{T}_i,\widetilde{\eta}_i,\pi_i(s,t),s,t)\right\}^2. \tag{4}$$

En combinant cette estimation au résultat de normalité asymptotique, on en déduit alors des intervalles de confiances ponctuels usuels du type

$$\left[\widehat{\theta}_\pi(s,t)\pm z_{1-\alpha/2}\frac{\widehat{\sigma}_{(s,t)}}{\sqrt{n}}\right], \tag{5}$$

où $z_{1-\alpha/2}$ est le quantile à $100(1-\alpha/2)\%$ d'une loi normale centrée réduite (i.e. 1.96 pour $\alpha=5\%$). Des intervalles de confiance simultanés pour tous les temps $s\in\mathscr{S}$ peuvent aussi être dérivés. Ils permettent d'estimer des régions de confiance contenant toute la courbe $\left\{\left(s,\theta_\pi(s,t)\right),s\in\mathscr{S}\right\}$ avec un niveau de confiance asymptotique à $100(1-\alpha)\%$. Pour cela, il suffit de remplacer le quantile $z_{1-\alpha/2}$ dans la définition de l'intervalle de confiance (5) par un quantile approprié, noté $\widehat{q}^{(\mathscr{S},t)}_{1-\alpha}$, qui prend en compte la corrélation des estimations aux différents temps de prédiction s, et que l'on peut calculer numériquement par l'algorithme suivant :

 ISSN: 2102-6238

1. Pour $b = 1, \ldots, B$, avec B grand (e.g. $B = 4\,000$) :
   (a) Générer n réalisations $(\omega_1^b, \ldots, \omega_n^b)$ d'une loi normale centrée réduite.
   (b) Calculer
$$\Upsilon^b = \sup_{s \in \mathscr{S}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \omega_i^b \frac{\widehat{\mathrm{IF}}_{\theta_\pi}(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i(s,t), s, t)}{\widehat{\sigma}_{(s,t)}} \right|.$$
2. Calculer $\widehat{q}_{1-\alpha}^{(\mathscr{S},t)}$ comme le $100(1-\alpha)$ème percentile de $\{\Upsilon^1, \ldots, \Upsilon^B\}$.

Ce type d'algorithme est fréquemment utilisé en analyse des données de survie depuis la parution de l'article de Lin et al. (1994) et s'apparente à du "*Wild bootstrap*" (Martinussen and Scheike, 2006; Beyersmann et al., 2013).

Des régions de confiance pour des différences entre les capacités pronostiques de deux modèles pronostics distincts $\pi^{(1)}$ et $\pi^{(2)}$ peuvent aussi se calculer de façon similaire. Par exemple, on peut remplacer $\widehat{\theta}_\pi(s,t)$ par $\widehat{\theta}_{\pi^{(1)}}(s,t) - \widehat{\theta}_{\pi^{(2)}}(s,t)$ dans (5) et $\widehat{\mathrm{IF}}_{\theta_\pi}(\cdot)$ par $\widehat{\mathrm{IF}}_{\theta_{\pi^{(1)}}}(\cdot) - \widehat{\mathrm{IF}}_{\theta_\pi^{(2)}}(\cdot)$ dans (4), pour obtenir l'intervalle de confiance d'une différence ainsi qu'un test associé ayant pour hypothèse nulle $\mathscr{H}_0^{(s)} : \theta_{\pi^{(1)}}(s,t) = \theta_{\pi^{(2)}}(s,t)$. L'hypothèse nulle $\mathscr{H}_0 = \cap_{s \in \mathscr{S}} \mathscr{H}_0^{(s)}$, i.e., $\mathscr{H}_0 = \forall_{s \in \mathscr{S}}\ \theta_{\pi^{(1)}}(s,t) = \theta_{\pi^{(2)}}(s,t)$, peut aussi être considérée en utilisant un intervalle de confiance simultané (Blanche et al., 2015).

5. Application au pronostic de la démence

On revient ici sur l'exemple des données Paquid et Trois-Cités présentées en section 2. Les prédictions que l'on cherche ici à évaluer et comparer sont celles basées sur deux modèles pronostiques distincts, notés $\pi^{(I)}$ et $\pi^{(M)}$. Pour tout s, ils associent à des informations disponibles au temps de prédiction s, notées $\mathscr{Y}^{(I)}(s)$ et $\mathscr{Y}^{(M)}(s)$, les risques d'apparition d'une démence dans les t prochaines années, notés $\pi^{(I)}(s,t)$ et $\pi^{(M)}(s,t)$. Ici $\mathscr{Y}^{(I)}(s)$ et $\mathscr{Y}^{(M)}(s)$ représentent l'information disponible sur l'âge, le sexe, le niveau d'étude d'une personne, et l'évolution de ses scores à l'IST pour $\mathscr{Y}^{(I)}(s)$, et au MMSE pour $\mathscr{Y}^{(M)}(s)$, au cours de s années de suivi.

5.1. Calcul des prédictions

Brièvement, l'approche de modélisation utilisée ici est complètement paramétrique et basée sur une approche à classes latentes modélisant conjointement l'évolution d'un test psychométrique et les risques concurrents de démence et de décès sans démence (Proust-Lima et al., 2015). Un modèle pronostique a été estimé pour chacun des deux tests psychométriques en utilisant les données de la cohorte Paquid, par maximisation de la vraisemblance. Essentiellement, les évolutions aux tests psychométriques sont modélisées par des modèles linéaires à effets aléatoires et les risques instantanés de démence et de décès sont modélisés par des modèles à risques proportionnels. La structure d'association qui joint les sous-modèles linéaires mixtes aux sous-modèles à risques proportionnels est hérité de leur dépendance commune aux classes latentes. Pour d'avantage de détails, on pourra consulter Blanche et al. (2015).

Comme ces modèles paramétrisent complètement la distribution jointe de l'évolution d'un test psychométrique et des temps et causes d'événement, les risques de démence individuels

 ISSN: 2102-6238

TABLEAU 1. *Statistiques descriptives : nombre de sujets à risque en s et effectifs (et fréquences, en %) de ceux qui sont, parmi eux, observés diagnostiqués déments entre s et $s+t$, observés vivants et non déments en $s+t$, observés décédés sans démence entre s et $s+t$ ou observés censurés entre s et $s+t$. Données des Trois-Cités pour $s = 0,1,\ldots,4$ ans et $t = 5$ ans.*

	$n(s)$	Déments	Survivants (non déments)	Décédés (non déments)	Censurés
$s=0$	3 880	182 (4.7)	2 838 (73.1)	248 (6.4)	612 (15.8)
$s=1$	3 782	224 (5.9)	2 689 (71.1)	257 (6.8)	612 (16.2)
$s=2$	3 521	218 (6.2)	2 541 (72.2)	205 (5.8)	557 (15.8)
$s=3$	3 373	179 (5.3)	2 284 (67.7)	187 (5.5)	723 (21.4)
$s=4$	2 963	224 (7.6)	1 579 (53.3)	188 (6.3)	972 (32.8)

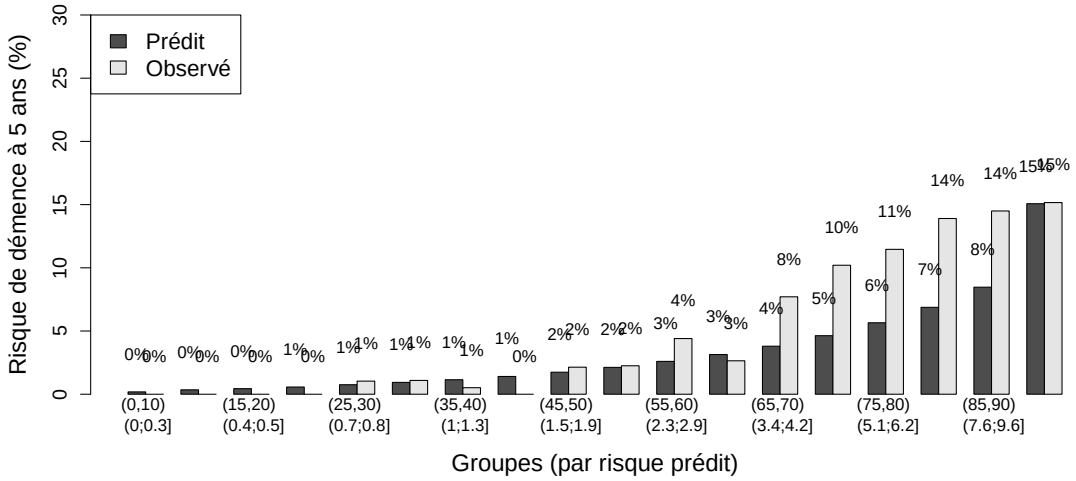
$\pi_i^{(I)}(s,t)$ et $\pi_i^{(M)}(s,t)$ se calculent aisément numériquement, pour tout sujet $i = 1,\ldots,n$ inclus dans l'échantillon des Trois-Cités, temps de prédiction s et horizon t. Dans la suite, on considérera un unique horizon de prédiction $t = 5$ ans et des temps de prédiction $s \in \mathscr{S} = \{0, 0.5, 1, \ldots, 4\}$ ans.

5.2. *Estimations des capacités pronostiques*

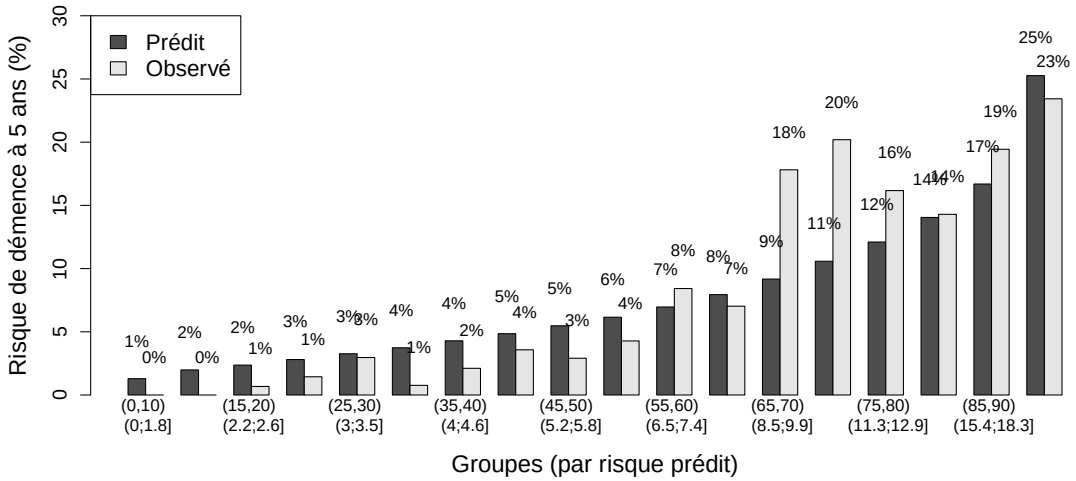
Quelques évaluations des capacités pronostiques des prédictions issues des modèles $\pi^{(I)}$ et $\pi^{(M)}$, précédemment estimés avec les données de Paquid, sont ici présentées. Elles sont basées sur les externes données de la cohorte des Trois-Cités. On procède donc ici à une évaluation externe qui mime la situation réaliste dans laquelle des prédictions sont calculées pour des sujets d'une population différente et plus récente, bien que supposée similaire, de celle dont sont issus les sujets de Paquid. Dans notre contexte, les validations externes sont en effet souvent jugées préférables (Steyerberg, 2009; Vickers and Cronin, 2010). Les données contiennent de nombreuses observations de décès sans démence (i.e. l'événement concurrent de la démence) et de données censurées. Au tableau 1, on constate d'ailleurs qu'on observe plus d'événements concurrents de décès sans démence et de données censurées que de démence, qui est ici l'événement d'intérêt.

La figure 1 présente une évaluation graphique de la calibration des prédictions issues du modèle $\pi^{(M)}$, à $s = 0$ et $s = 4$ ans. Cette évaluation est basée sur les observations des $n(0) = 3\ 880$ et $n(4) = 2\ 963$ sujets à risque de démence en $s = 0$ et $s = 4$ ans. Les risques d'événement observés dans chaque groupe sont ici calculés avec l'estimateur non-paramétrique de Aalen-Johansen (Aalen and Johansen, 1978) pour prendre en compte à la fois les données censurées et le risque concurrent de décès sans démence. Les groupes sont définis en fonction des valeurs des quantiles à 0%, 10%, 15%, 20%,...,85%, 90% et 100% de la distribution des risques prédits $\{\pi_i^{(M)}(s,5), i = 1,\ldots,n(s)\}$, pour $s = 0,4$ ans. La valeur de certains quantiles est indiquée en abscisse, par exemple 2.2% pour le quantile à 15%, pour $s = 4$ ans, sur la figure 1b. Globalement, ces évaluations graphiques tendent à suggérer une sous-évaluation des risques prédits de démence à 5 ans pour $s = 0$. À $s = 4$ ans, la calibration apparaît cependant plutôt bonne, hormis pour les deux groupes des sujets ayant un risque de démence à 5 ans prédit entre 8 à 11.3%. Ces résultats, en demi-teinte, sont probablement dûs au fait que les populations dont sont issues les deux cohortes sont quelque peu différentes. Les modèles $\pi^{(I)}$ et $\pi^{(M)}$, qui ont été estimés avec les données de Paquid, semblent ainsi ne pas être parfaitement calibrés pour des sujets issus d'une population différente, plus récente et bien plus urbaine, comme celle des sujets inclus dans la

 ISSN: 2102-6238



(a) $s = 0$



(b) $s = 4$ ans.

FIGURE 1: *Évaluation graphique de la calibration des prédictions basées entre autres sur l'évolution du MMSE. Temps de prédiction $s = 0$ et $s = 4$ ans, horizon de prédiction $t = 5$ ans, données de la cohorte des Trois-Cités, $n(0) = 3\,880$ et $n(4) = 2\,963$.*

cohorte des Trois-Cités.

L'estimation de courbes ROC à $s = 4$ ans est présentée sur la figure 2. Elle permet, entre autres, de visualiser des estimations des taux de vrais positifs (sensibilité) et de faux positifs (1-spécificité) qui seraient associés à un test pronostique d'une démence à $t = 5$ ans. Ces tests seraient basés sur un des modèles pronostiques $\pi^{(I)}$ ou $\pi^{(M)}$ et $s = 4$ ans de suivi. Par exemple, on peut lire sur le graphique que si on définit un test comme positif si $\pi^{(M)}(4,5) > r$, tel qu'on est un taux d'environ 20% de faux positifs, i.e., $1 - \mathrm{Sp}_{\pi^{(M)}}(r;4,5) = 20\%$, (ce qui correspond ici

 ISSN: 2102-6238

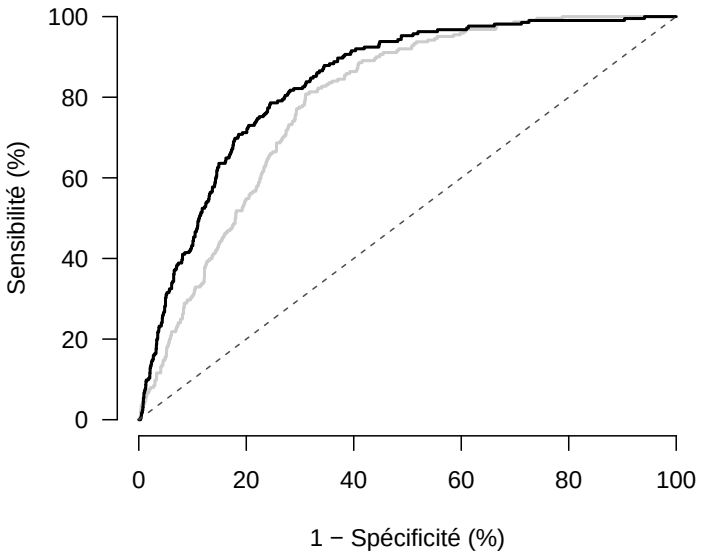
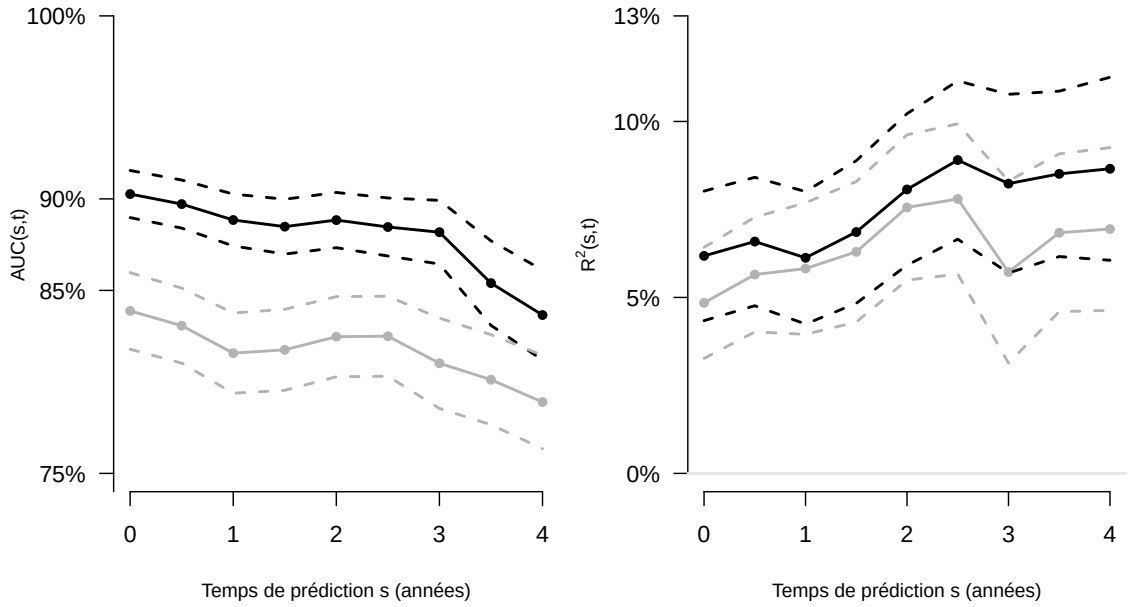


FIGURE 2: *Courbe ROC pour les prédictions issues des deux modèles, dont l'un est basé sur l'évolution du MMSE ($\pi^{(M)}$, en gris) et l'autre sur celle de l'IST ($\pi^{(I)}$, en noir). Temps de prédiction $s = 4$ ans, horizon de prédiction $t = 5$ ans, données de la cohorte des Trois-Cités, $n(s) = 2\,963$.*

à une valeur seuil du risque prédit de démence de $r = 12\%$) alors on estime qu'on aura un taux d'environ 55% de vrais positifs, i.e, $\mathrm{Se}_{\pi^{(M)}}(r;4,5) = 55\%$. Pour ce même taux de faux positifs de 20%, en utilisant $\pi^{(I)}$ au lieu de $\pi^{(M)}$, on estime qu'on pourrait avoir un bien meilleur taux de vrais positifs, d'environ 70%. L'estimation des courbes ROC suggère aussi que le modèle pronostique $\pi^{(I)}$ sera meilleur que $\pi^{(M)}$ pour définir un test pronostique de démence à $t = 5$ qui utilise $s = 4$ ans de suivi. Et ceci quel que soit le choix de la valeur du taux de faux positif (i.e., il sera uniformément plus puissant). La courbe ROC associée à $\pi^{(I)}$ est en effet au-dessus de celle associée à $\pi^{(M)}$ pour (presque) toutes les valeurs possibles de la spécificité.

La figure 3 présente l'évolution des critères résumés de capacités pronostiques $\mathrm{AUC}_\pi(s,5)$ et $\mathrm{R}^2_\pi(s,5)$ pour $\pi = \pi^{(I)}$ et $\pi = \pi^{(M)}$, lorsque $s \in \mathscr{S}$ croît. Elle permet d'obtenir un résumé concis de l'estimation des neuf évolutions que l'ont auraient pu présenter en figure 2 en choisissant une valeur $s \in \mathscr{S}$ quelconque. Les tendances décroissantes tendent à suggérer que les capacités de discrimination des prédictions diminuent lorsque s croît, alors que pourtant l'information disponible pour fournir des prédictions personnalisées augmente. Plutôt inattendue, cette décroissance pourrait s'expliquer par le fait que la relation entre le risque prédit et le risque observé ne soit pas toujours tout à fait croissante, et que ce problème croisse avec s. Au vu des figures 1a et 1b, c'est une explication qui paraît naturelle. La tendance de l'évolution des critères R^2 est quant à elle croissante en s. Ceci apparaît assez cohérent avec la figure 1 et la décomposition du Brier score (1). La variabilité des risques prédits apparaît en effet croissante lorsque s croît à la figure 1, tandis que le terme de biais de la décomposition (1) semble lui rester plutôt stable. Les intervalles de confiance ponctuels à 95% ajoutés aux graphiques permettent aussi de visualiser la précision des estimations, qui décroît naturellement avec s et la taille $n(s)$ des échantillons de sujets à risque

 ISSN: 2102-6238



(a) Évolution de l'AUC. (b) Évolution du R^2

FIGURE 3: *Évolutions de critères d'évaluation de capacités pronostiques.*
$AUC_\pi(s,t)$ et $R^2_\pi(s,t)$ pour $\pi = \pi^{(I)}$, en noir, et $\pi = \pi^{(M)}$, en gris, pour $s \in \mathscr{S}$. Les estimations ponctuelles sont présentées en trait plein et les intervalles de confiances ponctuels à 95% en tirets.

en s (tableau 1).

La figure 4 présente, entre autres, des intervalles de confiance à 95% ponctuels et simultanées pour l'évolution de la différence $\text{AUC}_{\pi^{(I)}}(s,t) - \text{AUC}_{\pi^{(M)}}(s,t)$, pour $s \in \mathscr{S} = \{0, 0.5, \ldots, 4\}$ ans. Comme pour chaque $s \in \mathscr{S}$ les intervalles de confiance ponctuels ne contiennent pas la valeur zéro, alors pour chaque test ayant pour hypothèse nulle $\mathscr{H}_0^{(s)} : \text{AUC}_{\pi^{(I)}}(s,t) = \text{AUC}_{\pi^{(M)}}(s,t)$, cette hypothèse nulle est rejetée. Par ailleurs, comme la bande de confiance simultanée ne croise pas la droite d'équation $y = 0$, alors le test ayant pour hypothèse nulle $\mathscr{H}_0 = \cap_{s\in\mathscr{S}} \mathscr{H}_0^{(s)}$ est lui aussi significatif, au risque de première espèce $\alpha = 5\%$. Ainsi, en termes d'AUC (de discrimination), on conclut que les capacités pronostiques de l'outil pronostique basé sur une évolution du populaire MMSE, construit à partir des données de la cohorte Paquid, apparaissent globalement moins bonnes que celles de celui basé sur le prometteur IST.

6. Remarques de conclusion

Dans ce manuscrit, on a présenté quelques critères résumés pour évaluer les capacités pronostiques de prédictions dynamiques, et exposé quelques méthodes d'inférence associées, qui s'adaptent au contexte dynamique et à la présence de risques concurrents et de données censurées.

On s'est ici intéressé à des critères statistiques assez généraux, qui sont parfois critiqués pour leur manque de pertinence pour juger de l'intérêt clinique de prédictions (Vickers, 2008). Ces critères statistiques n'ont en effet pas l'ambition de fournir un résumé exhaustif permettant d'éva-

 ISSN: 2102-6238

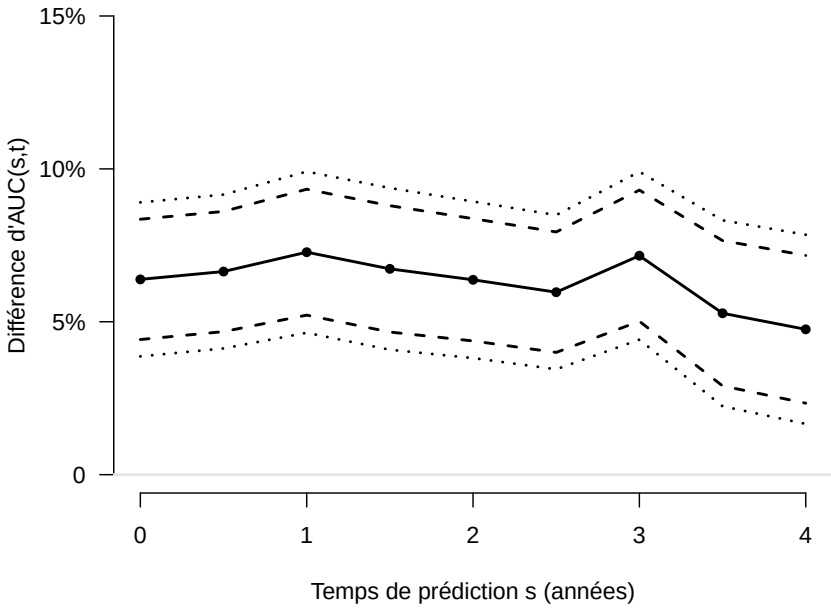


FIGURE 4: *Différences $AUC_{\pi^{(I)}}(s,t) - AUC_{\pi^{(M)}}(s,t)$, $s \in \mathscr{S}$, comparant les capacités pronostiques des prédictions issues des deux modèles pronostiques.*

luer complètement des modèles pronostiques dans le but d'applications cliniques particulières. Pour répondre à cet objectif, d'autres méthodes seraient en effet à privilégier (Pepe and Janes, 2013). Ils sont cependant des résumés intéressants dans le cadre d'une première évaluation, au stade où souvent les circonstances dans lesquelles les prédictions pourraient être cliniquement utiles restent encore à définir (Hand, 2010). Leur utilisation est également pertinente lorsqu'on s'intéresse à évaluer les prédictions issues de plusieurs modèles pronostiques et que l'on cherche d'abord à identifier les plus prometteurs d'entre eux, pour pouvoir par la suite en étudier de façon plus approfondie un nombre limité (Pepe and Janes, 2013).

On a présenté des méthodes d'inférence qui supposent que l'on dispose de données externes. Cela suppose que les modèles pronostiques permettant de calculer des prédictions ont été précédemment estimés à l'aide de données indépendantes. Il arrive cependant qu'on souhaite utiliser les mêmes données pour estimer les modèles et évaluer leurs prédictions. Cela se produit notamment lorsqu'on ne peux pas faire autrement, bien que ce serait préférable (Vickers and Cronin, 2010). Cette situation est alors significativement plus compliqué à traiter, tant du point de la construction de régions de confiance dès qu'une procédure de validation croisée est utilisée (Uno et al., 2007; Tian et al., 2007), que du point de vue de l'interprétation des résultats (Vickers and Cronin, 2010; Hastie et al., 2009, sec. 7.10–7.12).

Toutes les statistiques, en synthétisant des données, sacrifient quelques-uns des détails constituant la richesses des données. Plus l'information est résumée et plus elle est facile à communiquer, mais moins elle est riche. L'une des manières de rapporter une information assez exhaustive, permettant d'apprécier plus amplement les capacités pronostiques de prédictions, est probablement de reporter de nombreux graphiques de calibration, comme ceux présentés à la figure 1.

 ISSN: 2102-6238

Des estimations d'AUC, du Brier score, de sensibilités, de spécificités et d'autres statistiques intéressantes peuvent d'ailleurs se déduire de la lecture de ce type de graphiques (Huang et al., 2007; Pepe et al., 2008b; Viallon and Latouche, 2011).

Les critères ici présentés ont déjà été largement étudiés et, bien qu'ils présentent de nombreuses limites, il existe un consensus sur leur intérêt. Ce consensus repose, entre autres, sur des fondements théoriques solides (Gneiting and Raftery, 2007; Hilden and Gerds, 2014). À l'inverse, d'autres critères, plus récemment proposés et pourtant attractifs à première vue ont par la suite été largement discrédités (Hilden, 2014; Hilden and Gerds, 2014). Cette récente mésaventure illustre la difficulté de définir de nouveaux critères d'évaluation réellement préférables à ceux ici présentés.

Remerciements

Je remercie chaleureusement Hélène Jacqmin-Gadda et Cécile Proust-Lima pour les belles collaborations à l'origine de ce papier, ainsi que la fondation Bettencourt Schueller pour son soutien.

Références

Aalen, O. O. and Johansen, S. (1978). An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics*, pages 141–150.

Aisen, P., Andrieu, S., Sampaio, C., Carrillo, M., Khachaturian, Z., Dubois, B., Feldman, H., Petersen, R., Siemers, E., Doody, R., et al. (2011). Report of the task force on designing clinical trials in early (predementia) AD. *Neurology*, 76(3) :280–286.

Amieva, H., Le Goff, M., Millet, X., Orgogozo, J. M., Pérès, K., Barberger-Gateau, P., Jacqmin-Gadda, H., and Dartigues, J. F. (2008). Prodromal Alzheimer's disease : successive emergence of the clinical symptoms. *Annals of neurology*, 64(5) :492–498.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New york : Springer Verlag, New York.

Beyersmann, J., Termini, S. D., and Pauly, M. (2013). Weak convergence of the wild bootstrap for the Aalen–Johansen estimator of the cumulative incidence function of a competing risk. *Scandinavian Journal of Statistics*, 40(3) :387–402.

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. *Biometrics*, 71(1) :102–113.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1) :1–3.

Camm et al, A. (2010). Guidelines for the management of atrial fibrillation : The task force for the management of atrial fibrillation of the european society of cardiology (esc). *European heart journal*, 31 :2369–2429.

Dartigues, J., Gagnon, M., Barberger-Gateau, P., Letenneur, L., Commenges, D., Sauvel, C., Michel, P., and Salamon, R. (1992). The Paquid epidemiological program on brain ageing. *Neuroepidemiology*, 11(1) :14–18.

Datta, S., Bandyopadhyay, D., and Satten, G. A. (2010). Inverse probability of censoring weighted u-statistics for right-censored data with an application to testing hypotheses. *Scandinavian Journal of Statistics*, 37(4) :680–700.

Fournier, M.-C., Dantan, E., and Blanche, P. (2016). R^2-curve in the dynamic prediction context. *in preparation*.

Freedman, A. N., Seminara, D., Gail, M. H., Hartge, P., Colditz, G. A., Ballard-Barbash, R., and Pfeiffer, R. M. (2005). Cancer risk prediction models : a workshop on development, evaluation, and application. *Journal of the National Cancer Institute*, 97(10) :715–723.

Gail, M. H. and Pfeiffer, R. M. (2005). On criteria for evaluating models of absolute risk. *Biostatistics*, 6(2) :227–239.

Gerds, T. A., Cai, T., and Schumacher, M. (2008). The performance of risk prediction models. *Biometrical Journal*, 50(4) :457–479.

 ISSN: 2102-6238

Geskus, R. B. (2011). Cause-Specific Cumulative Incidence Estimation and the Fine and Gray Model Under Both Left Truncation and Right Censoring. *Biometrics*, 67(1) :39–49.

Gill, R. D. (1994). Lectures on survival analysis. In *Bakry, Dominique (ed.) et al., Lectures on probability theory. Ecole d'Eté de Probabilités de Saint-Flour XXII-1992. Summer School, 9th- 25th July, 1992, Saint-Flour, France. Berlin : Springer-Verlag. Lect. Notes Math. 1581, 115-241* .

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477) :359–378.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18 :2529–2545.

Hand, D. J. (2010). Evaluating diagnostic tests : the area under the roc curve and the balance of errors. *Statistics in Medicine*, 29(14) :1502–1510.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Second edition, New York : Springer.

Hilden, J. (2014). Commentary : On NRI, IDI, and "good-looking" statistics with nothing underneath. *Epidemiology*, 25(2) :265–267.

Hilden, J. and Gerds, T. A. (2014). A note on the evaluation of novel biomarkers : do not rely on integrated discrimination improvement and net reclassification index. *Statistics in Medicine*, 33(19) :3405–3414.

Huang, Y., Sullivan Pepe, M., and Feng, Z. (2007). Evaluating the predictiveness of a continuous marker. *Biometrics*, 63(4) :1181–1188.

Hung, H. and Chiang, C. (2010). Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics*, 38(1) :8–26.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282) :457–481.

Lin, D., Fleming, T., and Wei, L. (1994). Confidence bands for survival curves under the proportional hazards model. *Biometrika*, 81(1) :73–81.

Lopez, O. (2011). Nonparametric estimation of the multivariate distribution function in a censored regression model with applications. *Communications in Statistics-Theory and Methods*, 40(15) :2639–2660.

Martinussen, T. and Scheike, T. (2006). *Dynamic regression models for survival data*. Springer.

Nicolaie, M., van Houwelingen, J., de Witte, T., and Putter, H. (2013). Dynamic pseudo-observations : A robust approach to dynamic prediction in competing risks. *Biometrics*, 69(4) :1043–1052.

Parast, L., Cheng, S.-C., and Cai, T. (2012). Landmark Prediction of Long-Term Survival Incorporing Short-Term Event Time Information. *Journal of the American Statistical Association*, 107(500) :1492–1501.

Pepe, M. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA.

Pepe, M., Feng, Z., and Gu, J. (2008a). Comments on 'Evaluating the added predictive ability of a new marker : From area under the ROC curve to reclassification and beyond'by MJ Pencina et al., Statistics in Medicine (DOI : 10.1002/sim. 2929). *Statistics in Medicine*, 27(2) :173–181.

Pepe, M. and Janes, H. (2013). Methods for evaluating prediction performance of biomarkers and tests. In Lee, M.-L., Gail, G., Cai, T., Pfeiffer, R., and Gandy, A., editors, *Risk Assessment and Evaluation of Predictions*. Springer.

Pepe, M. S., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I. M., and Zheng, Y. (2008b). Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology*, 167(3) :362–368.

Proust-Lima, C., Amieva, H., Dartigues, J.-F., and Jacqmin-Gadda, H. (2007). Sensitivity of four psychometric tests to measure cognitive changes in brain aging-population–based studies. *American Journal of Epidemiology*, 165(3) :344–350.

Proust-Lima, C., Dartigues, J.-F., and Jacqmin-Gadda, H. (2015). Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death : a latent process and latent class approach. *Statistics in Medicine*.

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.

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

Satten, G. and Datta, S. (2001). The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, 55(3) :207–210.

Steyerberg, E. (2009). *Clinical prediction models : a practical approach to development, validation, and updating*.

 ISSN: 2102-6238

Springer.
Taylor, J. M., 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.
The 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 :316–325.
Tian, L., Cai, T., Goetghebeur, E., and Wei, L. (2007). Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika*, 94(2) :297–311.
Tjur, T. (2009). Coefficients of determination in logistic regression models–a new proposal : The coefficient of discrimination. *The American Statistician*, 63(4) :366–372.
Tsiatis, A. (2006). *Semiparametric theory and missing data*. New York : Springer Verlag.
Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data : an overview. *Statistica Sinica*, 14(3) :809–834.
Uno, H., Cai, T., Tian, L., and Wei, L. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478) :527–537.
van Houwelingen, H. and Putter, H. (2012). *Dynamic Prediction in Clinical Survival Analysis*. CRC Press.
van Houwelingen, H. C. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1) :70–85.
Viallon, V. and Latouche, A. (2011). Discrimination measures for survival outcomes : connection between the AUC and the predictiveness curve. *Biometrical Journal*, 53(2) :217–236.
Vickers, A. J. (2008). Decision analysis for the evaluation of diagnostic tests, prediction models, and molecular markers. *The American Statistician*, 62(4).
Vickers, A. J. and Cronin, A. M. (2010). Everything you always wanted to know about evaluating prediction models (but were too afraid to ask). *Urology*, 76(6) :1298–1301.
Zheng, Y. and Heagerty, P. J. (2005). Partly conditional survival models for longitudinal data. *Biometrics*, 61(2) :379–391.

 ISSN: 2102-6238