Thèse n° 2098 Année 2013

Université Bordeaux 2

THÈSE

pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITÉ BORDEAUX 2

Mention : Sociétés, Politique, Santé Publique
Spécialité : Santé Publique - Option : Biostatistique

École Doctorale Sociétés, Politique et Santé Publique (ED SP2)

Contribution à l'évaluation de capacités pronostiques en présence de données censurées, de risques concurrents et de marqueurs longitudinaux.

Inférence et applications à la prédiction de la démence.

Présentée et soutenue publiquement
Le 10 décembre 2013
par
Paul Blanche
Né le 14 Juillet 1986 à Équemauville

Membres du jury :

M. Laurent Bordes	Professeur	Pau	Examinateur
M. Daniel Commenges	Directeur de Recherche	Bordeaux	Président
M. Jean François Dartigues	Professeur	Bordeaux	Examinateur
M. Yohann Foucher	Maître de Conférence	Nantes	Rapporteur
Mme Agathe Guilloux	Maître de Conférence	Paris	Rapporteur
Mme Hélène Jacqmin-Gadda	Directeur de Recherche	Bordeaux	Directeur de thèse

Résumé

Ce travail a eu pour objectif de proposer des méthodes statistiques pour évaluer et comparer les capacités prédictives de divers outils pronostiques. Le Brier score et principalement les courbes ROC dépendant du temps ont été étudiés. Tous deux dépendent d'un temps t, représentant un horizon de prédiction. Motivé par les applications à la prédiction de la démence et des données de cohortes de personnes âgées, ce travail s'est spécifiquement intéressé à des procédures d'inférence en présence de données censurées et de risques concurrents. Le risque concurrent de décès sans démence est en effet important lorsque l'on s'intéresse à prédire une démence chez des sujets âgés. Pour obtenir des estimateurs consistants, nous avons utilisé une méthode appelée « Inverse Probability of Censoring Weighting » (IPCW). Dans un premier travail, nous montrons qu'elle permet d'étendre simplement les estimateurs pour données non censurées et de prendre en compte une censure éventuellement dépendante de l'outil pronostique étudié. Dans un second travail, nous proposons des adaptations pour les situations de risques concurrents. Quelques résultats asymptotiques sont donnés et permettent de dériver des régions de confiance et des tests de comparaison d'outils pronostiques. Enfin, un troisième travail s'intéresse à la comparaison d'outils pronostiques dynamiques, basés sur des marqueurs longitudinaux. Les mesures de capacités pronostiques dépendent ici à la fois du temps s auquel on fait la prédiction et de l'horizon de prédiction t. Des courbes de capacités pronostiques selon s sont proposées pour leur évaluation et quelques procédures d'inférence sont développées, permettant de construire des régions de confiance et des tests de comparaison de ces courbes. L'application des méthodes proposées a permis de montrer que des outils prédictifs de la démence basés sur des tests cognitifs ou des mesures répétées de ces tests ont de bonnes capacités pronostiques.

Mots clés : Alzheimer, Brier score, courbe ROC, démence, marqueurs longitudinaux, prédiction, censure, risques concurrents.

Abstract

The objective of this work is to develop statistical methods that can be used to evaluate and compare the prognostic ability of different prognostic tools. To measure prognostic ability, mainly the time-dependent ROC curve is studied and also the Brier score for a prediction horizon t. Motivated by applications where the aim is to predict the risk of dementia in cohort data of elderly people, this work focuses on inference procedures in the presence of right censoring and competing risks. In elderly populations death is a highly prevalent competing risk. To define consistent estimators of the prediction ability measures, we use the inverse probability of censoring weighting (IPCW) approach. In our first work, we show that the IPCW approach provides consistent estimators of prediction ability based on right censored data, even when the censoring distribution is marker-dependent. In our second work, we adapt the estimators to settings with competing risks. Asymptotic results are provided and we derive confidence regions and tests for comparing different prognostic tools. Finally, in a third work we focus on comparing dynamic prognostic tools which use information from repeated marker measurements to predict future events. The prognostic ability measures now depend on both the time s at which predictions are made and on the prediction horizon t. Curves of the prognostic ability as a function of s are developed for the evaluation of dynamic risk predictions. Inference procedures are adapted and so are confidence regions and tests to compare the curves. The applications of the proposed methods to cohort data show that the prognostic tools that use cognitive tests, or repeated measurements of cognitive tests, have high prognostic abilities.

Key words: Alzheimer, Brier score, ROC curve, dementia, longitudinal markers, prediction, censoring, competing risks.

« Quels que soient les progrès des connaissances humaines, il y aura toujours place pour l'ignorance et par suite pour le hasard et la probabilité. »

Emile Borel, *Le hasard* (1914).

Remerciements

À Madame Hélène Jacqmin-Gadda

Merci pour la confiance que vous m'avez accordée, pour vos enseignements, et pour vos conseils et vos encouragements du début de mon stage de fin d'études jusqu'à la fin de ma thèse. Merci pour cet accompagnement tout au long de cette expérience extraordinairement enrichissante, qui me permet maintenant de m'envoler vers de nouveaux horizons. Soyez assurée de ma profonde gratitude.

À Monsieur Yohann Foucher

J'ai lu ton article paru dans *Statistics in Medicine* en 2010 durant les premiers mois de ma thèse. Il m'a permis d'apprendre beaucoup. Ton expertise sur les sujets abordés dans cette thèse m'a aussi beaucoup apporté lors de nos trop rares discussions.

Je suis très honoré que tu aies accepté de juger ce travail. Tes connaissances et ton expérience des courbes ROC dépendant du temps m'apporteront assurément un éclairage nouveau sur mon travail. Pour avoir accepté d'être le rapporteur de cette thèse, soit assuré de ma profonde gratitude.

À Madame Agathe Guilloux

L'isup m'a bien formé. Cette école m'a appris autant à analyser soigneusement des données qu'à comprendre des notions plus théoriques, par exemple sur les fondements de la statistique inférentielle. Vos cours, et notamment le cours de Bootstrap, que j'ai beaucoup apprécié, y ont largement participé.

Je suis très honoré que vous ayez accepté de juger ce travail. Vos connaissances approfondies des données de survie m'apporteront beaucoup. Pour vos enseignements passés et pour avoir accepté d'être rapporteur de ma thèse, soyez assurée de ma profonde reconnaissance.

À Monsieur Laurent Bordes

Merci de me faire l'honneur de juger ce travail. Vos connaissances approfondies des statistiques non paramétriques et des données de survie m'apporteront assurément un regard éclairé sur mon travail.

À Monsieur Jean-François Dartigues

J'ai été très touché par vos mails exprimant votre intêret et votre enthousiasme pour nos travaux méthodologiques et leurs applications épidémiologiques. Pour avoir partagé mon enthousiasme et me faire l'honneur de participer à ce jury, soyez assuré de toute ma gratitude.

À Monsieur Daniel Commenges

Vous avez développé, formé et dirigé l'équipe Biostatistique qui m'a accueillie. Autant directement en répondant à mes nombreuses questions, qu'indirectement en ayant formé l'équipe qui m'a formé à son tour, j'ai beaucoup appris de vous. Pour m'avoir permis d'apprendre tant de choses et pour avoir accepté de présider mon jury, soyez assuré de toute ma gratitude.

À mes proches, à mes collègues et à tous les autres qui ont compté...

Aux collègues et amis

À Thomas A. Gerds, pour tant de choses que je ne peux ici énumérer sans m'éterniser... Pour tout ce que tu m'as appris, aussi bien en statistique théorique, en statistique appliquée, qu'en programmation R. Merci aussi pour tes nombreux conseils et encouragements, ainsi que pour avoir radicalement amélioré mon quotidien de thésard, en partageant tes connaissances des subtilités d'Emacs. Merci aussi pour les nombreuses sessions sports partagées, les soirées concerts/bières, et la découverte du sympathique sauna suivant la baignade dans la mer Baltique à 4 degrés...

À Cécile Proust-Lima, pour tes encouragements et ton soutien précieux. Ta curiosité et ton ouverture d'esprit rares m'ont si souvent aidé lors de nos nombreuses discussions passionnées. Merci pour ta gentillesse et ta disponibilité !

À Audrey Mauguen et Yassin Mazroui, mes collègues de bureau et amis thésards avec qui j'ai probablement le plus eu l'occasion de discuter de nos sujets de thèses respectifs. Sans nos discussions régulières, tant de choses me seraient restées incomprises... À travers nos discussions, qui m'ont si souvent aidé dans mes réflexions, cette thèse est aussi un peu la vôtre. En lisant entre les lignes, vous retrouverez, j'en suis sûre, par-ci par-là, l'influence certaine de nos belles discussions !

À Boris (aussi dit Bobo, ou "le lion"), à l'ami et au collègue statisticien passionné, si cultivé. Merci pour ton aide, et particulièrement pour notre partage des "astuces R". Merci aussi pour nos discussions à refaire le monde et la statistique, et pour tous les bons moments festifs ou sportifs, si "ambiancés" !

À Jérémie (aussi dit Didi), pour tous les bons moments passés ensemble, du stage à la fin de la thèse, au boulot, comme sur un terrain de sport, ou autour d'un petit café ou d'un bon Pago !

À Julie, mon amie et voisine de bureau, d'un grand soutien quotidien pour tant de choses... que je ne peux pourtant pas ici facilement citer... contrairement à ton aide "titanesque", à travers les si nombreuses relectures (de polys de cours et de thèse), bien qu'elles soient bien plus importantes ;-).

À Hind, tant pour ses déguisements et ses coiffures remarquables, que pour ses sourires rassurants ! Tes visites régulières au bureau 45 ont toujours apporté leur "petit brin de fantaisie" sympathique et vecteur d'une bonne humeur forte agréable !

À DD, leader du "BB-club" et des soirées animées ! ("*elle préfère l'amour en mer, c'est juste une question de...* " péniche ou de bateau corsaire... )

À Robin, mon ami "surfeur du dimanche", compagnon de natation et seul collègue appréciant Emacs à sa juste valeur ! Merci pour toutes ces discussions de Geek ! (`C-c C-e l-o`, `M-q`, `M-i`, `<Rg + tab`, ... on se comprend...)

À Pierre, pour ses visites distrayantes au bureau 45, son expérience inégalée des modèles de survie appliqués à Paquid, et surtout pour ces superbes barbecues enflammés !

Aux sympathiques biostats de l'USMR, à Julien, Anne-So, Hélène et Louise, avec qui c'est si plaisant de discuter aussi bien boulot que kitesurf ou vélo !

À Juan, Benjamin, Linda, Célia, et la p'tite Lucie, avec qui la fête est plus folle !

Aux "Paquidettes", Mélanie et Fanny, pour leur sympathie, leur disponibilité, et pour avoir partagé leurs connaissances uniques des données de Paquid et Trois-Cités.

À l'équipe du Mésocentre, notamment à Pierre et ses milliers de coeurs, sans qui mes études de simulations auraient souvent été 250 fois plus lentes...

À tous mes collègues de l'équipe Biostat, Benoit ("Ben", le surfeur), Karen (la "fan du petit Québec"), Loic (le jeune papa à l'accent amusant), Henry, Jérémie B., Mbéry, Réjane, Rodolphe (le "king of Tokyo") ...

À mes collègues de l'équipe enseignement, Fleur, Valérie, Marthe-Aline et Alioum, pour avoir partagé avec moi leur expérience.

Aux étudiants (danois et bordelais), à qui c'était si plaisant et distrayant de faire TD...

Aux non bordelais

À mes amis de Copenhague, notamment à Eleni, Pierro, Margarita et Ulla qui ont ensoleillé mes séjours danois.

À Vivian Viallon et Aurélien Latouche, pour m'avoir offert ma première collaboration, qui fut autant enrichissante que fort sympathique !

À tous ceux qui m'ont encouragé à faire, puis poursuivre, des études scientifiques, de la prépa à Caen au doctorat à Bordeaux, en passant par la fac de maths et l'ISUP. Merci à Stéphane ("L'Ancien") et Vic ("le Père Castor"), mes "ainés", qui m'ont fait découvrir les mathématiques et la culture scientifique au sens large. Merci aussi à Philippe Saint Pierre, réel pilier de la filière biostatistique de l'ISUP, pour ses enseignements et ses nombreux conseils avisés. À Sophie Tézenas du Montcel, qui a encadré mon premier stage à la Pitié-Salpêtrière et m'a encouragé à continuer les biostatistiques à Bordeaux.

À tous mes proches, ma famille et mes amis, qui ont partagé avec moi quelques concessions et sacrifices pour la réalisation de ce travail. Notamment à mes parents, grands-parents, ma soeur et mon beau-frérot, pour leurs encouragements et leur soutien. Mention spéciale pour Aurélien et Sandrine : merci de votre aide efficace pour la relecture !

Aux collègues et amis stagiaires de mes débuts Bordelais : "la petite Céline", Nuria, Pierre, Damien...

Enfin, aux milliers d'anonymes, développeurs bénévoles de logiciels libres d'une qualité remarquable, sans lesquels ce travail, comme tant d'autres, n'auraient jamais pu voir le jour. En particuliers aux développeurs de R, Linux, LaTeX, Emacs et ses puissantes extensions ESS et Orgmode.

Valorisations scientifiques

Publications

Publications principales issues de la thèse (en tant que principal auteur)

[1] P. Blanche, J-F Dartigues, H. Jacqmin-Gadda (2013). *Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring*, 55(5) :687–704. Biometrical Journal.

[2] P. Blanche, J-F Dartigues, H. Jacqmin-Gadda (2013). *Estimating and Comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks*, Statistics in Medicine, in press.

[3] P. Blanche, C. Proust-Lima, L. Loubère, C. Berr, J-F Dartigues, H. Jacqmin-Gadda (2013). *Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risk events*, submitted.

Autres publications sur la thèmatique de la thèse (en tant que coauteur)

[i] P. Blanche, A. Latouche, V. Viallon, *Time-dependent AUC with right-censored data : a survey study* (2013). to appear in *Risk Assessment and Evaluation of Predictions*, edited by M-L Lee, M. Gail, G. Satten, T. Cai, R. Pfeiffer and A. Gandy, Springer. (Preprint : `http://arxiv.org/abs/1210.6805`).

[ii] H. Jacqmin-Gadda, P. Blanche, E. Chary, C. Tourraine, J-F Dartigues (2013). *ROC curve estimation for time-to-event with semi-competing risks and interval censoring*, en révision mineure pour Statistical Methods in Medical Research.

[iii] H. Jacqmin-Gadda, P. Blanche, E. Chary, L. Loubère, H. Amieva, J-F Dartigues (2013). *Prognostic score to predict risk of dementia over 10 years accounting for death competing risks*, submitted.

[iv] M. Wolbers, P. Blanche, M. Koller, J. Witteman, T. Gerds, *Concordance for prognostic models with competing risks*, en révision pour Biostatistics.

Communications scientifiques

Présentations orales en conférence

- (Invité) P. Blanche, C. Proust-Lima, L. Loubère and H. Jacqmin-Gadda, *AUC for dynamic models*, Workshop Dynamic predictions for repeated markers and repeated events, Bordeaux, France, October 2013.
- P. Blanche, C. Proust-Lima, L. Loubère and H. Jacqmin-Gadda, *Estimating and comparing dynamic predictive accuracy of joint models for time-to-event and longitudinal data*, 4th Conference of the International Biometric Society Channel Network, St Andrews, Scotland, July 2013.
- H. Jacqmin-Gadda and P. Blanche, *ROC curve estimation for time-to-event with competing risks*, Atelier INSERM 223 : Évaluation des modèles prédictifs : adéquation aux observations et valeur prédictive, Bordeaux, France, Mai 2013.
- P. Joly, P. Blanche, C. Touraine, *Modèle de régression pour des probabilités cumulées en présence de risques concurrents et de censure par intervalles*, 45ème Journées de Statistique, Toulouze, Mai 2013.
- H. Jacqmin-Gadda, P. Blanche, E. Chary, *ROC curve estimation for time-to-event with competing risks under various censoring schemes*, Workshop Building and Evaluating Prognostic Models-Computational : Techniques and Strategies, Mainz, Germany, September 2012.
- P. Blanche, H. Jacqmin-Gadda, *Comparing areas under time-dependent ROC curves under competing risk*, 33st Annual conference of the International Society for Clinical Biostatistics, Bergen, Norway, August 2012.
- P. Blanche, H. Jacqmin-Gadda, *Estimating and comparing areas under time-dependent ROC curves in presence of censoring and competing risks*, Statistical Models and Methods for Reliability and Survival Analysis and Their Validation, Bordeaux, France, July 2012.
- E. Chary, P. Blanche, J-F. Dartigues, H. Jacqmin-Gadda, *Prédiction de la démence à 10 ans à partir de test neuropsychologiques*, Réunion francophone sur la maladie d'Alzheimer et des syndromes apparentés, Toulouse, France, mai 2012.
- P. Blanche, H. Jacqmin-Gadda, *Courbes ROC pour données censurées : comparaison de méthodes proposées et alternative*, Journées du GDR Statistique et Santé, Paris, France, Mai 2011.
- H. Jacqmin-Gadda, P. Blanche, C. Proust-Lima, *Evaluating discrimination abilities of a joint model for time-to-event and longitudinal marker*, International Biometric Conference, Florianapolis, Brazil, December 2010.

Présentation affichée en conférence

- P. Blanche, H. Jacqmin-Gadda, *Comparison of different estimators for time-dependent ROC curve*, 3rd Conference of the International Biometric Society Channel Network, Bordeaux, France, April 2011.

Communications invitées en séminaire

- P. Blanche, C. Proust-Lima, L. Loubère and H. Jacqmin-Gadda, *Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risk events*, Seminar of the Department of Biostatistics, University of Copenhagen, October 2013.
- P. Blanche, H. Jacqmin-Gadda, *Comparing time-dependent ROC curves for censored event times with competing risks*, Seminar of the Department of Biostatistics, University of Copenhagen, May 2012.
- P. Blanche, H. Jacqmin-Gadda, *Estimateurs de la courbe ROC pour un évènement dépendant du temps en présence de censure dépendant du marqueur*, Séminaire de l'équipe de Biostatistique, Pharmacoépidémiologie, et Mesures Subjectives en Santé, Université de Nantes, Novembre 2011.

Récompense scientifique

- Prix de la meilleure présentation orale étudiante, 4th Conference of the International Biometric Society Channel Network, St Andrews, Scotland, July 2013.

Table des matières

I. Introduction

I.1 À propos de la démence

Cette thèse étant un travail de biostatistique appliquée, nous commençons tout d'abord par décrire quelques éléments du contexte qui l'a motivée.

I.1.1 La démence : définition et contexte actuel

La démence est souvent définie par une altération de la mémoire et d'au moins une autre fonction cognitive avec des répercussions sur la vie quotidienne. Cliniquement, le diagnostic de démence est plus précisément encadré par le *Diagnostic and Statistical Manual of Mental Disorders*, publié par l'*American Psychiatric Association*, qui reprend ce concept avec plus de précisions. La démence est une maladie chronique, pour laquelle aucun traitement satisfaisant n'existe et dont on ne guérit pas. Bien que l'évolution diffère d'un patient à l'autre, la maladie finit par avoir un impact très important sur l'état de santé général du patient. Progressivement, il devient totalement dépendant en raison de la perte de la capacité à s'habiller, à se laver et à aller aux toilettes. La maladie crée aussi un état de faiblesse immunitaire et des complications infectieuses sont d'ailleurs souvent à l'origine du décès.

Aujourd'hui, on estime que 38% des femmes et 24% des hommes âgés de plus de 85 ans seraient atteints de démence (Gallez, 2005). Avec l'allongement de l'espérance de vie, on estime aussi que le nombre de sujets déments pourrait augmenter de 75% d'ici 2030 (Jacqmin-Gadda *et al.*, 2013a). La démence est donc actuellement un problème de santé publique majeur. Parmi les diverses formes de démence existantes, la maladie d'Alzheimer est la plus courante (environ 60 à 70% des démences).

En partie de ce constat, le *Plan Alzheimer* a été lancé en France en 2008, sur décision du Président de la République. Centré sur la personne malade et sa famille, il avait notamment pour objectif « de fournir un effort sans précédent sur la recherche » et de « favoriser un diagnostic plus précoce ».

I.1.2 Traitements préventifs et démence

Alors que la recherche se focalisait encore récemment sur le traitement de la maladie d'Alzheimer une fois diagnostiquée, elle semble maintenant essentiellement s'orienter vers les traitements préventifs, à administrer avant le diagnostic de démence, et vers les programmes de prévention (Aisen *et al.*, 2011; Alzheimer's Association, 2012).

Par exemple, des essais cliniques se sont récemment intéressés à l'effet préventif d'extraits de Ginkgo Biloba (arbre du sud-est de la Chine) sur le risque de démence (Snitz *et al.*, 2009; Andrieu *et al.*, 2008). Bien que les résultats des essais furent négatifs, des résultats d'analyses secondaires suggèrent que le Ginkgo Biloba pourrait, s'il est consommé sur le long terme, réduire le risque de démence.

En supposant un effet (modeste) des traitements préventifs de la démence, la courte durée des essais (5 ans) et la faible incidence de la démence dans les échantillons étudiés pourraient rendre la puissance des essais cliniques faible. Cela pourrait en partie expliquer les précédents résultats négatifs (Vellas *et al.*, 2011). Réfléchissant aux problématiques des essais cliniques préventifs sur la démence, un groupe de réflexion (« task-force ») américano-européen a récemment proposé des recommandations pour la planification des prochains essais (Vellas *et al.*, 2011). Parmi elles, la première est d'« enrichir la population de l'étude de sujets à fort risque de décliner pendant la période de l'étude » pour à la fois augmenter la puissance de l'étude et pouvoir réduire sa durée.

Ainsi, pour améliorer la puissance d'essais cliniques préventifs, mais aussi pour améliorer la prise en charge médicale (médicamenteuse ou non) de sujets déclinant vers une démence, il apparaît aujourd'hui essentiel de pouvoir cibler les populations à haut risque de démence.

I.1.3 Populations à haut risque de démence

Aujourd'hui, les pistes probablement les plus prometteuses pour définir une population à haut risque de démence semblent être l'exploitation de la connaissance des facteurs de risques connus, l'utilisation de biomarqueurs et l'utilisation de tests psychométriques.

I.1.3.1 Facteurs de risque de la démence

Aujourd'hui, quelques facteurs de risque de la démence semblent être clairement identifiés (Alzheimer's Association, 2012). Le premier d'entre eux est l'âge, bien que la démence ne fasse pas partie du processus de vieillissement normal. L'histoire familiale, via la génétique, et particulièrement le gène codant pour l'apolipoprotéine E (apoE) serait aussi associée au risque de démence.

Les facteurs de risque vasculaires et le bas niveau d'étude seraient également associés à un plus fort risque de démence. L'hypothèse d'une « réserve cognitive » est parfois évoquée pour expliquer l'association entre le niveau d'étude et la démence. Le régime alimentaire, l'intensité des activités sociales, les antécédents de traumatismes crâniens ou le sexe sont aussi suspectés d'être associés au risque de démence, bien que la communauté scientifique soit encore partagée à leur sujet.

I.1.3.2 Biomarqueurs et imagerie médicale

Les biomarqueurs, et notamment ceux permettant de mesurer l'accumulation de β-amyloid (une peptide néfaste pour le système nerveux) ou l'état de dégradation de cellules nerveuses, semblent riches de potentiel. Aujourd'hui, on estime qu'ils pourraient aussi être envisagés pour prédire une démence (Alzheimer's Association, 2012).

L'imagerie cérébrale par résonance magnétique (IRM) pourrait aussi se révéler utile (The 3C Study Group, 2003). En effet, elle permet d'observer les différentes substances du cerveau (substance blanche et substance ou matière grise) ainsi que les dégâts vasculaires qui sont associés au risque de démence.

I.1.3.3 Tests psychométriques

Les tests psychométriques qui permettent de mesurer un niveau cognitif pourraient aussi s'avérer fort utiles pour identifier des sujets à risque de démence. En utilisant de tels tests, Amieva *et al.* (2005, 2008) ont notamment montré que la cognition des futurs déments se

distinguait significativement de celle des sujets ne développant pas de démence longtemps avant le diagnostic (jusqu'à 12 ans avant). Bien que d'autres études soient requises pour les confirmer et les compléter, ces résultats pourraient se révéler très utiles. En effet, faire passer un test cognitif à un sujet est à la fois non-invasif, peu onéreux, rapide et peu contraignant. Distinguant des différences d'évolution, les résultats de Amieva *et al.* (2005, 2008) suggèrent aussi que des mesures répétées de ces tests, qui permettent d'estimer une évolution du niveau cognitif du sujet, pourraient aussi être plus prédictives qu'une unique mesure.

I.1.3.4 Évaluer les capacités prédictives et combiner les connaissances : l'utilisation des données et de la statistique

Chaque facteur de risque, biomarqueur ou test psychométrique est indépendamment plus utile pour prédire une démence que de tirer au sort, à pile ou face, en prédisant une démence par l'obtention d'un *pile* et en prédisant son contraire par un *face*. Cependant, on peut s'interroger sur leurs capacités pronostiques et se demander : A quel point prédisent-ils « bien » la démence ? La question est légitime puisqu'en épidémiologie il est bien connu qu'une seule variable, quelle que soit sa nature, ne prédit que très rarement efficacement un événement clinique (Pepe *et al.*, 2004).

Avant d'envisager l'utilisation clinique de prédicteurs de la démence, il est alors essentiel de quantifier leurs capacités prédictives (Stephan *et al.*, 2010). Par ailleurs, pour prédire la démence, comme pour prédire des risques cardiovasculaires ou tout autre événement clinique, la combinaison de différentes sources d'informations et l'utilisation de modèles permet généralement d'augmenter considérablement les capacités de prédiction.

Pour construire et évaluer des modèles de prédiction, l'utilisation de données de cohortes au moyen de méthodes statistiques appropriées représente alors une opportunité intéressante qui a déjà connu plusieurs succès. Par exemple, les données de la cohorte de Framingham ont permis de construire et de valider des modèles de prédiction d'événements cardiovaculaires aujourd'hui populaires (Wilson *et al.*, 1998). A l'origine de ce travail de thèse, il a donc été supposé que, de la même façon, les données de cohortes de personnes âgées sont potentiellement riches d'intérêt

pour l'étude de la prédiction de la démence.

I.2 Les cohortes Paquid et Trois-Cités

Permettant à la fois (i) de construire des modèles de prédictions et (ii) d'évaluer des capacités prédictives, les données de cohortes qui ont motivées ce travail de thèse, et dont les applications sont issues, sont brièvement présentées dans cette section.

Les applications des travaux méthodologiques de cette thèse sont essentiellement basées sur les données de l'étude Paquid (Chapitres III, IV et V). Les données de l'étude des Trois-Cités seront cependant aussi utilisées au Chapitre V, comme données de validation externe, pour comparer des modèles pronostiques estimés sur les données de la cohorte Paquid.

I.2.1 Paquid

Paquid (pour "*Personnes âgées Quid ?*") est l'une des premières cohortes européennes visant à étudier le vieillissement cognitif normal et pathologique des personnes âgées. En particulier, l'identification de facteurs de risque et de manifestations pré-cliniques de la maladie d'Alzheimer fait partie des objectifs principaux initiaux (Dartigues *et al.*, 1992).
Paquid est une étude de cohorte prospective incluant 3 777 personnes âgées de 65 ans ou plus en 1988. Les sujets inclus ont été tirés au sort sur les listes électorales dans le but de constituer un échantillon représentatif de la population générale.
Après une visite initiale, les sujets sont revus approximativement 1, 3, 5, 8, 10, 13, 15, 17 et 20 ans après par des psychologues. A la visite initiale comme aux suivantes, une multitude d'informations sont recueillies. Notamment, les scores des sujets à de nombreux tests cognitifs sont enregistrés, permettant, entre autres, d'estimer des dynamiques d'évolutions cognitives (Amieva *et al.*, 2005, 2008; Jacqmin-Gadda *et al.*, 1997; Proust-Lima *et al.*, 2007; Dantan *et al.*, 2011). Parmi eux, trois tests cognitifs présentés en Section I.2.3 ont été étudiés dans les diverses applications de cette thèse.
A chaque visite, un dépistage de démence est aussi réalisé par un psychologue. En cas de

suspicion de démence, le diagnostic final est établi par un neurologue lors d'un examen clinique. Enfin, les informations relatives aux âges de décès des sujets sont recueillies en contactant les médecins traitants ou les familles.

I.2.2 Trois-Cités

L'étude des Trois-Cités (3C) est également une étude de cohorte prospective visant à étudier la maladie d'Alzheimer. En particulier, l'un des objectifs initiaux était l'identification de personnes à haut risque de développer cette maladie pour pouvoir leur proposer des mesures de prévention, si on en disposait. Lancée en 1999, cette étude inclut 9 294 personnes âgées de 65 ans ou plus qui ont été recrutées par tirage au sort sur les listes électorales des villes de Bordeaux, Dijon et Montpellier, d'où le nom Trois-Cités. (The 3C Study Group, 2003).

Tous les deux ans et comme pour les sujets de Paquid, les sujets de l'étude des Trois-Cités ont été vus par des enquêteurs, ont passé des tests cognitifs et ont été diagnostiqués pour la démence. Les informations relatives aux âges et causes de décès ont également été recueillies en contactant les médecins traitants, les centres hospitaliers ou les proches. Par ailleurs, des données de neuro-imagerie (inexploitées dans cette thèse) ont aussi été recueillies, notamment pour étudier les liens entre défaut de vascularisation cérébrale et maladie d'Alzheimer.

Bien que très comparable à l'étude Paquid, notons cependant que la population de cette étude diffère quelque peu de celle de Paquid, du fait que l'étude soit plus récente (1999 versus 1988) et qu'elle n'inclut que des sujets vivant dans de larges agglomérations (Dijon, Montpellier et Bordeaux versus l'ensemble de la Dordogne et de la Gironde, milieu rural compris). Une conséquence notable est donc que le niveau d'étude des sujets, qui est fortement associé au niveau cognitif, est plus élevé dans l'étude des Trois-Cités que dans l'étude Paquid.

Enfin, signalons qu'une documentation détaillée de cette étude et des publications associées est disponible à l'adresse : `http://www.three-city-study.com/`.

I.2.3 Tests psychométriques utilisés dans nos applications

Le *Mini-Mental State Examination*, le *Test de fluence verbal d'Isaacs* et le *Digit symbol substitution test* sont les trois tests psychométriques qui sont étudiés dans les applications des méthodologies présentées dans cette thèse. Ils ont tous les trois été passés par les sujets de l'étude Paquid. Cependant, seuls les deux premiers ont été proposés dans l'étude des Trois-Cités.

I.2.3.1 Mini-Mental State Examination

Le "Mini Mental State Examination" (MMSE), proposé par Folstein *et al.* (1975) est un test assez complet évaluant simultanément plusieurs dimensions de la cognition, dont la mémoire, le calcul, l'orientation dans le temps et l'espace, et le langage. Actuellement, c'est le test cognitif le plus utilisé pour l'évaluation de troubles cognitifs et le dépistage de la démence des personnes âgées. Le test consiste à répondre à 30 questions. Attribuant un point par réponse juste, le score de ce test varie de 0 à 30.

Bien qu'extrêmement intéressant comme indice global de la cognition, il est cependant connu pour souffrir d'un effet plafond et est peu sensible aux changements de niveaux cognitifs des sujets à haut niveau cognitif.

I.2.3.2 Test de fluence verbal d'Isaacs

Le test de fluence verbal d'Isaacs et Kennie (1973) évalue la fluidité et la vitesse de production verbale. Il consiste à demander à un sujet de donner le plus de mots possibles appartenant à une catégorie sémantique particulière en une minute. Quatre catégories sémantiques sont considérées : les villes, les fruits, les animaux et les couleurs. Dans nos analyses, comme dans la plupart des précédentes réalisées sur ce test avec les données de Paquid, on s'intéresse à la somme des scores par catégorie à 15 secondes, avec censure à 10 mots par catégorie. Attribuant un point par mot juste, le score varie alors de 0 à 40.

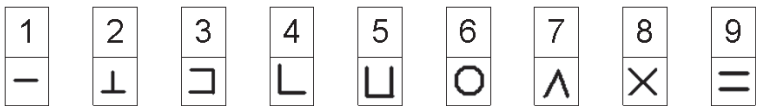
Les résultats de Proust-Lima *et al.* (2007) suggèrent que ce test serait relativement sensible aux petites variations de cognition, et ceci quel que soit le niveau cognitif des sujets. En effet,

il ne présente pas d'effet plafond ni plancher. Par ailleurs, Amieva *et al.* (2008) suggèrent que c'est l'un des tests cognitifs pour lequel l'évolution moyenne différerait le plus précocement entre les futurs déments et les autres.

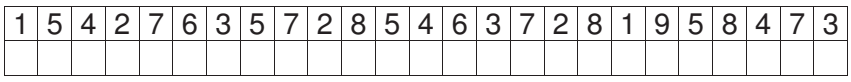
I.2.3.3 Digit symbol substitution test

Le « Digit Symbol Substitution Test » (DSST) proposé par Wechsler (1981) mesure l'attention et la vitesse psychomotrice. Comme décrit à la Figure I.1, étant donnée une grille de correspondance appariant des chiffres et des symboles, le test consiste à remplir des cases blanches avec le symbole correspondant au chiffre au-dessus de la case. Un sujet passant le test doit remplir le plus de cases possibles en 90 secondes. Attribuant un point par symbole correctement substitué, ce score varie de 0 à 90.

Les résultats de Proust-Lima *et al.* (2007) suggèrent que, pour les sujets de niveau cognitif moyen à haut, ce test serait relativement sensible et approprié à l'identification de petits changements de cognition.



(a) Code de substitution



(b) Portion d'une grille de test à remplir par le sujet testé

FIGURE I.1 – Test de substitution de symbole de Wechsler (1981)

I.3 Objectifs de la thèse

I.3.1 Motivation épidémiologique

La prédiction de la démence est compliquée car la maladie est encore peu connue. Pourtant, elle apparaît aujourd'hui comme une opportunité de premier plan pour améliorer la prise en charge des personnes âgées susceptibles de la développer. C'est pourquoi elle mérite d'être étudiée.

Parmi les prédicteurs potentiels de la démence, les tests psychométriques sont les plus intéressants car rapides, peu coûteux et non invasifs. De plus, les travaux de Amieva *et al.* (2008, 2005) ont montré une différence d'évolution de scores cognitifs longtemps avant l'apparition de la démence. Les capacités pronostiques d'un unique test cognitif apparaissent donc intéressantes à évaluer. Par ailleurs, il serait dommage de ne pas utiliser l'âge et le niveau d'études qui sont des facteurs de risques majeurs de la démence et qui sont toujours accessibles au médecin traitant. La pertinence d'un score pronostique, combinant à la fois une ou plusieurs mesures de tests psychométriques et ces facteurs de risques, apparaît ainsi également intéressante à investiguer.

Pour évaluer de tels outils pronostiques au moyen de données de cohortes, les sorties d'étude qui induisent un phénomène de censure des données devront être prises en compte par les méthodes statistiques d'estimation. Par ailleurs, la prédiction de la démence s'effectuant pour des personnes âgées, son évaluation devra tenir compte du risque concurrent non négligeable de décès sans démence.

D'un point de vue épidémiologique, ce travail de thèse souhaite donc contribuer à l'étude des capacités pronostiques de prédicteurs de la démence construits à partir de tests psychométriques, en tenant compte des phénomènes de censure et du risque concurrent de décès sans démence.

I.3.2 Objectif

Cette thèse a pour ambition de contribuer au développement de méthodes d'évaluation de capacités prédictives en présence de données censurées, de risques concurrents et de marqueurs

longitudinaux.

Plus précisément, l'objectif est de proposer des méthodes d'inférence de capacités pronostiques, incluant l'estimation ponctuelle et la construction de régions de confiance et tests statistiques, permettant notamment de faire des comparaisons d'outils prédictifs. Une place centrale sera accordée aux méthodes d'inférence pour les courbes ROC (« Receiving Operating Characteristic »).

C'est avant tout un travail de méthodologie biostatistique, bien que l'application à la prédiction de la démence soit omniprésente et à l'origine de tous les développements proposés.

I.3.3 Plan du mémoire

Ce travail s'inscrivant dans un contexte déjà riche de connaissances sur le sujet, une tentative de résumé de l'état de l'art est présentée au Chapitre II : *État de l'art*. Notamment, on y rappelle les particularités des données censurées et des risques concurrents. On discute ensuite succinctement les modèles de prédiction puis on rappelle et discute les différentes possibilités pour évaluer des outils prédictifs. On présente notamment les courbes ROC sur lesquelles porte la majeure partie du travail de thèse. Enfin, on termine l'état des connaissances en présentant brièvement les principales méthodes d'estimation adaptées aux données censurées, ainsi que quelques outils de statistique asymptotique utilisés pour dériver nos développements.

On présente ensuite notre travail de thèse, dont l'objectif principal était les développements méthodologiques pour les courbes ROC. Trois chapitres sont présentés, chacun ayant pour objectif de contribuer à l'une des trois thématiques suivantes :

- Chapitre III : *Courbe ROC dépendant du temps et censure dépendante de l'outil prédictif à évaluer*
- Chapitre IV : *Courbe ROC dépendant du temps et risques concurrents*
- Chapitre V : *Comparaisons de prédictions dynamiques basées sur des mesures répétées d'un marqueur*

Ces trois thématiques sont présentées sous la forme d'un article principal publié ou soumis à publication dans une revue internationale à comité de lecture, ainsi que de quelques complé-

ments.

Une discussion générale résumant les travaux et ouvrant vers de nouvelles perspectives est ensuite présentée au Chapitre VI, avant la bibliographie et une annexe qui clôt ce manuscrit.

II. Etat de l'art

II.1 Les données de survie : la censure et les risques concurrents

II.1.1 La censure indépendante et le risque instantané

En étudiant des durées, entre une entrée dans une étude clinique et un décès par exemple, on se trouve souvent confronté au problème de la censure, induit par exemple par des sorties d'étude. On dit qu'il y a un phénomène de censure (à droite), et qu'on est en présence de données censurées (à droite) lorsqu'on souhaite étudier un temps de survie T mais qu'on observe seulement le couple

$$(\widetilde{T}, \Delta) \qquad \text{avec} \quad \begin{cases} \widetilde{T} = \min(T, C) \\ \Delta = \mathbb{1}_{(T \leqslant C)} \end{cases},$$

moins informatif que la variable T. Le temps C est alors appelé temps de censure.

Depuis le début du XXème siècle, de nombreuses méthodes d'inférence statistique se sont intéressées à de telles données, donnant naissance à la branche de la statistique dite de l'« analyse de survie ». A notre connaissance, elles reposent toutes sur la notion de censure *indépendante*. Bien que le mot soit le même, *indépendant* en analyse de survie ne signifie cependant pas *indépendant* au sens probabiliste. On discute donc ici brièvement et informellement la notion de censure indépendante.

Au sens probabiliste, un temps de censure C est dit indépendant s'il n'influence pas le temps de survie T. Plus précisément, au sens probabiliste on dit que deux variables aléatoires réelles T et C sont indépendantes si pour tout intervalle I_i et I_j de $\mathbb{R}$, on a $\mathbb{P}(T \in I_i, C \in I_j) = \mathbb{P}(T \in I_i)\mathbb{P}(C \in I_j)$, ou de manière équivalente, $\mathbb{P}(T \in I_i | C \in I_j) = \mathbb{P}(T \in I_i)$. L'indépendance telle qu'on l'entend le plus souvent au sens de l'analyse de survie est une condition qui peut être considérée comme bien plus faible mais qui est aussi bien plus complexe.

Initialement proposée dans la thèse de Odd Aalen (Aalen, 1975), la définition d'une censure indépendante actuellement considérée comme la plus satisfaisante résulte de l'émergence de l'élégante théorie des martingales et de son application en analyse de survie (Aalen *et al.*, 2009). Elle est liée aux propriétés des martingales car elle est définie conditionnellement au

passé. Sa définition rigoureuse est donnée par Andersen *et al.* (1993, Définition III.2.1) et Aalen *et al.* (2008, Sec. 2.2.8). Elle est ici omise, car nécessitant l'introduction de quelques concepts et notations mathématiques assez techniques (filtrations, martingales, décomposition de Doob-Meyer et théorème d'innovation).

Heuristiquement, on peut cependant définir une censure comme *indépendante* au sens de l'analyse de survie si, sachant qu'une personne est en vie au temps t (i.e. $T > t$), l'information additionnelle qu'une personne soit non censurée ne change pas son risque instantané de décès (Andersen et Keiding, 2012). Le sens des mots « additionnelle » et « risque instantané » est alors important à clarifier. Un risque instantané conditionnellement à des covariables $X = (X_1, \ldots, X_p)$, est défini par

$$\lambda(t|X) = \lim_{dt \downarrow 0} \frac{1}{dt} \mathbb{P}(t \leqslant T < t + dt | T \geqslant t, X). \tag{II.1}$$

Essentiellement, l'information « additionnelle » signifie que l'information est additionnelle à ce qui est connu et utilisé pour modéliser le risque instantané de décès au temps t, i.e. *au passé*. Grossièrement il s'agit de toute information complémentaire à $T > t$ et aux covariables X. C'est cette notion de conditionnement sur le *passé*, intrinsèque aux propriétés des martingales, qui est rigoureusement définie par Andersen *et al.* (1993, Définition III.2.1) et Aalen *et al.* (2008, Sec. 2.2.8).

Une censure indépendante au sens probabiliste est donc indépendante au sens de l'analyse de survie. Cependant, une censure peut être non indépendante au sens probabiliste bien qu'indépendante au sens de l'analyse de survie. Par exemple, c'est le cas lorsque la censure C et le temps de survie T sont tous les deux corrélés à des covariables communes $X = (X_1, \ldots, X_p)$ considérées dans le risque instantané, mais que T et C sont indépendants conditionnellement à X. La censure ne sera néanmoins pas indépendante si l'une des covariables $X_1, \ldots, X_p$ n'est pas incluse dans le modèle (Andersen *et al.*, 1993, Exemple III.2.9). Par exemple, la censure peut donc être différente dans les bras traité et placebo d'un essai clinique, lorsqu'on s'intéresse à modéliser l'effet d'un traitement.

D'un certain point de vue, la notion de censure indépendante en survie est assez similaire

aux notions de biais de confusion intrinsèques à l'épidémiologie, ou aux notions de données manquantes aléatoires, dites « données MAR », pour « Missing At Random » (Hedeker et Gibbons, 2006, Chap. 14), qui sont omniprésentes en modélisation de données répétées. Contrairement à la notion de censure indépendante au sens probabiliste, ces dernières dépendent toutes les trois à la fois de la nature des données mais aussi des modèles que l'on souhaite estimer.

Pour des raisons techniques et sans perte de généralités importantes, dans les Chapitres III, IV et V, on ne parlera d'indépendance ou d'indépendance conditionnelle qu'au sens probabiliste.

II.1.2 Le risque instantané, quantité clé en analyse de survie

Diverses quantités peuvent être utilisées pour analyser les propriétés d'un temps de survie. Si la censure indépendante est définie en fonction du risque instantané de l'équation (II.1), c'est parce que ces quantités peuvent souvent en être dérivées. En particulier, la fonction de survie de T conditionnellement à X, définie par $S(t|X) = \mathbb{P}(S > t|X)$, peut se réécrire comme

$$S(t|X) = \exp\left\{-\int_0^t \lambda(u|X)du\right\} = \exp\Big\{-\Lambda(t|X)\Big\},$$

où $\Lambda(t|X) = \int_0^t \lambda(u|X)du$ est appelé le risque cumulé. Réciproquement, on notera que le risque instantané peut s'écrire en fonction de la survie via

$$\lambda(t|X) = -\frac{\partial}{\partial t}\log\Big\{S(t|X)\Big\}. \tag{II.2}$$

La densité de probabilité en t, notée $f(t|X)$, peut aussi se réécrire comme

$$f(t|X) = -\frac{\partial}{\partial t}S(t|X) = \lambda(t|X)\exp\left\{-\int_0^t \lambda(u|X)du\right\},$$

et pour toute fonction $\varphi(\cdot)$ d'intérêt, on a alors aussi

$$\begin{aligned}\mathbb{E}\Big[\varphi(T)\Big|X\Big] &= \int_0^\infty \varphi(u)f(u|X)du \\ &= \int_0^\infty \varphi(u)\lambda(u|X)\exp\left\{-\int_0^u \lambda(s|X)ds\right\}du.\end{aligned}$$

Ainsi, en présence d'une censure indépendante qui essentiellement assure que $\lambda(t|X)$ peut être estimé sans biais, les nombreuses quantités d'intérêt qui en sont dérivées peuvent également être estimées sans biais en utilisant ces relations.

II.1.3 Risques concurrents

La situation « classique » en survie considère qu'un sujet est susceptible de subir un unique événement (appelé décès). La situation dite de « risques concurrents » (aussi dite de « risques compétitifs ») considère quant à elle qu'un sujet est à risque de plusieurs événements différents, mutuellement exclusifs. Par exemple si l'on étudiait une population de fumeurs, on pourrait considérer qu'un sujet peut subir trois risques concurrents : un décès par infarctus, un décès par cancer du poumon ou un décès dû à une autre cause. Ces deux situations de survie « classique » et de risques concurrents, ainsi que d'autres plus générales, peuvent être modélisées par des modèles multi-états (Andersen et Keiding, 2002), comme l'illustrent les Figures II.1 et II.2.

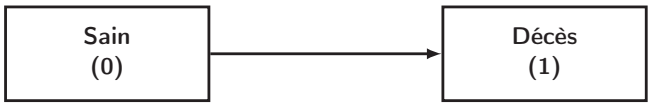


FIGURE II.1 – La survie « classique » : un modèle à deux états.

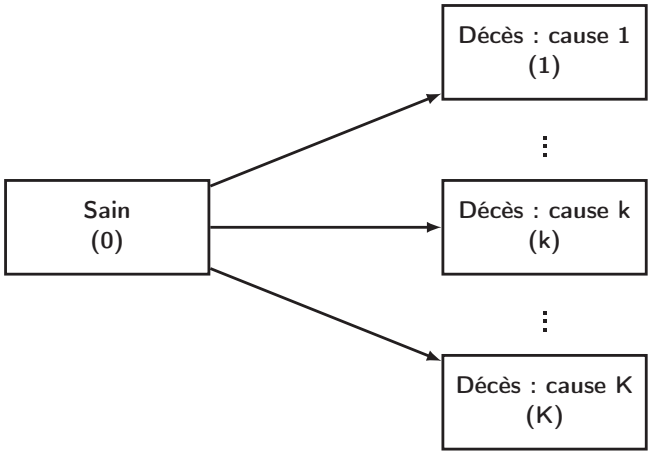


FIGURE II.2 – Les risques concurrents.

Un autre exemple, celui qui nous intéressera principalement dans les Chapitres IV et V, considère qu'un sujet sain peut subir les deux événements concurrents de démence et de décès sans démence (Figure II.3). Deux différences majeures entre la notion de risque concurrent et la notion de censure sont :

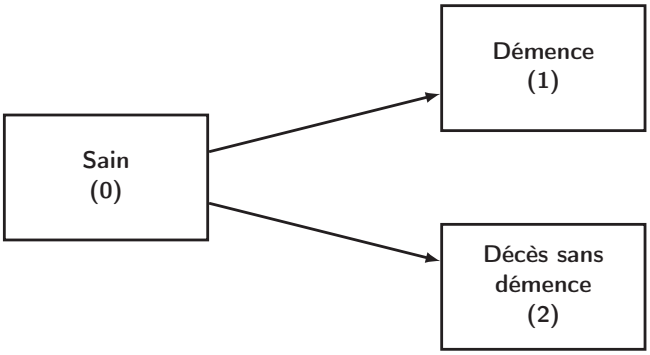


FIGURE II.3 – Les risques concurrents de démence et décès sans démence.

(i) Les risques concurrents correspondent à des phénomènes réels (que l'on souhaite étudier ou au moins prendre en compte dans les définitions des quantités d'intérêt), et non pas à des phénomènes de nuisances liés au schéma d'observation des données.

(ii) Les risques concurrents sont mutuellement exclusifs : un sujet ne peut par définition jamais expérimenter deux risques concurrents. Par contraste, il est supposé qu'un sujet censuré finit toujours par mourrir après son temps de censure, bien que l'on n'observe pas cette donnée.

Pour modéliser les risques concurrents, on utilise généralement un vecteur (T, η), composé d'un temps d'évènement T, et d'un type (ou cause) d'événement η, avec $\eta \in \{1, \ldots, K\}$, si l'on suppose K types d'événement. Comme en survie « classique », en pratique on observe souvent des données censurées (généralement à droite). On observe alors

$$(\widetilde{T}, \widetilde{\eta}) \qquad \text{avec} \quad \begin{cases} \widetilde{T} = \min(T, C) \\ \widetilde{\eta} = \Delta\eta = \eta \mathbb{1}_{(T \leqslant C)} \end{cases},$$

moins informatif que le couple (T, η). D'autres modélisations des risques compétitifs ont aussi été proposées, comme celle des « temps latents », bien qu'elle soit critiquée depuis longtemps (Tsiatis, 1975) et qu'elle soit aujourd'hui considérée comme peu appropriée en biostatistique (Andersen et Keiding, 2012).

Dans la situation des risques concurrents, le risque instantané de l'équation (II.1) peut se

décomposer en la somme de risques spécifiques à chaque cause (« cause specific hazard »),

$$\lambda(t|X) = \sum_{k=1}^{K} \lambda_k(t|X),$$

avec

$$\lambda_k(t|X) = \lim_{dt \downarrow 0} \frac{1}{dt} \mathbb{P}(t \leqslant T < t + dt, \eta = k | T \geqslant t, X). \tag{II.3}$$

Lorsque l'on s'intéresse à des méthodes de régression du risque spécifique à la cause k, i.e. $\lambda_k(t|X)$, la situation des risques concurrents peut être rapportée à la situation classique sans risque concurrent (Andersen *et al.*, 2012). En effet, pour des raisons de factorisation de la vraisemblance, l'estimation peut être réalisée en considérant les données des sujets subissant les événements concurrents comme censurées (Kalbfleisch et Prentice, 2002, Section 8.2.3). Heuristiquement, on dit que cela est possible car les risques concurrents n'influencent pas le risque à « court terme » de l'événement k, i.e. le risque instantané $\lambda_k(t|X)$. Cependant, le risque à « long terme », c'est-à-dire la probabilité de subir l'événement k dans les t années à venir, appelée l'incidence cumulée et notée $F_k(t|X)$, est quant à elle influencée par les risques concurrents (Putter *et al.*, 2007; Andersen *et al.*, 2012). En effet, la fonction d'incidence cumulée d'une cause k en t est :

$$\begin{aligned} F_k(t|X) &= \mathbb{P}(T \leqslant t, \eta = k | X) \\ &= \int_0^t \lambda_k(u|X) S(u|X) du \\ &= \int_0^t \lambda_k(u|X) \exp\left(- \sum_{j=1}^{K} \int_0^u \lambda_j(s|X) ds \right) du. \end{aligned} \tag{II.4}$$

Le risque à long terme $F_k(t|X)$ est donc impacté par tous les risques causes spécifiques, via le terme $S(u|X)$ dans l'intégrale de l'équation (II.4). L'interprétation naturelle étant que pour subir l'événement k, il faut ne pas avoir subi l'un des $K-1$ événements concurrents avant. En particulier, cette relation montre que même si la covariable X ne modifie pas le risque $\lambda_k(u|X)$, elle peut, en modifiant l'un des $\lambda_j(u|X)$, $j \neq k$, modifier $F_k(t|X)$, i.e. le risque à long terme. Considérons l'exemple de la démence de la Figure II.3, avec $\eta \in \{1, 2\}$, $\eta = 1$ pour la démence et $\eta = 2$ pour le décès sans démence, où X représente la variable binaire indicatrice du sexe.

Même si le risque instantané de démence $\lambda_1(u|X)$ ne dépend pas du sexe, $F_1(t|X)$ en dépend fortement, puisque $\lambda_2(u|X)$ en dépend (les femmes ayant une espérance de vie plus longue).

Dans certains cas, le sens de l'effet d'une variable sur le risque à court terme et le risque à long terme d'un événement peut même s'inverser. Notons que cette relation non monotone, entre risques à court et long terme, n'a pas lieu en l'absence de risques concurrents dès lors que l'effet de la variable est supposé indépendant du temps (i.e. avec un modèle de Cox (1972) par exemple), comme le rappelle l'équation (II.2) page 17. C'est l'une des spécificités principales des risques concurrents qui a donné lieu à une multitude de discussions et de travaux (Gray (1988); Fine et Gray (1999); Andersen et Keiding (2012); Gerds *et al.* (2012), entre autres).

II.2 Modèles pronostiques

II.2.1 Généralités

Divers et variés, les modèles pronostiques connaissent aujourd'hui un intérêt croissant en statistique médicale (Steyerberg, 2009). Dans cette thèse, on s'intéresse à ceux ayant pour objectif de prédire un évènement binaire D (noté $D = 1$ s'il a lieu, $D = 0$ sinon) en fonction d'une information disponible, généralement modélisée par un vecteur de covariables X. L'événement D peut éventuellement être défini en fonction d'un temps de survie T et d'un horizon de prédiction t, par exemple si $D \equiv D(t) = \mathbb{1}_{(T \leqslant t)}$. En présence de risques concurrents, la cause de décès peut aussi être utilisée pour le définir, par exemple si $D \equiv D(t) = \mathbb{1}_{(T \leqslant t, \eta = 1)}$.

Contrairement aux modèles d'association, explicatifs, étiologiques ou causaux, qui sont utilisés pour quantifier et tester des effets de covariables X sur D, les modèles pronostiques n'ont pour ambition que d'estimer la probabilité $\mathbb{P}(D = 1|X)$. Ils ne sont donc pas, en première intention du moins, limités par des difficultés de modélisation comme celles liées à l'interprétation du modèle, de ses paramètres ou aux notions de confusion par exemple.

C'est pourquoi, par exemple, en présence de risques concurrents, avec $D \equiv D(t) = \mathbb{1}_{(T \leqslant t, \eta = 1)}$, les approches modélisant directement la probabilité de l'événement $D(t) = 1$ sachant X, i.e. l'incidence cumulée $F_1(t|X)$, comme celles de Fine et Gray (1999) ou Scheike

et al. (2008), sont souvent préférées à des modèles combinant des modèles pour les hasards spécifiques à chaque cause, pour faire de la prédiction (Andersen et Keiding, 2012). En effet, bien que peu interprétables d'un point de vue étiologique, elles ont entre autres l'avantage d'être plus parcimonieuses et de supposer moins d'hypothèses.

Dans les applications de cette thèse, on ne s'intéressera qu'à des modèles pronostiques issus de modèles de régression. Cependant, la grande liberté de modélisation permise par l'absence de contrainte d'interprétation des paramètres permet aussi d'utiliser de nombreuses méthodes issues de l'apprentissage automatique, présentées dans la monographie Hastie *et al.* (2009) par exemple. Notons d'ailleurs que certaines d'entre elles sont très populaires, comme les forêts aléatoires de Breiman (2001) qui ont récemment été étendues aux données de survie et aux situations de risques concurrents (Ishwaran *et al.*, 2008; Mogensen et Gerds, 2013).

Par ailleurs, pour construire des modèles pronostiques paramétriques ou semi-paramétriques, notons qu'on attache généralement peu d'importance à la validité de certaines hypothèses courantes, telles que la linéarité (ou log-linéarité) de l'effet de covariables par exemple. C'est aussi l'une des différences majeures entre les modèles pronostiques et les modèles d'association ou de causalité. S'il est vrai qu'un modèle pronostique prédira d'autant mieux que les hypothèses sous-jacentes qu'il suppose sont vérifiées, en général rien n'implique qu'il prédira mal si certaines de ces hypothèses sont violées (Steyerberg, 2009, Sec. 6.1.1). On retrouve donc en prédiction, comme dans de nombreux domaines de la statistique, le célèbre message de Georges Box : « Au fond, tous les modèles sont faux mais certains sont utiles » [1].

En biostatistique, pour modéliser l'effet de covariables X sur l'occurrence de $D = 1$, on modélise souvent $\mathbb{P}(D = 1|X)$ avec un modèle logistique. Pour modéliser l'occurrence de $D \equiv D(t) = \mathbb{1}_{(T \leqslant t)} = 1$ en fonction de X, i.e. $\mathbb{P}(T \leqslant t|X)$, on utilise souvent un modèle de Cox. Pour $\mathbb{P}(T \leqslant t, \eta = 1|X)$, on peut similairement utiliser un modèle de Fine et Gray (1999) par exemple. De nombreuses alternatives sont aussi disponibles et foisonnent dans la littérature. Toutes ces approches sont particulièrement adaptées lorsqu'on suppose un unique

1. Citation originale : « Essentially, all models are wrong, but some are useful » (Box, 1979)

temps auquel on fait le pronostic, appelé « temps landmark » dans la suite, auquel on dispose de l'information sur les covariables X.

Une autre famille de modèles pronostiques nous intéressera cependant particulièrement au Chapitre V : celle des modèles pronostiques dit « dynamiques », dont l'idée est décrite ci-après.

II.2.2 Modèle de prédiction dynamique en présence de marqueurs longitudinaux

II.2.2.1 L'idée de la prédiction dynamique

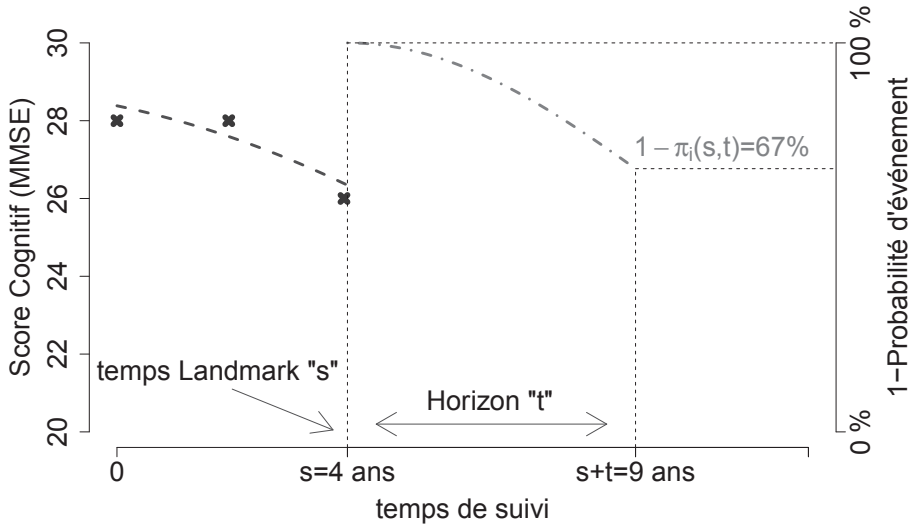


FIGURE II.4 – Prédiction dynamique en utilisant des mesures répétées d'un test cognitif (voir texte pour la légende).

Notre intérêt particulier pour les modèles dynamiques est illustré à la Figure II.4. La situation est la suivante. Les sujets des cohortes Paquid et Trois-Cités sont suivis régulièrement (environ tous les deux ans). A chaque visite, leur cognition est évaluée et des scores cognitifs sont enregistrés (représentés par les croix de la Figure II.4). Après quelques années de suivi, disons s années, on souhaite prédire l'apparition ou non d'un événement dans les t années suivantes, que l'on notera $D(s,t) = 1$ s'il a lieu et $D(s,t) = 0$ sinon. Pour cela, on souhaite utiliser

toute l'information à notre disposition au temps de prédiction s, le temps « landmark ». En particulier, on veut utiliser l'information sur les mesures répétées des tests cognitifs enregistrés jusqu'en s, et qui nous donne une idée de l'évolution cognitive du sujet, (représentée en tiret sur la Figure II.4). C'est parce que l'information utilisée pour la prédiction croît avec le temps landmark s, que l'on parle de prédiction dynamique.

Formellement, on s'intéresse donc à la prédiction dynamique estimant une probabilité de la forme

$$\pi(s,t) = \mathbb{P}\big(D(s,t) = 1 \big| T > s, \mathcal{H}(s)\big), \tag{II.5}$$

où T représente un temps d'événement et $\mathcal{H}(s)$ (« $\mathcal{H}$ » pour histoire) représente toute l'information disponible et utile pour prédire $D(s,t) = 1$. Pour $s = 4$ ans fixé, la fonction décroissante $t \mapsto 1 - \pi(s,t)$ est représentée en alternance de tiret et de pointillé sur la Figure II.4, pour $t \in [0,5]$ ans.

Dans la littérature, $D(s,t)$ représente le plus souvent l'indicateur de l'événement $\{s < T \leqslant s + t\}$ (Proust-Lima et Taylor, 2009; Proust-Lima *et al.*, 2012; Rizopoulos, 2011; van Houwelingen, 2007, entre autres).

Cependant, dans le cas de la prédiction de la démence, en raison de la présence du risque concurrent de décès sans démence (Figure II.3), on aura $D(s,t) = \{s < T \leqslant s + t, \eta = 1\}$, avec T correspondant à la durée entre l'âge de l'apparition de la démence, ou du décès sans démence, et l'âge à l'entrée dans l'étude, et avec $\eta = 1$ indiquant l'événement démence, et $\eta = 2$ l'événement décès sans démence. Comme Nicolaie *et al.* (2013a,b), entre autres, on s'intéressera donc à ce type de prédictions dynamiques au Chapitre V.

II.2.2.2 Les modèles multi-états et l'approche landmark

Une première approche pour estimer des prédictions dynamiques consiste à utiliser, de manière adaptée, des modèles à l'origine proposés pour modéliser uniquement l'effet de variables qui ne changent pas au cours du temps (variables « baseline »).

En particulier, les modèles multi-états peuvent être utilisés pour construire des prédictions dynamiques de $D(s,t) = 1$, en fonction d'une information dynamique $\mathcal{H}(s)$ (van Houwelingen

et Putter, 2012, Section III.9). Brièvement, l'idée consiste à définir l'événement $D(s,t) = 1$ comme l'entrée dans un état absorbant, et à modéliser l'information $\mathcal{H}(s)$ par la trajectoire du sujet à travers plusieurs états intermédiaires. Par exemple, Cortese et Andersen (2010) et Cortese *et al.* (2013) utilisent la méthode pour prédire un décès chez des patients leucémiques, gréffés de la moelle, où l'information dynamique $\mathcal{H}(s)$ inclut notamment l'apparition de complications dues à la greffe avant le temps s.

Une autre approche consiste à utiliser une « grille » de temps s, et à estimer autant de modèles prédictifs qu'il y a de temps s dans la grille, pour modéliser directement $D(s,t) = 1$ en fonction de $\mathcal{H}(s)$. C'est l'approche dite « landmark » (van Houwelingen, 2007; van Houwelingen et Putter, 2012, Section III.8). Éventuellement, les différents modèles correspondant aux différents temps landmark s peuvent être combinés en un seul « super model » en suivant l'idée des équations d'estimation pour données groupées de Liang et Zeger (1986) (van Houwelingen, 2007). Considérant le même exemple d'application des patients leucémiques greffés, Nicolaie *et al.* (2013b) modélisent ainsi directement $D(s,t) = 1$ en fonction de $\mathcal{H}(s)$ par un modèle logistique.

Ces deux approches sont très intéressantes en présence de données de survie, et particulièrement lorsque l'information dynamique $\mathcal{H}(s)$ représente des covariables dépendantes du temps observées en temps continu, et que ces covariables sont non bruitées. C'est le cas, par exemple, lorsque $\mathcal{H}(s)$ inclut les indicatrices d'une rechute d'un cancer, d'un traitement ou d'une aggravation sévère d'une maladie avant le temps s. Elles sont cependant plus difficiles à utiliser et peu appropriées lorsque $\mathcal{H}(s)$ représente les observations de biomarqueurs ou de tests cognitifs jusqu'en s, qui ne sont observés qu'à des temps discrets, et généralement avec des erreurs de mesure.

II.2.2.3 Les modèles conjoints

Une autre approche pour estimer des prédictions dynamiques consiste à modéliser conjointement l'information dynamique $\mathcal{H}(s)$ et l'événement $D(s,t)$.

Plus spécifiquement, reprenons le cas de la démence qui nous intéresse principalement, et

pour lequel on cherche à estimer

$$\begin{aligned}\pi(s,t) &= \mathbb{P}(D(s,t)=1|T>s,\mathcal{H}(s)) \\ &= \mathbb{P}(s<T\leqslant s+t,\eta=1|T>s,\mathcal{H}(s)),\end{aligned} \tag{II.6}$$

avec $\mathcal{H}(s) = \{X, \mathcal{Y}(s)\}$, où X représente des covariables fixes et $\mathcal{Y}(s)$ représente les scores cognitifs du sujet observés jusqu'en s. Brièvement, l'idée consiste à modéliser conjointement (T,η) et la trajectoire du test cognitif, notée $Y(\cdot)$, sachant les covariables fixes X. Pour cela, on modélise la corrélation entre (T,η) et $Y(\cdot)$ par une structure latente γ, dont on paramétrise la distribution, et on utilise deux sous-modèles paramétriques pour les distributions de (T,η) sachant γ et de $Y(\cdot)$ sachant γ. De nombreux modèles sont alors possibles. Les plus communs sont présentés dans les revues de la littérature exhaustives de Tsiatis et Davidian (2004) et Rizopoulos (2012). Le modèle conjoint ainsi défini étant complétement paramétrique, les estimations des paramètres peuvent être obtenues par maximisation de la vraissemblance. Pour estimer les prédictions dynamiques, l'approche la plus simple consiste ensuite à réécrire $\pi(s,t)$ en fonction de l'information dynamique $\mathcal{H}(s) = \{X, \mathcal{Y}(s)\}$ et des paramètres du modèle, puis à remplacer les paramètres par leur estimation.

Dans cette thèse, notre intérêt particulier pour la modélisation conjointe réside dans le fait que, en modélisant la trajectoire de l'évolution des tests cognitifs, elle s'adapte bien au fait que les scores cognitifs ne soient mesurés qu'à des temps de visites discrets, différents pour chaque sujet, et au fait que les scores cognitifs représentent les observations bruitées de niveaux cognitifs sous-jacents. Nous reparlerons de prédictions dynamiques basées sur des modèles conjoints au Chapitre V.

Par ailleurs, notons qu'historiquement le développement des modèles conjoints a principalement été guidé par deux motivations autres que celle de la prédiction dynamique (Jacqmin-Gadda *et al.*, 2004). Une première était l'étude d'évolution de biomarqueurs en présence de données manquantes non aléatoires dues à des décès. Une autre était l'étude du lien entre une covariable dépendante du temps « interne » (aussi dite « endogène »), typiquement un biomarqueur, et le risque instantané de décès. En effet, les approches plus usuelles basées sur la

maximisation de la vraisemblance partielle de Cox (1972) ont été développées pour des variables dépendantes du temps « externes » (aussi dites « exogènes »), tel qu'un niveau de pollution atmosphérique, et sont souvent peu pertinentes en présence de variables dépendantes du temps « internes » (Fisher et Lin, 1999; Kalbfleisch et Prentice, 2002, Section 6.3).

II.3 Évaluation des modèles pronostiques

« *Prédire est difficile, surtout lorsqu'il s'agit de l'avenir* » (Niels Bohr [2]). Pour tenter de faire de leur mieux lorsqu'il s'agit de prédiction, les statisticiens ont proposé de nombreux modèles pronostiques. Cependant, « *Les statisticiens, comme les artistes, ont la fâcheuse habitude de tomber amoureux de leurs modèles* » (George Box [3]). Il est donc essentiel de pouvoir évaluer et comparer différentes prédictions issues de différents modèles de manière juste et objective.

On présente donc dans cette section les principaux critères d'évaluation et de comparaison de prédictions utilisées en biostatistique, dont beaucoup sont décrits par les récentes revues de la littérature de Gerds *et al.* (2008); Steyerberg *et al.* (2010) et Steyerberg (2009) pour les cas généraux, de Wolbers *et al.* (2009) pour le cas particulier des situations de risques concurrents et de Schoop *et al.* (2011) pour celui des prédictions dynamiques.

On s'intéressera dans cette thèse uniquement aux critères les plus courants, pouvant être utilisés pour comparer les prédictions issues de n'importe quel modèle ou technique de prédiction. On considérera que les prédictions peuvent être issues d'un modèle de régression paramétrique, semi-paramétrique, non-paramétrique, d'une méthode d'apprentissage automatique, etc.

II.3.1 Capacités à discriminer d'un marqueur

Parmi les critères les plus usités en biostatistique pour évaluer un modèle pronostique, on trouve ceux utilisés à l'origine pour évaluer et comparer des marqueurs, ici décrits.

2. La citation originale étant : « *Prediction is very difficult, especially about the future* »
3. La citation originale étant : « *Statisticians, like artists, have the bad habit of falling in love with their models.* »

II.3.1.1 La courbe ROC

Il est souvent utile d'évaluer le potentiel d'un marqueur, noté M, à discriminer une population de sujets malades d'une population de sujets non malades, ou plus généralement ayant ou non subi un événement quelconque. On notera D (pour « Diseased ») l'indicateur de maladie, avec $D = 1$ pour malade, et $D = 0$ pour non malade. On appellera un sujet malade un *cas* et un sujet non malade un *contrôle*. La plupart du temps, c'est une valeur haute du marqueur qui est associée au diagnostic positif de maladie. C'est par exemple le cas lorsque M représente une charge virale dans les études concernant le virus de l'immunodéficience humaine (VIH), ou lorsque M représente un taux de PSA (pour « prostate-specific antigen ») dans les études sur le cancer de la prostate. Par convention et sans perte de généralité, on considérera dans la suite que c'est le cas (autrement on peut considérer $-M$ pour renverser l'association). Pour alléger la présentation, on considère aussi ici que M a une distribution continue, telle que la probabilité d'observer des ex-aequos est nulle, bien que les modifications requises sans ces hypothèses soient mineures.

Pour un seuil c choisi, on peut alors dichotomiser M et construire un test diagnostique défini comme positif si $M > c$ et négatif si $M \leqslant c$. On définit alors la sensibilité, notée $Se(c)$, et la spécificité, notée $Sp(c)$, comme les probabilités de « vrais positifs » et de « vrais négatifs »,

$$Se(c) = \mathbb{P}(M > c|D = 1) \quad \text{et} \quad Sp(c) = \mathbb{P}(M \leqslant c|D = 0).$$

Notons que ces mesures ne dépendent pas de la prévalence $\mathbb{P}(D = 1)$ de la maladie. Elles peuvent ainsi être estimées même à partir d'études cas-témoins, ce qui est particulièrement pratique lorsque la mesure de M est coûteuse ou invasive par exemple. C'est une des raisons principales de leur popularité, en plus de leur facilité d'interprétation. Pour étudier la capacité de discrimination du marqueur sur l'ensemble des valeurs seuil c possibles, on s'intéresse souvent à la courbe ROC. Historiquement introduite pour l'analyse de signaux radars durant la Seconde Guerre mondiale, elle est depuis les années 80 très utilisée aussi bien dans le milieu du marketing que celui de la recherche médicale.

La courbe ROC est définie par l'ensemble des points $\Big\{\big(Se(c), 1 - Sp(c)\big), c \in \mathbb{R}\Big\}$. Par

définition, cette courbe est située dans le rectangle $[0,1] \times [0,1]$, telle que $ROC(0) = 0$, $ROC(1) = 1$, et est croissante de pente $dROC(c)/dc$ égale au rapport de vraisemblances $f_{M|D=1}(c)/f_{M|D=0}(c)$, avec $f(\cdot)_{M|D=d}$, $d = 0,1$, les densités conditionnelles de M sachant $D = d$. La courbe ROC est égale à la première bissectrice si M ne discrimine aucunement $D = 1$ de $D = 0$, i.e. si $\forall c \in \mathbb{R}, \quad \mathbb{P}(M > c|D = 1) = \mathbb{P}(M > c|D = 0)$, et est constante égale à 1 sur $]0,1]$ si M discrimine parfaitement $D = 1$ de $D = 0$, i.e. si $\forall c \in \mathbb{R}, \quad \mathbb{P}(M > c|D = 1) = 1$ et $\mathbb{P}(M > c|D = 0) = 0$.

Plus généralement, plus l'aire sous cette courbe, notée AUC (pour « Area Under Curve »), est grande et plus M discrimine bien $D = 1$ de $D = 0$. En effet, l'aire sous la courbe ROC représente la probabilité qu'un sujet malade ait une valeur du marqueur M supérieure à celle d'un sujet non malade,

$$\begin{aligned} AUC = \int_0^1 ROC(u)du &= \int_0^1 Se((1-Sp)^{-1}(u))du \\ &= \int_{-\infty}^{\infty} Se(c) f_{M|D=0}(c)dc \\ &= \int_{-\infty}^{\infty} \mathbb{P}(M > c|D = 1) f_{M|D=0}(c)dc \\ &= \mathbb{P}\big(M_i > M_j | D_i = 1, D_j = 0\big), \end{aligned}$$

avec i et j les indices de deux sujets indépendants.

En outre, deux autres quantités sont aussi parfois intéressantes pour un seuil c fixé : la valeur positive prédictive, notée PPV (pour « Positive Predictive Value ») et la valeur prédictive négative, notée NPV (pour « Negative Predictive Value »). Elles sont étroitement liées à $Se(c)$ et $Sp(c)$, et définies par

$$PPV(c) = \mathbb{P}(D = 1|M > c) \quad \text{et} \quad NPV(c) = \mathbb{P}(D = 0|M \leqslant c).$$

Contrairement à $Se(c)$ et $Sp(c)$, elles dépendent de la prévalence $\mathbb{P}(D = 1)$. Elles ne quantifient donc pas directement les capacités intrinsèques de M à discriminer des sujets malades ou non. Elles sont cependant intéressantes pour évaluer l'intérêt clinique de M. Notons que par définition, elles ne peuvent pas être estimées à partir de données d'études cas-témoins.

D'un point de vue de l'estimation, si l'on dispose d'un n-échantillon i.i.d $\{(M_i, D_i), i = 1, \ldots, n\}$ représentatif, alors pour tout seuil $c \in \mathbb{R}$, il est pratique de définir les effectifs d'un tableau de contingence,

$$n_D^+(c) = \sum_{i=1}^n \mathbb{1}_{(M_i > c, D_i = 1)} \quad \text{et} \quad n_{\bar{D}}^+(c) = \sum_{i=1}^n \mathbb{1}_{(M_i > c, D_i = 0)},$$
$$n_D^-(c) = \sum_{i=1}^n \mathbb{1}_{(M_i \leqslant c, D_i = 1)} \quad \text{et} \quad n_{\bar{D}}^-(c) = \sum_{i=1}^n \mathbb{1}_{(M_i \leqslant c, D_i = 0)},$$

et de ses marges,

$$n^+(c) = n_D^+(c) + n_{\bar{D}}^+(c) \quad \text{et} \quad n^-(c) = n_D^-(c) + n_{\bar{D}}^-(c),$$
$$n_D(c) = n_D^+(c) + n_D^-(c) \quad \text{et} \quad n_{\bar{D}}(c) = n_{\bar{D}}^+(c) + n_{\bar{D}}^-(c),$$

comme résumé par la Table II.1.

		Malade (oui/non)		
		Oui ($D = 1$)	Non ($D = 0$)	
Test	Positif ($M > c$)	$n_D^+(c)$	$n_{\bar{D}}^+(c)$	$n^+(c)$
	Négatif ($M \leqslant c$)	$n_D^-(c)$	$n_{\bar{D}}^-(c)$	$n^-(c)$
		$n_D(c)$	$n_{\bar{D}}(c)$	

TABLE II.1 – Tableau de contingence associé à $Se(c)$, $Sp(c)$, $PPV(c)$, et $NPV(c)$.

A partir de ces notations, on définit alors les estimateurs empiriques :

$$\widehat{Se}(c) = \frac{n_D^+(c)}{n_D(c)} \quad \text{et} \quad \widehat{Sp}(c) = \frac{n_{\bar{D}}^-(c)}{n_{\bar{D}}(c)},$$
$$\widehat{PPV}(c) = \frac{n_D^+(c)}{n^+(c)} \quad \text{et} \quad \widehat{NPV}(c) = \frac{n_{\bar{D}}^-(c)}{n^-(c)}.$$

De plus, en sommant des rectangles de largeurs de la taille des accroissements de $1 - \widehat{Sp}(\cdot)$ et de hauteurs des sensibilités correspondantes, on peut montrer que l'aire sous la courbe définie par l'ensemble des points $\left\{\left(\widehat{Se}(c), 1 - \widehat{Sp}(c)\right), c \in \mathbb{R}\right\}$ est égale à la simple expression :

$$\widehat{AUC} = \frac{1}{n_D n_{\bar{D}}} \sum_{i=1}^n \sum_{j=1}^n \mathbb{1}_{(M_i > M_j, D_i = 1, D_j = 0)}.$$

A partir de ces estimations, on peut aisément construire des intervalles de confiance et des tests. Les propriétés statistiques de $\widehat{Se}(c)$, $\widehat{Sp}(c)$, $\widehat{PPV}(c)$ et $\widehat{NPV}(c)$ découlent des propriétés usuelles des proportions. En revanche, $\widehat{AUC}$ est une proportion moins usuelle : une U-statistique, dont les propriétés ont été étudiées par DeLong *et al.* (1988), qui en ont dérivé des intervalles de confiance et des tests. Pour la plupart, les propriétés asymptotiques de $\widehat{AUC}$ font aujourd'hui l'objet des exemples et exercices du livre sur les U-statistique Kowalski et Tu (2008). Enfin, notons que de nombreux détails sur les courbes ROC sont présentés par l'ouvrage exhaustif de Pepe (2003).

II.3.1.2 Courbe ROC dépendant du temps

À l'origine et comme présenté précédemment, les courbes ROC ont d'abord été proposées pour évaluer des tests diagnostiques, c'est-à-dire évaluer la capacité d'un marqueur à diagnostiquer une maladie à la date de la mesure du marqueur. Par la suite, Heagerty *et al.* (2000) et Heagerty et Zheng (2005), entre autres, ont suggéré l'utilisation des courbes ROC pour évaluer des marqueurs pronostiques, c'est-à-dire pour évaluer la capacité d'un marqueur à détecter des futurs malades, dans les quelques jours, mois ou années suivant la mesure du marqueur par exemple.

Majoritairement développées dans les années 2000, les méthodologies proposées ont été motivées par l'intérêt croissant des cliniciens pour les diagnostics précoces et le ciblage de populations à risque, aujourd'hui considérés comme des opportunités majeures de progrès dans le traitement et la prise en charge de nombreuses pathologies, notamment de nombreux cancers.

D'un point de vue inférentiel, la principale différence entre les méthodes pour courbes ROC diagnostiques et pronostiques vient de la possibilité d'observer des sujets perdus de vue dans le cas pronostique, qu'on discutera en Section II.4.1.

II.3.1.2.1 Sans risques concurrents

Les notions de *cas* et de *contrôle* étant à la base du concept des courbes ROC, c'est de leurs définitions que résultent les différentes courbes ROC dépendant du temps proposées dans la

littérature. Considérant un marqueur M mesuré au temps $t = 0$, et l'apparition d'un événement (une maladie) au temps de survie T, Heagerty et Zheng (2005) distinguent les définitions de divers couples de cas *incidents* ou *cumulatifs* et de contrôles *statiques* ou *dynamiques* pour la courbe ROC dépendant du temps, temps que nous noterons t. On les décrit au Tableau II.2.

Cas		Contrôle		notation
Incident	$T = t$	Dynamique	$T > t$	$\mathbb{I}/\mathbb{D}$
Cumulatif	$T \leqslant t$	Dynamique	$T > t$	$\mathbb{C}/\mathbb{D}$
Incident	$T = t$	Statique	$T > \tau$	$\mathbb{I}/\mathbb{S}$

TABLE II.2 – Définitions de couples de cas et de contrôles dépendant du temps, avec $t > 0$ et τ fixé, tel que pour tous les temps t d'intérêt $\tau \gg t$ (sans risques concurrents).

Diverses définitions de la sensibilité, notée $Se(c,t)$, de la spécificité notée $Sp(c,t)$, de la courbe ROC et de l'AUC correspondant, notées $ROC(t)$ et $AUC(t)$, en découlent alors. Respectivement, on définit les sensibilités *incidente* et *cumulative* par

$$Se^{\mathbb{I}}(c,t) = \mathbb{P}(M > c|T = t) \quad \text{et} \quad Se^{\mathbb{C}}(c,t) = \mathbb{P}(M > c|T \leqslant t)$$

et les spécificités *dynamique* et *statique* par

$$Sp^{\mathbb{D}}(c,t) = \mathbb{P}(M \leqslant c|T > t) \quad \text{et} \quad Sp^{\mathbb{S}}(c,\tau) = \mathbb{P}(M \leqslant c|T > \tau)$$

avec $t > 0$ et τ fixé, tels que pour tous les temps t d'intérêt $\tau \gg t$.

La définition $\mathbb{I}/\mathbb{D}$ est souvent considérée comme utile pour évaluer en détail les caractéristiques de discrimination intrinsèques au marqueur M, sans forcément d'application de M à l'esprit. En pratique, on s'intéresse en effet rarement à l'occurrence d'un événement exactement au temps t, mais plutôt dans une fenêtre de temps. Cependant, on verra à la Section II.3.1.3 que cette définition est étroitement liée à celle du C-index, ce qui explique probablement en partie sa popularité.

La superposition des tracés de plusieurs courbes ROC$^{\mathbb{I}/\mathbb{D}}$, à différents temps t, peut s'avérer informative pour étudier la dynamique des capacités de prédiction de M quand t varie. Cependant, l'interprétation n'est pas aisée, puisque les groupes des cas et des contrôles changent

tous les deux avec t. La définition $\mathbb{I}/\mathbb{S}$ n'a pas cet inconvénient, puisque le groupe contrôle ne change pas avec t. De plus, elle est intéressante si l'on cherche à distinguer les cas incidents d'un groupe contrôle qui ne serait pas à risque sur une longue période ($\tau >> t$).

La définition $\mathbb{C}/\mathbb{D}$, quant à elle, est la plus pertinente pour nos applications. C'est celle qui est utile pour évaluer la capacité d'un marqueur à prédire l'apparition d'un évènement ($D \equiv D(t) = \mathbb{1}_{(T \leqslant t)} = 1$) ou non ($D \equiv D(t) = 0$) dans les t années (ou autre unité de temps) suivant la mesure du marqueur. C'est donc celle-ci qui est essentielle pour évaluer les capacités d'outils ou de marqueurs pronostiques d'aide à la décision. Par exemple, cette définition est pertinente lorsqu'on s'intéresse à la prédiction d'un événement tel qu'une démence, un risque cardio-vasculaire ou un cancer dans les 5 ans à venir. Notons que pour cette définition l'$AUC(t)$ correspond alors à

$$\begin{aligned} AUC^{\mathbb{C}/\mathbb{D}}(t) &= \int_0^1 ROC^{\mathbb{C}/\mathbb{D}}(u)du \\ &= \int_0^1 Se^{\mathbb{C}}((1 - Sp^{\mathbb{D}})^{-1}(u))du = \mathbb{P}\big(M_i > M_j | T_i \leqslant t, T_j > t\big), \end{aligned}$$

avec i et j les indices de deux sujets indépendants, et correspond donc toujours à un indice de concordance simple à interpréter.

Enfin, pour plus de détails sur les courbes ROC dépendant du temps ainsi que les méthodes statistiques associées proposées dans la littérature, on pourra consulter l'intéressante discussion de Pepe *et al.* (2008a) ou la récente revue de la littérature Blanche *et al.* (2013b), disponible en annexe.

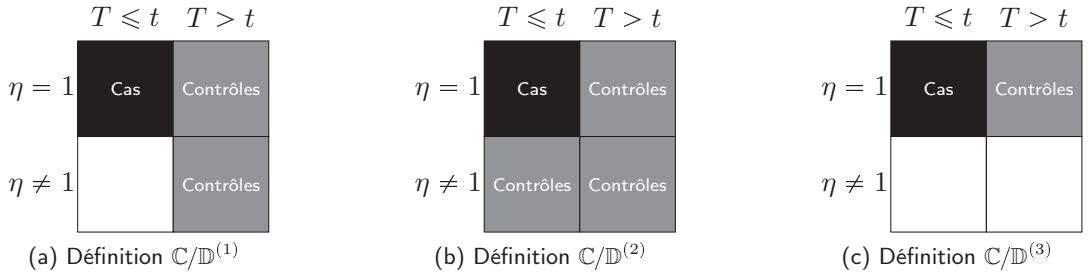
II.3.1.2.2 En présence de risques concurrents

En présence de risques concurrents, on s'intéresse souvent à un événement principal, noté $\eta = 1$, plutôt qu'aux événements concurrents, notés $\eta \neq 1$. Les définitions de *cas* et de *contrôles* doivent alors être adaptées, comme nous le reverrons plus en détail au Chapitre IV.

Saha et Heagerty (2010) et Foucher *et al.* (2010) proposèrent les premiers des définitions avant Zheng *et al.* (2012a). Comme en l'absence de risques concurrents, on peut aussi définir des cas *incidents* et *dynamiques*. On peut également définir des contrôles *statiques*. Ce concept

Cas		Contrôle		Notation
Incident	$T = t, \eta = 1$	Dynamique (1)	$T > t$	$\mathbb{I}/\mathbb{D}^{(1)}$
Cumulatif	$T \leqslant t, \eta = 1$	Dynamique (1)	$T > t$	$\mathbb{C}/\mathbb{D}^{(1)}$
Cumulatif	$T \leqslant t, \eta = 1$	Dynamique (2)	$T > t$ ou $T \leqslant t, \eta \neq 1$	$\mathbb{C}/\mathbb{D}^{(2)}$
Cumulatif	$T \leqslant t, \eta = 1$	Dynamique (3)	$T > t, \eta = 1$	$\mathbb{C}/\mathbb{D}^{(3)}$

TABLE II.3 – Définitions de couples de cas et de contrôles dépendant du temps, avec $t > 0$ (avec risques concurrents).



(a) Définition $\mathbb{C}/\mathbb{D}^{(1)}$ (b) Définition $\mathbb{C}/\mathbb{D}^{(2)}$ (c) Définition $\mathbb{C}/\mathbb{D}^{(3)}$

FIGURE II.5 – Sous groupes de la population définis par différentes définitions « Cumulatives/-Dynamiques » des cas (en noir) et des contrôles (en gris) en présence de risques concurrents.

de contrôles *statiques* n'a cependant pas été repris, probablement compte tenu de son intérêt plus mineur. Cependant, pour la définition des contrôles *dynamiques*, plusieurs définitions sont possibles, dépendant de la façon de considérer les sujets subissant les événements concurrents. Les définitions sont présentées au Tableau II.3.

Notons que, contrairement à la définition usuelle des cas et des contrôles ($D = 1$ versus $D = 0$ en Section II.3.1.1), ou à la définition $\mathbb{C}/\mathbb{D}$ en l'absence des risques concurrents, les groupes des cas et des contrôles des définitions $\mathbb{C}/\mathbb{D}^{(1)}$ et $\mathbb{C}/\mathbb{D}^{(3)}$ ne forment pas une partition de l'ensemble de la population. Seule la définition $\mathbb{C}/\mathbb{D}^{(2)}$ en forme une, comme illustré à la Figure II.5. C'est en fait la raison principale qui nous fera le plus souvent préférer cette définition $\mathbb{C}/\mathbb{D}^{(2)}$ parmi les trois *cumulatives/dynamiques*. Par ailleurs, notons que la définition dynamique (3) des cas, proposée par Foucher *et al.* (2010), peut être difficilement identifiable en pratique, avec une durée de suivi limitée, en particulier non paramétriquement, ou sans

hypothèses d'extrapolation non testables, puisque implicitement elle nécessite l'estimation de $\mathbb{P}(\eta = 1) = \lim_{t\to\infty} \mathbb{P}(T \leqslant t, \eta = 1)$.

II.3.1.3 C-index

Une autre mesure de capacité d'un marqueur à discriminer les futurs malades des futurs non malades est également populaire : le C-index proposé par Harrell *et al.* (1996), que nous noterons $\mathcal{C}$ ($\mathcal{C}$ pour Concordance). A l'origine, il était défini comme

$$\mathcal{C} = \mathbb{P}(M_i > M_j | T_i < T_j),$$

avec i et j les indices de deux sujets indépendants. Notons qu'on a aussi les expressions $\mathcal{C} = \mathbb{P}(T_i < T_j | M_i > M_j)$ ou $\mathcal{C} = 2\mathbb{P}(M_i > M_j, T_i < T_j)$, puisque $\mathbb{P}(M_i > M_j) = \mathbb{P}(T_i < T_j) = 0.5$ (par indépendance et symétrie), et que quelques lignes de calculs montrent que l'indice $\mathcal{C}$ n'est rien d'autre qu'une transformation linéaire du « τ de Kendall », i.e. $\mathcal{C} = 0.5(1-\tau)$ (Harrell *et al.*, 1996). Contrairement aux aires sous les courbes dépendantes du temps, sa définition ne dépend pas d'un temps t spécifique. Ceci peut parfois être vu comme un avantage lorsqu'aucun temps t n'est d'un intérêt particulier comme temps de prédiction, ou lorsque l'on cherche à avoir un résumé global des capacités de prédiction d'un marqueur, sur tous les temps t de prédiction possibles.

En pratique, il existe généralement un temps τ_0, correspondant par exemple à la durée maximale du suivi d'un sujet dans une étude, telle qu'on n'observe que les temps d'événement T vérifiant $T < \tau_0$. Pour éviter tout problème d'identifiabilité, Uno *et al.* (2011), Heagerty et Zheng (2005) et Gerds *et al.* (2013b) ont donc préféré définir un C-index tronqué, noté $\mathcal{C}(\tau_0)$ et défini par

$$\mathcal{C}(\tau_0) = \mathbb{P}(M_i > M_j | T_i < T_j, T_i < \tau_0).$$

Par ailleurs, Heagerty et Zheng (2005) notent qu'on peut voir $\mathcal{C}(\tau_0)$ comme une moyenne pondérée des différents $AUC^{\mathbb{I}/\mathbb{D}}(s)$ pour $s \in [0, \tau_0]$. En effet, en notant que

$$AUC^{\mathbb{I}/\mathbb{D}}(t) = \mathbb{P}(M_i > M_j | T_i = t, T_j > t)$$

on a alors

$$\begin{aligned}
\mathcal{C}(\tau_0) &= \mathbb{P}(M_i > M_j | T_i < T_j, T_i < \tau_0) \\
&= \frac{\mathbb{P}(M_i > M_j, T_i < T_j, T_i < \tau_0)}{\mathbb{P}(T_i < T_j, T_i < \tau_0)} \\
&= \int_0^{\tau_0} \frac{\mathbb{P}(M_i > M_j, T_i = s, T_j > s)}{\mathbb{P}(T_i < T_j, T_i < \tau_0)} ds \\
&= \int_0^{\tau_0} \mathbb{P}(M_i > M_j | T_i = s, T_j > s) \frac{\mathbb{P}(T_i = s, T_j > s)}{\mathbb{P}(T_i < T_j, T_i < \tau_0)} ds = \int_0^{\tau_0} AUC^{\mathbb{I}/\mathbb{D}}(s) w_{\tau_0}(s) ds
\end{aligned}$$

avec $w_{\tau_0}(s) = f_T(s)S_T(s)/W(\tau_0)$, $f_T(\cdot)$ la densité de T et $S_T(\cdot)$ sa fonction de survie et $W(\tau_0) = \mathbb{P}(T_i < T_j, T_i < \tau_0) = \int_0^{\tau_0} f_T(s)S_T(s)ds$, car les sujets i et j sont indépendants. En particulier, $W(\infty) = \int_0^{\infty} f_T(s)S_T(s)ds = \{-S_T(\infty)^2 + S_T(0)^2\}/2 = 1/2$.

II.3.2 Capacités pronostiques d'un modèle

L'utilisation des mesures de capacités pronostiques de marqueurs (précédemment présentées) est souvent proposée pour évaluer les capacités de discrimination d'un modèle pronostique (Steyerberg *et al.*, 2010).

Deux autres mesures de discrimination ont récemment été proposées par Pencina *et al.* (2008) : les indices IDI, pour « Integrated Discrimination Improvement » et NRI, pour « Net Reclassification Index ». Bien qu'assez récents, ils sont aujourd'hui extrêmement populaires (en septembre 2013, l'article de Pencina *et al.* (2008) était cité plus de 1600 fois d'après *Google scholar*).

La motivation pour introduire l'IDI et le NRI venait de la constatation que les différences entre les AUCs de deux modèles pronostiques différents étaient souvent faibles (Pencina *et al.*, 2008). En particulier, lorsqu'une variable supplémentaire est ajoutée dans un modèle de prédiction pré-existant, même lorsqu'elle est significative, la différence entre les AUCs des deux modèles prédictifs est souvent minime et non significative. Pourtant, une différence d'AUC était (et est encore) souvent considérée comme un critère principal pour comparer des modèles pronostiques en biostatistique (Pepe et Janes, 2013). Ces deux nouvelles mesures de discrimination ont alors été proposées car elles étaient prétendues plus sensibles et plus pertinentes que l'AUC

pour la comparaison de modèles.

Dans toute cette Section II.3.2, on notera D une variable binaire ($D = 1$ si l'événement a lieu, $D = 0$ sinon), et X et Y des vecteurs de covariables. On supposera que l'on souhaite prédire l'événement $D = 1$ à partir de X uniquement ou de X et Y, et on notera risk(X) et risk(X, Y) les prédictions correspondantes. Estimant des probabilités, on supposera aussi, sans perte de généralité importante, que risk(X) et risk(Y, X) sont à valeurs dans $[0, 1]$.

II.3.2.1 L'IDI

Considérant deux modèles de prédiction, l'un utilisant uniquement les covariables X, noté risk(X), et l'autre les covariables X et Y, noté risk(X, Y), l'IDI comparant risk(X) et risk(X, Y) est défini par :

$$\begin{aligned} IDI =& \mathbb{E}\Big[\text{risk}(X,Y) - \text{risk}(X)\Big|D = 1\Big] \\ &+ \mathbb{E}\Big[-\Big\{\text{risk}(X,Y) - \text{risk}(X)\Big\}\Big|D = 0\Big]. \end{aligned} \tag{II.7}$$

Le premier « I » de IDI, pour « integrated », vient du fait que, pour tout modèle de prédiction ou marqueur M à valeur dans $[0, 1]$, l'intégrale de la sensibilité pour M, notée IS$_M$, est

$$\begin{aligned} IS_M = \int_0^1 Se(c)dc &= \int_0^1 \mathbb{P}(M > c|D = 1)dc \\ &= \int_0^1 \mathbb{E}\big[\mathbb{1}_{(M>c)}\big|D = 1\big]dc \\ &= \mathbb{E}\left[\int_0^1 \mathbb{1}_{(M>c)}dc\Big|\, D = 1\right] = \mathbb{E}\big[M\big|D = 1\big], \end{aligned}$$

et similairement l'intégrale de un moins la spécificité pour M, noté IP_M, est

$$IP_M = \int_0^1 \{1 - Sp(c)\}\, dc = \int_0^1 \mathbb{P}(M > c|D = 0)dc = \mathbb{E}\big[M\big|D = 0\big].$$

On a ainsi l'expression alternative :

$$IDI = \Big(IS_{\text{risk}(X,Y)} - IS_{\text{risk}(X)}\Big) - \Big(IP_{\text{risk}(X,Y)} - IP_{\text{risk}(X)}\Big)\,, \tag{II.8}$$

qui correspond en fait à la première définition de l'IDI donnée par Pencina *et al.* (2008). L'idée intuitive de cet indice est la suivante : si le prédicteur risk(X, Y) prédit mieux $D = 1$ que le prédicteur risk(X), alors risk(X, Y) a globalement une meilleure sensibilité que risk(X), ainsi le premier terme de (II.8) est positif et risk(X, Y) a aussi globalement une meilleure spécificité que risk(X), et ainsi le second terme est négatif. $IDI > 0$ est donc considéré par Pencina *et al.* (2008) comme signe que risk(X, Y) est meilleur que risk(X) et ils proposent donc de tester l'hypothèse nulle $\mathcal{H}_0 : IDI = 0$ pour tester le fait que risk(X, Y) prédit mieux que risk(X).

II.3.2.2 Le NRI

Le NRI est défini par

$$\begin{aligned} NRI =& \mathbb{E}\Big[\text{sign}\big(\text{risk}(X,Y) - \text{risk}(X)\big)|D = 1\Big] \\ &+ \mathbb{E}\Big[-\text{sign}\big(\text{risk}(X,Y) - \text{risk}(X)\big)|D = 0\Big] \end{aligned} \tag{II.9}$$

avec sign(x) qui vaut 1 si $x > 0$, 0 si $x = 0$ et -1 si $x < 0$.

Alors que l'IDI compare quantitativement risk(X, Y) et risk(X), le NRI peut s'interpréter comme « comptant » combien de fois risk(X, Y) est plus pertinent que risk(X) pour classer un sujet comme malade ($D = 1$) ou non ($D = 0$). Il mesure ainsi combien l'ajout de Y dans le modèle pronostique aide à mieux classer les sujets comme malades ou non, d'où son nom d'indice de reclassement. Similairement, $NRI > 0$ est donc considéré par Pencina *et al.* (2008) comme signe que risk(X, Y) est meilleur que risk(X) et ils proposent donc aussi de tester l'hypothèse nulle $\mathcal{H}_0 : NRI = 0$ pour tester le fait que risk(X, Y) prédit mieux que risk(X).

Observant le n-échantillon $\{(X_i, Y_i, D_i), i = 1, \ldots, n\}$ et connaissant les fonctions $X \mapsto$ risk(X) et $(X, Y) \mapsto$ risk(X, Y), on estime alors aisément l'IDI et le NRI par de simples proportions, et toute autre procédure d'inférence s'en déduit aisément. Récemment, des extensions des définitions et des procédures d'inférence pour l'IDI et le NRI ont aussi été proposées pour des définitions dépendant du temps, avec $D \equiv D(t) = \mathbb{1}_{(T \leqslant t)}$, reprenant l'idée des définitions $\mathbb{C}/\mathbb{D}$ des cas et des contrôles dépendant du temps de la Section II.3.1.2.1 (Chambless *et al.*, 2011; Zheng *et al.*, 2012b; Uno *et al.*, 2013; Zhou *et al.*, 2013).

Bien que simples, plausibles et attractifs car ils mènent à plus de résultats significatifs que des tests de différences d'AUCs, on verra à la section II.3.3 que ces deux critères NRI et IDI sont en fait d'une pertinence assez contestable.

II.3.2.3 Calibration

Lorsque l'on s'intéresse à des modèles prédictifs, on souhaite souvent savoir s'ils donnent la « juste probabilité », c'est-à-dire si les prédictions sont en cohérence avec les observations. Par exemple, si on prédit un risque de 20% de rechute d'un cancer pour 1000 patients, on s'attend à observer une rechute pour environ 200 d'entre eux. C'est la notion essentielle de calibration. On suppose ici encore que l'on souhaite prédire l'événement $D = 1$ à partir de variables explicatives X.

Comme rappelé par Pepe et Janes (2013), il existe deux définitions de la calibration d'un modèle prédictif $X \mapsto \mathsf{risk}(X)$: *forte* ou *faible*. On dit que $\mathsf{risk}(\cdot)$ est calibré au sens *fort* si

$$\forall x, \quad \mathbb{P}\big(D = 1 | X = x\big) = \mathsf{risk}(x),$$

i.e. si $\mathbb{P}\big(D = 1|X\big) = \mathsf{risk}(X)$, et au sens *faible* si

$$\forall r \in [0,1], \quad \mathbb{P}\big(D = 1|\mathsf{risk}(X) = r\big) = r.$$

Notons que la définition forte implique la définition faible (la réciproque étant évidement fausse). En effet, si $\mathbb{P}\big(D = 1|X\big) = \mathsf{risk}(X)$, alors

$$\begin{aligned}
\mathbb{P}\big(D = 1|\mathsf{risk}(X) = r\big) &= \mathbb{E}(\mathbb{1}_{(D=1)}|\mathsf{risk}(X) = r) \\
&= \mathbb{E}\Big[\mathbb{E}(\mathbb{1}_{(D=1)}|X)\Big|\mathsf{risk}(X) = r\Big] \\
&= \mathbb{E}\Big[\mathbb{P}(D = 1|X)\Big|\mathsf{risk}(X) = r\Big] = \mathbb{E}\Big[\mathsf{risk}(X)\Big|\mathsf{risk}(X) = r\Big] = r
\end{aligned}$$

(le passage de la première à la seconde ligne venant du fait que la σ-algèbre engendrée par $\mathsf{risk}(X)$ est incluse dans celle engendrée par X). Bien que la définition de la calibration au sens fort soit celle que l'on souhaite pour tout modèle prédictif, elle est particulièrement difficile à vérifier en pratique dès que la dimension de X est grande, en raison du phénomène du « fléau de

la dimension » (i.e. plus la dimension p de X est grande et plus les données sont « éparpillées » dans un espace de dimension p, ce qui nécessite l'observation d'échantillons de très grandes tailles pour faire de l'inférence). En pratique, on utilise alors souvent la définition au sens faible.

Il existe assez peu de méthodes de statistique inférentielle pour évaluer la calibration d'un modèle. Le test de Hosmer et Lemeshow (1980) et son extension aux données censurées (D'Agostino et Nam, 2004) sont parfois proposés. Cependant il est de plus en plus souvent considéré comme inintéressant (Pepe et Janes, 2013) car il est peu puissant quand la taille d'échantillon n est petite, et trop sensible à des défauts de calibration sans pertinence clinique quand n est grand (Paul *et al.*, 2013).

Aujourd'hui, les visualisations graphiques sont souvent privilégiées. Les approches principales reposent essentiellement sur l'idée de partitionner la population en sous-groupes de sujets de risques estimés équivalents (en « strates »), le plus souvent en utilisant les déciles de risk(X). Des diagrammes en bâtons permettent alors de comparer le risque moyen estimé dans chaque groupe à celui observé.

Suivant la même idée, Huang *et al.* (2007) et Pepe *et al.* (2008b) ont aussi proposé l'utilisation de la « predictiveness curve », définie par $\Big\{\big(R(q), q\big), q \in [0,1]\Big\}$, avec $R(q) = \mathbb{P}\big(D = 1 \big| \text{risk}(X) = F^{-1}_{\text{risk}(X)}(q)\big)$ où $F^{-1}_{\text{risk}(X)}(q)$ est le q-ème quantile de risk(X). Son point fort étant qu'en plus d'être utile pour étudier la calibration, elle peut également être utilisée pour étudier les capacités à discriminer d'un modèle prédictif. Notamment, des connections fortes avec la notion d'AUC existent (Viallon et Latouche, 2011).

Quelques extensions et autres travaux connexes ont aussi été récément proposés, suivant des idées similaires (Gerds *et al.*, 2013a, par exemple). Pour d'avantage de discussion sur la notion de calibration et les méthodologies statistiques associées, on pourra consulter Steyerberg (2009, Chap. 13.5) ou le récent tutoriel Austin et Steyerberg (2013).

II.3.2.4 Le Brier Score

Le Brier score (Brier, 1950), ou « Expected Brier score », est une forme d'erreur quadratique (aussi appelée erreur $\mathbb{L}^2$). Pour un prédicteur risk(X), on le définit par

$$BS_{\mathsf{risk}(X)} = \mathbb{E}\Big[D - \mathsf{risk}(X)\Big]^2,$$

où $\mathbb{E}$ dénote à la fois l'espérance sur D et sur X, que l'on note parfois $\mathbb{E}_{D,X}$. Contrairement à l'AUC, c'est un score faible au Brier score qui indique de bonnes capacités pronostiques. Comme toute erreur quadratique, il se décompose sous la forme d'une somme d'un terme de biais et d'un terme de variance,

$$BS_{\mathsf{risk}(X)} = \mathbb{E}\Big[\mathbb{E}(D|X) - \mathsf{risk}(X)\Big]^2 + \mathbb{E}\Big[D - \mathbb{E}(D|X)\Big]^2, \tag{II.10}$$

puisque le double produit est nul :

$$\begin{aligned}
2 \times \mathbb{E}\Big(\Big[\mathbb{E}(D|X) - &\mathsf{risk}(X)\Big]\Big[D - \mathbb{E}(D|X)\Big]\Big) \\
&= 2 \times \mathbb{E}\left(\mathbb{E}\left\{\Big[\mathbb{E}(D|X) - \mathsf{risk}(X)\Big]\Big[D - \mathbb{E}(D|X)\Big]\Big|X\right\}\right) \\
&= 2 \times \mathbb{E}\left(\Big[\mathbb{E}(D|X) - \mathsf{risk}(X)\Big]\underbrace{\mathbb{E}\Big\{D - \mathbb{E}(D|X)\Big|X\Big\}}_{=0}\right) = 0.
\end{aligned}$$

Le premier terme de (II.10) est souvent appelé terme d'« imprécision » ou de « calibration ». Il mesure la proximité entre la prédiction risk(X) et le « vrai » risque d'événement sachant X (calibration au sens fort). Notons qu'il est nul si risk(X) est un prédicteur parfait, c'est-à-dire s'il utilise l'information X de façon optimale, en étant égal à $\mathbb{E}(D|X) = \mathbb{P}(D = 1|X)$.

Le second terme, appelé « inséparabilité », qui est égal à $\mathbb{E}(\mathsf{Var}(D|X))$, ne dépend pas de risk$(X)$, mais dépend fortement de la capacité intrinsèque de discrimination de X. En effet,

$$\begin{aligned}
\mathbb{E}(\mathsf{Var}(D|X)) = \mathbb{E}\Big[D - \mathbb{E}(D|X)\Big]^2 &= \mathbb{E}\Big[D - \mathbb{E}(D)\Big]^2 - \mathbb{E}\Big[\mathbb{E}(D) - \mathbb{E}(D|X)\Big]^2 \\
&= \mathsf{Var}(D) - \mathsf{Var}(\mathbb{E}(D|X))
\end{aligned} \tag{II.11}$$

qui n'est rien d'autre que le théorème de la variance totale (Saporta, 2006, page 73). Ainsi, plus X discrimine bien $D = 1$ de $D = 0$, plus $\mathbb{E}(D|X)$ varie, plus Var$(\mathbb{E}(D|X))$ sera grand, et plus le Brier score sera petit.

Le Brier score est une mesure assez populaire (Pepe et Janes, 2013; Steyerberg, 2009; Steyerberg *et al.*, 2010), notamment pour évaluer des prédictions d'événements du type $D \equiv D(t) = \mathbb{1}_{(T \leqslant t)}$, où T est un temps de survie et t un horizon de prédiction (Graf *et al.*, 1999; Gerds et Schumacher, 2006; Schoop *et al.*, 2008; Lawless et Yuan, 2010; Parast *et al.*, 2012, entre autres).

Il est cependant aussi critiqué pour sa difficulté d'interprétation. Notamment, les équations (II.10) et (II.11) montrent qu'il dépend fortement de la prévalence $p = \mathbb{P}(D = 1)$, entre autres à travers Var$(D) = p(1-p)$. L'impact des qualités de calibration et de discrimination de risk(X) sur le Brier score peut alors être difficile à différencier de celui de la prévalence. Pour remédier à cette difficulté d'interprétation, on peut le « normaliser » pour définir un critère de type « R^2 » (Graf *et al.*, 1999), par

$$R^2 = 1 - \frac{BS_{\mathsf{risk}(X)}}{BS_0},$$

avec BS_0 le Brier score du modèle nul n'incluant aucune covariable, c'est-à-dire le prédicteur qui prédit un risque égal à $p = \mathbb{P}(D = 1)$ pour tous les sujets (ou une estimation $\widehat{p}$ de p). On dit souvent que c'est un critère de type « variance expliquée ». En effet, si risk(X) est calibré (au sens fort), c'est à dire si risk$(X) = \mathbb{E}(D|X)$, alors le premier terme de (II.10) est nul et $BS_{\mathsf{risk}(X)} = \mathbb{E}(\mathsf{Var}(D|X))$. De plus, si le modèle nul prédit $p = \mathbb{P}(D = 1)$ pour tout le monde, alors $BS_0 = \mathsf{Var}(D) = p(1-p)$. Ainsi, dans ce cas il en résulte que :

$$\begin{aligned} R^2 &= 1 - \frac{BS_{\mathsf{risk}(X)}}{BS_0} \\ &= \frac{\mathsf{Var}(D) - \mathbb{E}(\mathsf{Var}(D|X))}{\mathsf{Var}(D)} = \frac{\mathsf{Var}\big(\mathbb{E}(D|X)\big)}{\mathsf{Var}(D)} = \frac{\mathsf{Var}\big(\mathsf{risk}(X)\big)}{\mathsf{Var}(D)}. \end{aligned}$$

Pour l'estimation du Brier score, si on observe un n-échantillon $\{(X_i, D_i), i = 1, \ldots, n\}$ et que l'on connaît la fonction $X \mapsto$ risk(X), alors on estime aisément $BS_{\mathsf{risk}(X)}$ par la moyenne

empirique

$$\widehat{BS}_{\mathsf{risk}(X)} = \frac{1}{n}\sum_{i=1}^{n}\Big[D_i - \mathsf{risk}(X_i)\Big]^2,$$

et R^2 par $\widehat{R}^2 = 1 - \widehat{BS}_{\mathsf{risk}(X)}/\widehat{BS}_0$. En présence de données censurées, des méthodes assez simples ont aussi été proposées (Graf *et al.*, 1999; Gerds et Schumacher, 2006, entre autres). Nous en reparlerons au Chapitre V.

II.3.3 Mesures de capacités pronostiques propres et impropres

II.3.3.1 Définition

Lorsque l'on s'intéresse à l'évaluation des capacités pronostiques d'un modèle, il est important d'utiliser des mesure dites « propres ». Au sens rigoureux de Gneiting et Raftery (2007), la définition d'une mesure propre requiert l'introduction d'un certain formalisme mathématique élégant, mais quelque peu complexe et insuffisant pour le cas de l'AUC. Comme Schoop *et al.* (2011), on définira donc une mesure propre heuristiquement.

Définition II.1 (Mesure de capacités pronostiques propre). *Considérons une variable binaire D, à valeur dans $\{0,1\}$, et un vecteur de covariables X utilisé pour prédire l'événement $D = 1$. On dira qu'une mesure de capacités pronostiques est* propre *si l'outil pronostique défini par la « vraie » probabilité de l'événement, i.e. $\mathbb{P}(D = 1|X)$, a le meilleur score à cette mesure parmi tous les modèles prédictifs risk(X) possibles utilisant X.*

Comme rappelé par Schoop *et al.* (2011) et Selten (1998), on peut considérer le fait qu'une mesure de capacités pronostiques soit propre comme un pré-requis minimum.

Dans cette Section, on se propose de vérifier si les diverses mesures de capacités pronostiques précédement présentées sont propres.

II.3.3.2 À propos du Brier score et de l'erreur $\mathbb{L}^1$

La décomposition du Brier score en la somme d'un terme de calibration et d'un terme d'inséparabilité de l'équation (II.10) suffit à montrer que le Brier score est propre au sens de

notre définition II.1. En effet, le Brier score est optimisé pour le prédicteur « parfait » défini par risk$(X) = \mathbb{E}(D|X) = \mathbb{P}(D = 1|X)$.

Le Brier score est une erreur dite $\mathbb{L}^2$. Non présentées dans ce manuscrit, notons que des mesures basées sur une pénalisation de type $\mathbb{L}^1$, i.e. basées sur $\mathbb{E}\big|D - \text{risk}(X)\big|$ ont aussi été proposées (Schemper et Henderson, 2000; Henderson *et al.*, 2002). Bien qu'elles soient relativement populaires en biostatistique, elles ont le défaut d'être impropres. En effet, comme rappelé par Lawless et Yuan (2010), la fonction $x \mapsto \mathbb{E}|D - x|$ est minimisée par la médiane de D (sachant X) et non pas par sa moyenne (sachant X).

II.3.3.3 À propos de l'AUC

McIntosh et Pepe (2002) ont donné le résultat suivant, qui contribue à justifier la pertinence de l'AUC pour évaluer la discrimination d'un modèle prédictif, en montrant qu'il est propre au sens de notre définition II.1.

Proposition II.1 (Optimalité du « vrai » score de risque pour l'AUC). *Considérons une variable binaire D, à valeur dans $\{0, 1\}$, et un vecteur de covariables X utilisé pour prédire l'événement $D = 1$. Alors, la « vraie » probabilité $\mathbb{P}(D = 1|X)$ a le plus grand des AUC possibles pour un modèle prédictif de l'événement $D = 1$ utilisant les covariables X.*

Démonstration. Considérons le problème du point de vue de celui d'un test statistique basé sur X avec $\mathcal{H}_0 : D = 0$ versus $\mathcal{H}_1 : D = 1$. Pour tout $\alpha \in]0, 1[$, on définit le test du rapport de vraisemblance de niveau α comme celui qui rejette $\mathcal{H}_0$ en faveur de $\mathcal{H}_1$ quand

$$\text{LR}(X) > c(\alpha) \quad \text{avec} \quad \text{LR}(X) = \frac{f_{X|D=1}(X)}{f_{X|D=0}(X)}, \tag{II.12}$$

avec les notations $f(\cdot)_{X|D=d}$, $d = 0, 1$, pour les densités conditionnelles de X sachant $D = d$, et $c(\alpha)$ tel que $\alpha = \mathbb{P}(\text{LR}(X) > c(\alpha)|D = 0)$. Ici, le risque de première espèce α correspond à 1 moins la spécificité de LR(X) en $c(\alpha)$. Le lemme de Neymann-Pearson nous indique alors que ce test est uniformément le plus puissant, c'est-à-dire que parmi les tests de niveau α il a la plus grande puissance $1 - \beta = \mathbb{P}(\text{LR}(X) > c(\alpha)|D = 1)$, i.e. la plus grande sensibilité (voir

par exemple Saporta (2006, p. 330)). Or, la formule de Bayes donne

$$\begin{aligned}\mathbb{P}(D=1|X) &= \frac{f_{X|D=1}(X)\mathbb{P}(D=1)}{f_{X|D=1}(X)\mathbb{P}(D=1)+f_{X|D=0}(X)\mathbb{P}(D=0)} \\ &= \frac{\mathsf{LR}(X)q}{\mathsf{LR}(X)q+1} \qquad \text{(II.13)}\end{aligned}$$

avec $q = \mathbb{P}(D=1)/\mathbb{P}(D=0)$ la cote (« l'odds ») de la maladie dans la population. Comme (II.13) montre que $\mathbb{P}(D=1|X)$ est une fonction monotone croissante de $\mathsf{LR}(X)$, alors (II.12) est équivalent à

$$\mathbb{P}(D=1|X) > c^*(\alpha) \quad \text{avec} \quad c^*(\alpha) = \frac{c(\alpha)q}{c(\alpha)q+1}.$$

Par conséquent, pour toute valeur de 1 moins la spécificité, la plus grande sensibilité correspondante parmi celles des prédicteurs basés sur X est atteinte par celle du prédicteur $\mathbb{P}(D=1|X)$. Ainsi, la courbe ROC est uniformément la plus haute et l'aire sous la courbe est maximale. □

McIntosh et Pepe (2002) ont en fait donné un résultat légèrement différent. Ils s'intéressaient à l'optimalité de la combinaison linéaire de biomarqueurs issue d'un modèle logistique pour prédire une maladie. Comme ils le font remarquer, une conséquence du fait que la courbe ROC (et l'AUC) est invariante par transformation monotone d'un marqueur est que tout score en bijection avec $\mathbb{P}(D=1|X)$ partage cette propriété d'optimalité. En particulier, le « vrai » prédicteur linéaire d'un modèle logistique est optimal (sous l'hypothèse d'un modèle bien spécifié). De plus, ils notent que ce résultat est valable même lorsque le modèle prédictif est estimé sur des données de type études « cas-témoins » pour lesquelles $q = \mathbb{P}(D=1)/\mathbb{P}(D=0)$, et par conséquent $\mathbb{P}(D=1|X)$, ne peuvent pas être estimés.

Par ailleurs, notons que si deux modèles prédictifs ont des AUCs égales mais des courbes ROC différentes (se croisant), alors en suivant le même raisonnement que dans la preuve de la propriété II.1, on peut montrer que les modèles ne sont pas optimaux et qu'ils peuvent donc être améliorés.

On a également la remarque suivante en présence de risques concurrents, qui justifiera le choix de la définition des cas et des contrôles au Chapitre V.

Remarque II.1. *Comme la définition Cumulative/Dynamique (2) des cas et des contrôles, $\mathbb{C}/\mathbb{D}^{(2)}$ du Tableau II.3, est telle que tous les non-cas sont des contrôles (Figure II.5), alors la proposition II.1 implique que le prédicteur $\mathbb{P}(T \leqslant t, \eta = 1|X)$ a l'AUC maximale parmi celles de tous les prédicteurs utilisant les covariables X.*

II.3.3.4 À propos de l'IDI et du NRI

L'IDI et le NRI étant définis comme les différences de capacités prédictives de deux modèles, on ne peut pas directement les qualifier de propres ou d'impropres (i.e. non propres). Cependant on peut s'apercevoir qu'ils sont tous les deux la différence d'une mesure impropre appliquée aux deux modèles. De notre point de vue, ceci en fait donc, contrairement à l'AUC, des indices peu utiles et dangereux à utiliser.

Pour s'en rendre contre facilement, on peut d'abord considérer l'exemple de Hilden et Gerds (2013) résumé au Tableau II.4. Il consiste à supposer qu'une population est composée de trois sous-groupes homogènes de risque de décès dans les 10 ans disons de 30%, 60% et 80%. On suppose aussi, par exemple, que les probabilités d'appartenir à chaque groupe sont de 50%, 30% et 20 %. Pour mettre en évidence les problèmes intrinsèques de l'IDI et du NRI, on considère alors la comparaison du modèle « parfait », prédisant le vrai risque de décès de chaque sujet de chacun des groupes, à un modèle caricatural qui prédit des risques de 0%, 100% et 100% (Tableau II.4).

Sous-groupe de la population	I	II	III
Probabilité d'appartenance au groupe	50%	30%	20 %
Vrai risque	30%	60%	80%
Risque estimé « caricaturalement »	0%	100%	100%

TABLE II.4 – Exemple pour lequel l'IDI et le NRI indiquent qu'un modèle non calibré est meilleur qu'un modèle « parfait ».

Alors, en notant risk$_C$ le risque estimé caricaturalement et risk$_V$ le vrai risque, les deux calculs ci-dessous montrent que $IDI > 0$ et $NRI > 0$ et que donc ces deux indices préféreraient

le modèle estimant le risque « caricaturalement » que « parfaitement ».

$$\begin{aligned}
IDI =& \mathbb{E}\Big[\mathsf{risk}_C - \mathsf{risk}_V \Big| D = 1\Big] + \mathbb{E}\Big[- \Big\{\mathsf{risk}_C - \mathsf{risk}_V\Big\} \Big| D = 0\Big] \\
=& \frac{(0-0.3)\times 0.3\times 0.5 + (1-0.6)\times 0.6\times 0.3 + (1-0.8)\times 0.8\times 0.2}{\mathbb{P}(D=1)} \\
&+ \frac{-(0-0.3)(1-0.3)\times 0.5 - (1-0.6)(1-0.6)\times 0.3 - (1-0.8)(1-0.8)\times 0.2}{\mathbb{P}(D=0)} \\
\approx& 21.6\%
\end{aligned}$$

avec $\mathbb{P}(D=1) = 0.5\times 0.3 + 0.3\times 0.6 + 0.2\times 0.8 = 0.49$. De même, on a

$$\begin{aligned}
NRI =& \mathbb{E}\Big[\mathsf{sign}\big(\mathsf{risk}_C - \mathsf{risk}_V\big) | D = 1\Big] + \mathbb{E}\Big[- \mathsf{sign}\big(\mathsf{risk}_C - \mathsf{risk}_V\big) | D = 0\Big] \\
=& \frac{(-1)\times 0.3\times 0.5 + 1\times 0.6\times 0.3 + 1\times 0.8\times 0.2}{\mathbb{P}(D=1)} \\
&+ \frac{1\times(1-0.3)\times 0.5 + (-1)\times(1-0.6)\times 0.3 + (-1)\times(1-0.8)\times 0.2}{\mathbb{P}(D=0)} \approx 76.0\%
\end{aligned}$$

En conclusion, cet exemple montre que l'IDI et le NRI peuvent préférer des modèles mal calibrés. Ce défaut de l'IDI et du NRI est illustré et étudié plus en détails par Pepe *et al.* (2013) et Hilden et Gerds (2013). Ludiquement, ces deux papiers nous apprennent aussi comment « tricher » avec ces critères, en construisant des modèles volontairement mal calibrés utilisant un « nouveau » marqueur parfaitement inutile (i.e. non associé à l'occurrence de l'événement d'intérêt) de telle sorte qu'ils aient pourtant un IDI et un NRI positifs comparés à un modèle plus sérieux, parfaitement calibré et n'utilisant pas ce marqueur inutile.

Enfin, notons qu'un autre exemple, proposé par Pepe et Janes (2013), est décrit au Tableau II.5. Il montre aussi que la comparaison de deux modèles parfaitement équivalents (mais différents) et tous les deux calibrés peut aussi donner des résultats de NRI surprenants : un NRI positif !

Aux vues de ces deux exemples la pertinence de l'IDI et du NRI apparaît quelque peu contestable et c'est donc sans regret qu'ils seront absents des travaux principaux de cette thèse présentés aux Chapitres III, IV et V.

$D=1$		risk(X,Y) low	risk(X,Y) med	risk(X,Y) high	
risk(X)	low	10	10	0	20
	med	5	20	10	35
	high	5	5	35	45
		20	35	45	100

$D=0$		risk(X,Y) low	risk(X,Y) med	risk(X,Y) high	
	low	500	100	0	600
	med	100	200	0	300
	high	0	0	100	100
		600	300	100	900

TABLE II.5 – Exemple décrivant deux prédicteurs performants de façon équivalente (puisque les marges horizontales et verticales sont identiques), mais pour lequel le NRI est positif. Les effectifs décrivent les effectifs des sujets subissants ou non l'événement ($D = 1$ versus $D = 0$) et ayant un risque d'événement défini par trois items, faible (low), médian (med) ou haut (high), d'après deux modèles notés risk(X, Y) et risk(X). $NRI = (10 + 10 + 0 - \{5 + 5 + 5\})/100 - (100 + 0 + 0 - \{100 + 0 + 0\})/900 = 5\%$.

II.3.3.5 A propos du C-index

Le C-index, tronqué ou non, est parfois aussi utilisé pour évaluer un modèle prédictif de $D \equiv D(t) = \mathbb{1}_{(T \leqslant t)}$ (Gerds *et al.*, 2013b). Pourtant, à notre avis et bien que cela n'ait jamais été discuté dans la littérature à notre connaissance, quel que soit le temps de troncature τ_0 (qu'il soit égal au t de $D(t)$ ou non), le C-index tronqué $\mathcal{C}(\tau_0)$ n'est pas, en général, approprié pour cet objectif, et ne définit pas une mesure propre.

Avant de détailler l'exemple « jouet » qui illustre la critique, notons qu'elle vient principalement du fait que le C-index est, par définition, un indice de discrimination global, ne tenant pas compte de l'horizon de prédiction t de $D(t) = \mathbb{1}_{(T \leqslant t)}$. Pour mesurer les capacités de discrimination globales d'un marqueur qui, par définition, ne dépend pas de l'horizon de prédiction t, le C-index est cependant, à notre avis, tout à fait intéressant.

Considérons l'exemple de distributions du temps de survie T sachant X qui suivent des lois de Weibull, et dont les paramètres de forme dépendent de X. Plus précisément, on prend l'exemple d'un marqueur X à valeurs positives et on définit le risque instantané et la survie sachant X par $\lambda(t|X) = Xt^{X-1}$ et $S(t|X) = e^{-t^X}$. L'effet de X sur la courbe de $t \mapsto S(t|X)$ est illustré par la Figure II.6. Rappelons que pour tout marqueur M tel que la probabilité d'observer des ex-aequos est nulle, on définit le C-index tronqué par $\mathcal{C}(\tau_0) = \mathbb{P}(M_i > M_j | T_i <$

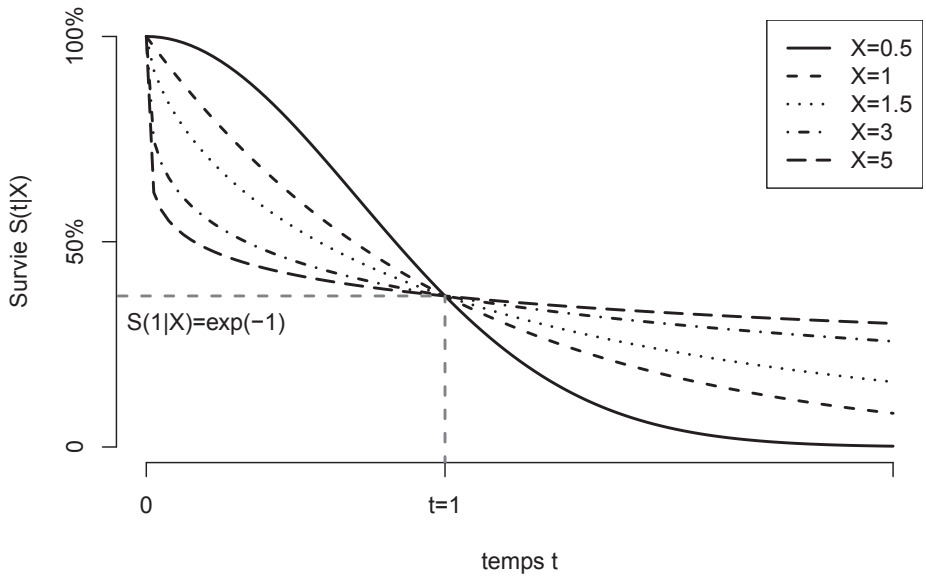


FIGURE II.6 – Effet du paramètre de forme d'une loi de Weibull sur sa fonction de survie. Courbes de $t \mapsto S(t|X) = e^{-t^X}$ pour différentes valeurs de X.

$T_j, T_i < \tau_0$), avec i et j les indices de deux sujets indépendants. Lorsque la probabilité d'observer des ex-aequos est non nulle, il est défini par $\mathcal{C}(\tau_0) = \mathbb{P}(M_i > M_j | T_i < T_j, T_i < \tau_0) + 0.5 \times \mathbb{P}(M_i = M_j | T_i < T_j, T_i < \tau_0)$ (Harrell *et al.*, 1996). Or pour cet exemple, quelle que soit la valeur de X on a $S(t = 1|X) = e^{-1}$. Donc si on définit le « marqueur » $M = S(t = 1|X)$, qui est le prédicteur « parfait » utilisant X pour prédire $D(t) = 1$, alors pour tout τ_0, $\mathcal{C}(\tau_0) = 0.5$. Or, au vu de la Figure II.6 il est clair que pour $M = X$, en général $\mathcal{C}(\tau_0) \neq 0.5$, puisque X a un fort effet sur la distribution de T sachant X.

Cet exemple montre ainsi que le C-index n'est, à notre avis, pas approprié pour évaluer des modèles prédictifs de $D \equiv D(t) = \mathbb{1}_{(T \leqslant t)}$, puisque le modèle prédictif « parfait » n'a pas toujours le C-index le plus élevé parmi tous les modèles prédictifs utilisant X.

Cependant, notons qu'on a vu en section II.3.1.3 que

$$\mathcal{C}(\tau_0) = \int_0^{\tau_0} AUC^{\mathbb{I}/\mathbb{D}}(s) w_{\tau_0}(s) ds,$$

avec une fonction $w_{\tau_0}(\cdot)$ qui ne dépend pas de la distribution du marqueur M. Or, on aussi vu à la proposition II.1 que la « vraie » probabilité d'être un cas sachant le marqueur était le modèle pronostique qui avait l'AUC maximale. Pour le cas de l'AUC utilisant les définitions $\mathbb{I}/\mathbb{D}$ des cas et des contrôles, c'est donc le risque instantané en s sachant X, i.e. $\lambda(s|X) = \lim_{ds\downarrow 0} \frac{1}{ds}\mathbb{P}(s \leqslant T < s + ds | T \geqslant s, X)$, qui a l' $AUC^{\mathbb{I}/\mathbb{D}}(s)$ maximale parmi tous les modèles prédictifs utilisant le marqueur X. Par ailleurs, pour certains modèles, comme les modèles à hasards proportionnels, les transformations $X \mapsto \lambda(s|X)$ et $X \mapsto S(t|X)$ « préservent les rangs » pour tout s, t, au sens où

$$\forall s, t, \quad X_i > X_j \quad \Leftrightarrow \quad \lambda(s|X_i) > \lambda(s|X_j) \quad \Leftrightarrow \quad S(t|X_i) > S(t|X_j).$$

Ainsi, en supposant qu'une telle classe de modèles ait générée les données, le C-index ne peut pas avoir le défaut précédemment montré.

II.3.4 Évaluer les capacités pronostiques : marginalement ou conditionnellement ?

En pratique, pour construire des modèles utilisant de nombreuses covariables, on utilise le plus souvent des prédicteurs linéaires (Andersen et Skovgaard, 2010). Il n'est alors pas toujours évident de faire certains choix de modélisation comme ceux d'inclure ou non des termes d'interaction entre certaines covariables ou de stratifier ou non sur d'autres. Par exemple, pour prédire un cancer du sein comme Gail (2008), on peut utiliser un prédicteur linéaire défini par une combinaison de marqueurs génétiques et de l'âge, ou bien stratifier sur l'âge et construire des prédicteurs linéaires conditionnels à l'âge. L'âge étant l'un des facteurs de risque connu important, les choix de modélisation relatifs à l'âge méritent une attention particulière. De même, pour la prédiction de la démence, on pourrait utiliser un unique prédicteur issu d'une combinaison linéaire de plusieurs tests psychométriques et de l'âge, ou bien envisager plusieurs combinaisons linéaires de tests psychométriques conditionnelles à l'âge.

Similairement, les capacités prédictives peuvent être évaluées « marginalement », c'est-à-dire sur l'ensemble de la population, ou « conditionnellement » à l'exposition ou non à des facteurs de risques importants, tels que l'âge. Kerr et Pepe (2011) ont récemment discuté en détail ces deux approches qui sont toutes les deux intéressantes. Kerr et Pepe (2011) concluent que le choix de l'une ou de l'autre n'est pas toujours évident et devrait probablement être plus souvent discuté. En particulier, la définition précise des questions cliniques auxquelles les analyses statistiques tentent d'apporter des réponses et la description de la richesse des données disponibles devraient probablement plus souvent guider le choix de l'une ou de l'autre.

Dans la suite de cette thèse, on ne considérera que des méthodologies statistiques pour l'évaluation de capacités pronostiques marginales. L'évaluation de capacités pronostiques conditionnelles pourra néanmoins être réalisée par les méthodes décrites au moyen de simples analyses de sous-groupes (si les effectifs sont suffisamment grands pour le permettre).

Notons cependant qu'une littérature croissante s'intéresse à l'évaluation de capacités pronostiques conditionnelles en utilisant des modèles de régression semiparamétriques, pour les courbes ROC par exemple. Entre autres, citons Cai et Pepe (2002); Dodd et Pepe (2003) et Thas *et al.* (2012) pour les courbes ROC et l'AUC en général et Cai *et al.* (2006); Song et Zhou (2008); Hung et Chiang (2010); Zheng *et al.* (2012a) et Zhou *et al.* (2013) pour les courbes ROC et l'AUC dépendant du temps en présence de données de survie.

II.3.5 Validation interne et validation externe

Un modèle pronostique qui ne prédit bien que sur les données avec lesquelles il a été estimé est, d'un point de vue pratique, inutile. C'est le problème du sur-ajustement (« overfiting ») (Steyerberg, 2009, Chapitre 5). Pour qu'un modèle pronostique soit utile, il faut qu'il ait de bonnes capacités pronostiques quand il est appliqué à de « nouveaux » sujets (Steyerberg, 2009). Pour valider ces capacités à prédire correctement pour de nouveaux sujets, on distingue essentiellement deux types de méthode de validation, appelées interne et externe. Dans cette thèse, on ne se placera que dans le cadre d'une validation externe. On présente cependant les deux approches ci-après, et on rediscutera de la validation interne dans les perspectives du

Chapitre VI.

Par simplicité, dans cette section on ne détaillera l'évaluation de capacités pronostiques que pour le Brier score, comme c'est le cas dans la majorité de la littérature. Les raisonnements peuvent généralement s'appliquer pour d'autres mesures de capacités pronostiques. La généralisation n'est cependant pas toujours immédiate.

Comme précédemment, on suppose dans cette section qu'on souhaite utiliser un vecteur de covariables X, pour prédire l'événement binaire D.

II.3.5.1 Validation externe

Une pratique courante consiste à estimer des modèles pronostiques sur une première base de données, disons de m sujets, appelée échantillon d'apprentissage et notée $\mathcal{L}_m = \{(X_k, D_k), k = 1, \ldots, m\}$ ($\mathcal{L}$ pour « Learning »), et à estimer les capacités pronostiques à l'aide d'une seconde, disons de n sujets, appelée échantillon de validation, et notée $\mathcal{V}_n = \{(X_i, D_i), i = 1, \ldots, n\}$ ($\mathcal{V}$ pour « Validation »).

On construit alors un modèle pronostique $\widehat{R}_{\mathcal{L}_m} = R(\mathcal{L}_m)$, où $R : \mathcal{L}_m \mapsto R(\mathcal{L}_m)$ est la fonction de construction du modèle à partir des données $\mathcal{L}_m$, puis on l'évalue conditionnellement à l'échantillon d'apprentissage $\mathcal{L}_m$, en estimant

$$BS_{\widehat{R}_{\mathcal{L}_m}} = \mathbb{E}_{D,X}\left[\left\{D - \widehat{R}_{\mathcal{L}_m}(X)\right\}^2 \middle| \mathcal{L}_m\right], \tag{II.14}$$

à l'aide des données $\mathcal{V}_n$, généralement par l'estimateur empirique

$$\widehat{BS}_{\widehat{R}_{\mathcal{L}_m}} = \frac{1}{n}\sum_{i\in\mathcal{V}_n}\left\{D_i - \widehat{R}_{\mathcal{L}_m}(X_i)\right\}^2.$$

L'estimation par validation externe est particulièrement souhaitable d'un point de vue clinique et elle est naturelle lorsque l'on dispose de deux bases de données relativement similaires.

Notons que, par définition, l'estimation $\widehat{BS}_{\widehat{R}_{\mathcal{L}_m}}$ dépend à la fois de $\mathcal{L}_m$ et de $\mathcal{V}_n$. La dépendance en $\mathcal{V}_n$ peut être quantifiée en construisant des intervalles de confiance de $BS_{\widehat{R}_{\mathcal{L}_m}}$. Cependant, la dépendance en $\mathcal{L}_m$, souhaitable pour l'interprétation, ne peut pas être quantifiée.

II.3.5.2 Validation interne

Lorsqu'une unique base de données est disponible, il est fréquent de scinder la base de données en deux, pour créer « artificiellement » les échantillons d'apprentissage et de validation à partir d'un unique échantillon (Steyerberg, 2009). Un inconvénient est cependant que les résultats peuvent dépendre fortement de la façon (aléatoire ou non) de les scinder. Des méthodes alternatives sont alors de plus en plus privilégiées.

De nombreuses méthodes de validation interne existent lorsqu'on dispose uniquement d'un seul échantillon $\mathcal{B}_N = \{(X_k, D_k), k = 1, \ldots, N\}$. Avant d'en mentionner quelques-unes, il est important d'insister sur une différence notable entre les approches de validation interne et externe. Pour des raisons d'identifiabilité, elles n'estiment souvent pas la même quantité : les méthodes de validation interne essaient généralement d'estimer une quantité proche d'un Brier score moyen,

$$\mathbb{E}_{\mathcal{B}_N}\left[BS_{\widehat{R}_{\mathcal{B}_N}}\right] = \mathbb{E}_{\mathcal{B}_N}\left(\mathbb{E}_{D,X}\left[\left\{D - \widehat{R}_{\mathcal{B}_N}(X)\right\}^2 \middle| \mathcal{B}_N\right]\right), \tag{II.15}$$

(avec les notations $\mathbb{E}_{D,X}$ pour l'espérance sur le couple aléatoire (D, X) et $\mathbb{E}_{\mathcal{B}_N}$ l'espérance sur l'échantillon aléatoire $\mathcal{B}_N$, de taille N). En notant $r_N(X) = \mathbb{E}_{\mathcal{B}_N}\left[\widehat{R}_{\mathcal{B}_N}(X)\right]$ la prédiction moyenne (sachant X) estimée sur tous les échantillons $\mathcal{B}_N$ de taille N, Gerds et van de Wiel (2011) rappellent qu'on a la décomposition suivante pour le Brier score moyen,

$$\begin{aligned}\mathbb{E}_{\mathcal{B}_N}\left[BS_{\widehat{R}_{\mathcal{B}_N}}\right] =& \mathbb{E}_{\mathcal{B}_N}\left(\mathbb{E}_{D,X}\left[\left\{D - \widehat{R}_{\mathcal{B}_N}(X)\right\}^2 \middle| \mathcal{B}_N\right]\right)\\ =& \mathbb{E}_{D,X}\left[\left\{D - r_N(X)\right\}^2\right]\\ &+ \mathbb{E}_{\mathcal{B}_N}\left(\mathbb{E}_X\left[\left\{\widehat{R}_{\mathcal{B}_N}(X) - r_N(X)\right\}^2 \middle| \mathcal{B}_N\right]\right)\end{aligned} \tag{II.16}$$

puisque le double produit s'annule :

$$\begin{aligned}&2 \times \mathbb{E}_{\mathcal{B}_N}\left(\mathbb{E}_{D,X}\left[\left\{\widehat{R}_{\mathcal{B}_N}(X) - r_N(X)\right\}\left\{D - r_N(X)\right\} \middle| \mathcal{B}_N\right]\right)\\ &= 2 \times \mathbb{E}_{D,X}\left[\left\{D - r_N(X)\right\} \underbrace{\mathbb{E}_{\mathcal{B}_N}\left[\widehat{R}_{\mathcal{B}_N}(X) - r_N(X)\right]}_{=0}\right].\end{aligned}$$

L'équation (II.16) peut alors s'interpréter comme la somme du Brier score du modèle moyen $X \mapsto r_N(X)$ et de la moyenne, sur l'espace des covariables X, de la variance du risque estimé $\widehat{R}_{\mathcal{B}_N}(X)$ autour de $r_N(X)$. Remarquons alors qu'on retrouve le fameux dilemme « Biais-Variance » (voir par exemple Geman *et al.* (1992)), puisque le premier terme de $\mathbb{E}_{\mathcal{B}_N}\left[BS_{\widehat{R}_{\mathcal{B}_N}}\right]$ décroît lorsque $r_N(x)$ est proche de $\mathbb{E}(D|X=x)$ pour tout x, alors que simultanément, pour tout x, le second terme croît à mesure que $\widehat{R}_{\mathcal{B}_N}(x)$ varie autour de $r_N(x)$.

La minimisation de l'erreur de prédiction moyenne est donc un compromis entre une minimisation du biais de l'estimation du « vrai risque » et une minimisation de la variance de son estimation. La minimisation des deux simultanément n'est en effet généralement pas possible en pratique, puisque la complexité du monde réel implique souvent que le « vrai risque » est une fonction compliquée. Un estimateur du risque non biaisé, mais ayant une grande variabilité (car ayant beaucoup de paramètres par exemple) peut donc ne pas être préférable à un estimateur biaisé (plus simple, avec moins de paramètres). Ainsi, insistons sur le fait qu'on travaille généralement à taille d'échantillon N fixé et que, lorsqu'on compare deux stratégies de construction de modèles pronostiques, par exemple, les conclusions pourraient éventuellement s'inverser si N changeait.

Comme rappelé par Efron et Tibshirani (1997) entre autres, ce sont les capacités pronostiques conditionnelles aux données à notre disposition pour construire le modèle pronostique (II.14), plus que les capacités moyennes (II.15), qui nous intéressent le plus souvent. Avec des données de cohorte comme celles de Paquid, Trois-Cités ou Framingham par exemple, c'est l'estimation de modèles de risque conditionnel à ces données, et non pas à des données moyennes, dont elles représenteraient un cas particulier, qui nous intéresse. Cependant Efron et Tibshirani (1997) notent que ces capacités conditionnelles sont difficilement estimables avec un unique échantillon, contrairement aux capacités moyennes (II.15).

Parmi les méthodes les plus populaires pour estimer (II.15), on trouve la « cross-validation ». Brièvement, la « K-fold cross-validation » consiste à

1. Partitionner l'échantillon $\mathcal{B}_N$ en K échantillons de taille égale $\mathcal{B}^{(k)}$, tels que $\mathcal{B}_N = \cup_{k=1}^{K} \mathcal{B}^{(k)}$.
2. Pour $k = 1, \ldots, K$:
   a) Assembler $K-1$ des sous-échantillons créés pour construire une base d'apprentissage $\mathcal{L}_m^{(-k)} = \mathcal{B}_N \backslash \mathcal{B}^{(k)}$ et l'utiliser pour construire le modèle pronostique $\widehat{R}_{\mathcal{L}_m^{(-k)}}$
   b) Évaluer ses capacités pronostiques sur $\mathcal{B}^{(k)}$, en calculant
   $$\widehat{BS}_{\widehat{R}_{\mathcal{L}_m^{(-k)}}} = \frac{1}{n} \sum_{i \in \mathcal{B}^{(k)}} \left\{ D_i - \widehat{R}_{\mathcal{L}_m^{(-k)}}(X_i) \right\}^2 .$$
3. Calculer les capacités moyennes par $\overline{\widehat{BS}}_{\widehat{R}_{\mathcal{L}_m^{(-k)}}} = \frac{1}{K} \sum_{k=1}^{K} BS_{\widehat{R}_{\mathcal{L}_m^{(-k)}}}$.

Un choix important pour cette méthode est celui du paramètre K. Plus il sera grand et plus la taille des échantillons d'apprentissage $\mathcal{L}_m^{(-k)}$ seront proches de N, et donc plus l'estimation $\overline{\widehat{BS}}_{\widehat{R}_{\mathcal{L}_m^{(-k)}}}$ sera une estimation raisonnable de $\mathbb{E}_{\mathcal{B}_N}\left[BS_{\widehat{R}_{\mathcal{B}_N}}\right]$. Les propriétés du choix « optimal » de K sont compliquées et encore mal connues (Arlot et Celisse, 2010). En pratique le choix de K est donc souvent guidé par le temps de calcul des procédures de construction du modèle pronostique. Le choix de $K = N$ mène à l'approche « leave-one-out », coûteuse en temps de calcul. Pour le réduire, des approximations asymptotiques ont, par exemple, été proposées pour le cas de modèles paramétriques (Commenges *et al.*, 2012). Pour le cas de l'AUC, Airola *et al.* (2011) ont récemment comparé plusieurs approches de cross-validation empiriquement, et leurs recommandations sont en faveur du « leave-pair-out ». Enfin, notons que pour une revue générale très approfondie sur la cross-validation, on pourra consulter Arlot et Celisse (2010).

Dans les années 2000, des méthodes utilisant la technique « Bootstrap » ont aussi été proposées pour améliorer la cross-validation, notamment par Efron et Tibshirani (1997). Depuis, l'approche a été reprise et adaptée à la présence de données censurées, notamment par Gerds et Schumacher (2007) et Schumacher *et al.* (2007) pour le Brier score, et Foucher et Danger (2012) pour les courbes ROC dépendant du temps.

Enfin, notons que la construction d'intervalles de confiance et de tests pour l'AUC ou le Brier score par exemple, n'est pas aisée lorsque les procédures d'estimation incluent des approches

de cross-validation. Uno *et al.* (2007); Parast *et al.* (2012); Commenges *et al.* (2012), entre autres, ont cependant commencé à s'y intéresser dans certains cas de modèles pronostiques paramétriques ou semiparamétriques, tandis que van de Wiel *et al.* (2009) ont récemment proposé un algorithme assez général.

Dans les travaux principaux de cette thèse, on ne fera pas de validation interne. Une validation externe sera cependant présentée au Chapitre V.

II.4 Estimation en présence de données censurées

II.4.1 Le problème des données censurées

Dans cette thèse, on s'intéresse principalement à l'estimation de courbes ROC dépendant du temps. Essentiellement, on considérera des définitions de cas ($D = 1$) et de contrôles ($D = 0$) avec $D \equiv D(t) = \mathbb{1}_{(T \leqslant t)}$ au Chapitre III et avec $D \equiv D(t) = \mathbb{1}_{(T \leqslant t, \eta = 1)}$ en présence de risques concurrents au Chapitre IV, qui correspondent à des définitions *cumulatives/dynamiques* selon la terminologie de la Section II.3.1.2.

La principale difficulté pour faire de l'inférence viendra du fait qu'on observe des données censurées $\{(\widetilde{T}_i, \Delta_i), i = 1, \ldots, n\}$, avec $\widetilde{T}_i = \min(T_i, C_i)$ et $\Delta_i = \mathbb{1}_{(T_i \leqslant C_i)}$. En effet, que l'on soit en présence de risques concurrents ou non, la difficulté illustrée par la Figure II.7 est similaire : pour les sujets perdus de vue avant le temps t, on ne sait pas si $D \equiv \mathbb{1}_{(T \leqslant t)}$ est égal à 1 ou 0. On ne peut donc pas estimer les sensibilités, les spécificités et les courbes ROC comme c'est décrites en Section II.3.1.1, en calculant de simples proportions issues de tableaux de contingence (Tableau II.1).

La difficulté sera essentiellement la même lorsqu'on considérera $D \equiv D(s,t) = \mathbb{1}_{(s < T \leqslant s+t, \eta = 1)}$ au Chapitre V, pour s'adapter à la présence de marqueurs longitudinaux, puisque considérer le temps « résiduel » $T' = T - s$ ramène à la même situation.

Techniquement, l'objectif principal de cette thèse sera de proposer des procédures d'inférence non-paramétriques pour les courbes ROC en présence de telles données censurées.

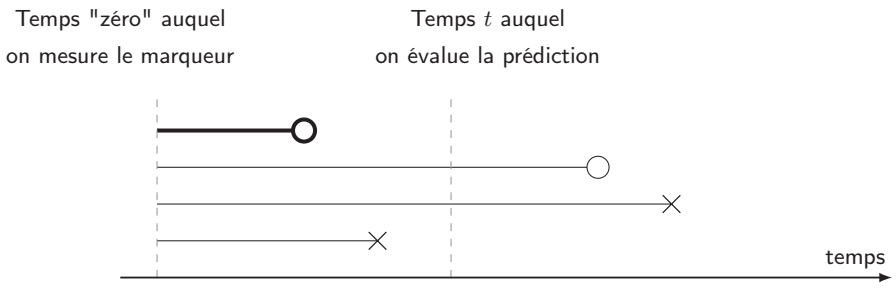


FIGURE II.7 – Le problème des données censurées et l'observation des cas et contrôles $\mathbb{C}/\mathbb{D}$. Chaque ligne horizontale représente le temps de survie observé $\widetilde{T}_i$ d'un sujet i. Les croix représentent des observations non censurées, les ronds des observations censurées. Contrairement aux autres sujets, le premier sujet, en gras, est perdu de vue avant t. On ne peut donc pas savoir s'il est un cas ou un contrôle en t.

II.4.2 Méthodes d'estimation en présence de données censurées

Pour estimer les courbes ROC en présence de données censurées, on peut utiliser des méthodes complétement paramétriques (Rizopoulos, 2011) ou semiparamétriques (Chambless et Diao, 2006; Song et Zhou, 2008; Zheng *et al.*, 2012a, entre autres). De notre point de vue, il est cependant généralement préférable d'adopter des approches non paramétriques, qui font moins d'hypothèses, et qui fournissent généralement des estimations non biaisées. Elles permettent ainsi une comparaison « juste » de modèles potentiellement très différents.

Dans cette section, on présente donc succintement l'idée des principales méthodes non-paramétriques permettant d'estimer des courbes ROC en présence de données censurées.

II.4.2.1 Les approches utilisant des estimateurs de Kaplan-Meier conditionnels

Lorsque l'on s'intéresse à des estimations non paramétriques en présence de données censurées, un estimateur « incontournable » est l'estimateur de Kaplan-Meier (Kaplan et Meier, 1958). Rappelons que, pour tout temps t, il estime $S(t) = \mathbb{P}(T > t)$, et que pour un n-échantillon $\{(\widetilde{T}_i, \Delta_i), i = 1, \ldots, n\}$, avec $\widetilde{T}_i = \min(T_i, C_i)$ et $\Delta_i = \mathbb{1}_{(T_i \leqslant C_i)}$, il est défini

par

$$\widehat{S}(t) = \prod_{i=1}^{n} \left(1 - \frac{\Delta_{(i)}}{n-i+1}\right)^{\mathbb{1}_{(\tilde{T}_{(i)} \leqslant t)}}, \tag{II.17}$$

avec les indices ordonnés $(i), i = 1, \ldots, n$, tels que $\widetilde{T}_{(1)} \leqslant \cdots \leqslant \widetilde{T}_{(n)}$, où les ex-aequos entre les temps de censure ou entre les temps d'événement sont ordonnés arbitrairement, et où les ex-aequos entre un temps censuré et un temps non censuré sont ordonnés tel que le temps non censuré précède le censuré.

Historiquement, Heagerty *et al.* (2000) proposèrent les premiers des estimateurs de la courbe ROC dépendant du temps en présence de données censurées. Pour tenir compte des données censurées dans l'estimation de la sensibilité (et de la spécificité), pour un seuil c donné, leur première idée consistait à utiliser l'estimateur de Kaplan-Meier (II.17) sur des sous-échantillons de sujets ayant (ou n'ayant pas) des valeurs du marqueur supérieur à c. Puis, motivés par quelques imperfections de cette première approche, comme nous le reverrons au Chapitre III, ils proposèrent ensuite l'utilisation d'un estimateur non paramétrique de la survie conditionnelle $S(t|X) = \mathbb{P}(T > t|X)$, où X est le marqueur pronostique. Beran (1981) et Akritas (1994) avaient introduit de tels estimateurs, qui peuvent, comme l'estimateur de Kaplan-Meier, être considérés comme des estimateurs du maximum de vraissemblance non paramétriques. Brièvement, pour tout x, l'idée des deux consiste à estimer $S(t|X = x)$ à partir de l'estimateur de Kaplan-Meier (II.17) sur des sous-échantillons de sujets « proches » d'un sujet ayant une valeur du marqueur égale à x. Les sujets « proches » peuvent être définis comme ceux ayant une valeur du marqueur égale à x à une « petite tolérance » près (Beran, 1981) ou comme étant la petite fraction des sujets de l'échantillon ayant la valeur au marqueur la plus proche de x (i.e. les « plus proches voisins ») (Akritas, 1994). Cette approche pour estimer des courbes ROC dépendant du temps a ensuite été reprise par Saha et Heagerty (2010), Chiang et Hung (2010) et Cai *et al.* (2011), entre autres. Pour les deux méthodes, le choix d'un paramètre de lissage (une fenêtre, « bandwidth ») est néanmoins nécessaire pour définir les sujets « proches », et il n'est pas toujours évident en pratique. Les procédures de choix optimal sont en effet complexes et difficiles à utiliser en pratique (Dabrowska, 1992; Akritas, 1994).

II.4.2.2 L'approche IPCW

L'appoche IPCW (pour « Inverse Probability of Censoring Weighting ») a aussi été proposée pour estimer les courbes ROC dépendant du temps et des Brier scores en présence de données censurées (Uno *et al.*, 2007; Hung et Chiang, 2010; Graf *et al.*, 1999, entre autres). Brièvement, l'idée est de considérer le problème des données censurées comme un cas particulier de données manquantes, et d'utiliser des estimateurs pondérés, similaires à ceux de Horvitz et Thompson (1952). Très populaires en théorie des sondages, les estimateurs de type « Horvitz-Thompson » pour données censurées ont été largement popularisés par Robins et Rotnitzky (1992) et par Satten et Datta (2001) en biostatistique. En particulier, Satten et Datta (2001) rappellent que l'estimateur de Kaplan-Meier (II.17) peut se réécrire comme un estimateur IPCW :

$$\widehat{S}(t) = 1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(\widetilde{T}_i \leqslant t)} \frac{\Delta_i}{\widehat{G}(\widetilde{T}_i)}, \tag{II.18}$$

où pour tout u, $\widehat{G}(u)$ est l'estimateur de Kaplan-Meier de $G(u) = \mathbb{P}(C > u)$, défini par (II.17), en remplaçant Δ_i par $1 - \Delta_i$ pour tout i. Intuitivement, la pondération $\Delta_i/\widehat{G}(\widetilde{T}_i)$ est justifiée par le fait que

$$\begin{aligned}
\mathbb{E}\left\{\mathbb{1}_{(\widetilde{T}_i \leqslant t)} \frac{\Delta_i}{G(\widetilde{T}_i)}\right\} &= \mathbb{E}\left\{\mathbb{E}\left[\mathbb{1}_{(T_i \leqslant t)} \frac{\mathbb{1}_{(T_i \leqslant C_i)}}{G(T_i)} \middle| T_i\right]\right\} \qquad (\text{car} \quad \Delta_i = 1 \Rightarrow \widetilde{T}_i = T_i) \\
&= \mathbb{E}\left\{\mathbb{1}_{(T_i \leqslant t)} \frac{\mathbb{E}\left[\mathbb{1}_{(T_i \leqslant C_i)} \middle| T_i\right]}{G(T_i)}\right\} = \mathbb{E}\left\{\mathbb{1}_{(T_i \leqslant t)} \frac{G(T_i)}{G(T_i)}\right\} = \mathbb{P}\left(T \leqslant t\right),
\end{aligned}$$

où l'hypothèse d'une censure indépendante est utilisée pour avoir $\mathbb{E}\left[\mathbb{1}_{(T_i \leqslant C_i)} \middle| T_i\right] = G(T_i)$. Cette approche IPCW est étroitement liée à celle des « intégrales de Kaplan-Meier », pour lesquelles des résultats très généraux ont été montrés (Stute, 1993, 1995; Gill, 1994, Chapitre 8). Nous verrons aux Chapitres III,IV et V que pour l'estimation de courbes ROC, d'AUC et de Brier score il existe cependant une légère différence avec l'approche des intégrales de Kaplan-Meier. En effet, le poids utilisé pour un sujet i n'est pas toujours égal à l'accroissement de $\widehat{S}(\cdot)$ en $\widetilde{T}_i$, i.e. à $\Delta_i/\widehat{G}(\widetilde{T}_i)$. Pour s'adapter au mieux au temps d'horizon de prédiction t, le poids d'un sujet i sera généralement choisi de la forme

$$\widehat{W}_i(t) = \frac{\mathbb{1}_{(\widetilde{T}_i > t)}}{\widehat{G}(t)} + \frac{\mathbb{1}_{(\widetilde{T}_i \leqslant t)}\Delta_i}{\widehat{G}(\widetilde{T}_i)}.$$

Enfin, notons que plusieurs auteurs ont proposé d'utiliser des covariables pour estimer les poids (van der Laan et Robins, 2003; Gerds et Schumacher, 2006; Tsiatis, 2006; Datta *et al.*, 2010; Lopez, 2011, entre autres). L'idée est de tenir compte de l'éventuelle dépendance entre le temps de censure C et les covariables. Les poids peuvent alors être estimés en utilisant les estimateurs de Kaplan-Meier conditionnels de Beran (1981) ou Akritas (1994), ou bien en utilisant des modèles de régression semiparamétriques. Nous en reparlerons au Chapitre III et IV. Par ailleurs, certains auteurs suggèrent que cette modification du calcul des poids peut aussi améliorer l'efficacité des estimateurs (au sens d'une variance plus faible) (van der Laan et Robins, 2003; Tsiatis, 2006).

II.4.2.3 Andersen-Klein « jackknife pseudo values »

Enfin, notons qu'une approche basée sur les pseudo-valeurs du « jackknife » a récemment été proposée pour estimer des modèles semiparamétriques très généraux, en présence de données censurées (Andersen *et al.*, 2003). On pourra consulter Andersen et Perme (2010) pour une introduction didactique à la méthode et à ses applications, et Graw *et al.* (2009) pour un premier travail plus théorique sur ses propriétés asymptotiques.

A notre connaissance, cette approche n'a jamais été utilisée pour estimer des courbes ROC dépendant du temps. Cependant, elle a récemment été proposée pour estimer non paramétriquement des Brier scores (Cortese *et al.*, 2013), ou pour étudier la calibration d'un modèle pronostique (Gerds *et al.*, 2013a). Bien que non étudiée dans cette thèse, notons qu'elle est généralement considérée comme une alternative intéressante à l'approche IPCW.

II.5 Statistique asymptotique : la boîte à outils

La statistique asymptotique consiste essentiellement à étudier le comportement d'estimateurs ou de tests lorsque la taille de l'échantillon n grandit indéfiniment. Elle permet de donner quelques résultats limites, c'est-à-dire quand $n \to \infty$, que l'on utilisera comme des approximations en pratique, lorsqu'on peut considérer n « assez grand ».

L'objectif de cette section est de rappeler quelques idées et outils de statistique asymptotique. Sans ambition d'une présentation exhaustive et d'une grande rigueur mathématique, on souhaite ici uniquement présenter brièvement, et informellement, les principaux outils et les principales idées utiles aux travaux des Chapitres IV et V.

Pour plus de détails et des preuves des résultats, on pourra consulter les livres de références suivants : Serfling (1980), Andersen *et al.* (1993), Kowalski et Tu (2008) et Martinussen et Scheike (2006) et van der Vaart (1998), desquels cette section est largement inspirée.

II.5.1 Quelques idées et outils incontournables

À la base des raisonnements de statistique asymptotique, on trouve les notions de convergence, dont deux définitions sont particulièrement utiles en statistique appliquée. On dira qu'un vecteur aléatoire X_n converge vers X :

- « En distribution », si $\forall x \quad \mathbb{P}(X_n \leqslant x) \to \mathbb{P}(X \leqslant x)$ quand $n \to \infty$ (on dit aussi en loi, ou faiblement). On la notera $X_n \xrightarrow{\mathcal{D}} X$.
- « En probabilité », si $\forall \epsilon > 0$, $\mathbb{P}(\|X_n - X\| > \epsilon) \to 0$ quand $n \to \infty$, où $\|X\|$ désigne la norme de X. On la notera $X_n \xrightarrow{\mathbb{P}} X$.

Pour raccourcir les expressions, on utilisera la notation $o_p(r_n)$, et particulièrement $X_n = o_p(r_n)$ pour signifier que $X_n = Y_n r_n$ avec $Y_n \xrightarrow{\mathbb{P}} 0$.
Le symbole $O_p(1)$ est aussi utilisé pour signifier « borné en probabilité », i.e. que $\forall \epsilon > 0$ il existe un entier N_ϵ et une constante M_ϵ tels que $\forall n > N_\epsilon \quad \mathbb{P}\big(|X_n| > M_\epsilon\big) < \epsilon$. On notera aussi $O_p(r_n)$ et en particulier $X_n = O_p(r_n)$ pour signifier que $X_n = Y_n r_n$, avec $Y_n = O_p(1)$.

En pratique, il est courant de souhaiter déduire les propriétés d'un estimateur à partir de connaissances portant sur des estimateurs plus simples dont il est dérivé. Pour cela, on dispose de quelques outils essentiels, dont font partie le « *continuous mapping theorem* » et le lemme de Slutsky.

Théorème II.1 (Continuous mapping theorem). *Soit $g : \mathbb{R}^k \mapsto \mathbb{R}^m$ une fonction continue en*

tout point d'un ensemble E tel que $\mathbb{P}(X \in E) = 1$.

1. *Si $X_n \xrightarrow{\mathcal{D}} X$, alors $g(X_n) \xrightarrow{\mathcal{D}} g(X)$.*
2. *Si $X_n \xrightarrow{\mathbb{P}} X$, alors $g(X_n) \xrightarrow{\mathbb{P}} g(X)$.*

Lemme II.1 (Slutsky). *Soient X_n, X et Y_n des vecteurs ou des variables aléatoires. Si $X_n \xrightarrow{\mathcal{D}} X$ et, pour une constante c, $Y_n \xrightarrow{\mathcal{D}} c$, alors :*

1. $X_n + Y_n \xrightarrow{\mathcal{D}} X + c$,
2. $X_n Y_n \xrightarrow{\mathcal{D}} cX$,
3. $X_n Y_n^{-1} \xrightarrow{\mathcal{D}} c^{-1} X$, *dès que* $c \neq 0$.

Dérivé des deux premiers, un troisième outil très utile repose sur l'utilisation de développements de Taylor au premier ordre.

Proposition II.2. *Soient $g : \mathbb{R}^m \mapsto \mathbb{R}$ une fonction continue et différentiable en $\mu = (\mu_1, \ldots, \mu_m)$ et un vecteur aléatoire $X_n = (X_{n1}, \ldots, X_{nm})$ tel que $\forall j \in \{1, \ldots, m\}$, $\sqrt{n}\Big(X_{nj} - \mu_j\Big) = O_p(1)$. Alors,*

$$\sqrt{n}\Big(g(X_n) - g(\mu)\Big) = \sqrt{n} \sum_{j=1}^{m} \frac{\partial}{\partial x_j} g(\mu)\Big(X_{nj} - \mu_j\Big) + o_p(1).$$

Démonstration. C'est une conséquence des Théorèmes 2 et 5 du Chapitre 1 de Kowalski et Tu (2008). □

Cette propriété est essentiellement une version de la célèbre « méthode delta » qui, dans sa forme la plus usuelle, affirme que pour toute fonction $g : \mathbb{R} \mapsto \mathbb{R}$ dérivable en θ, de dérivée g' telle que $g'(\theta) \neq 0$, si $\widehat{\theta} \sim \mathcal{N}(\theta, \sigma_\theta^2/n)$ alors $g(\widehat{\theta}) \sim \mathcal{N}\big(g(\theta), g'(\theta)^2 \sigma_\theta^2/n\big)$.

Enfin, pour utiliser ces outils et dériver les propriétés d'estimateurs originaux, il est souvent nécessaire de partir des propriétés connues d'estimateurs plus usuels, comme par exemple, de celles connues pour les moyennes de variables i.i.d. En effet, dans le cas d'un échantillon i.i.d $X_1, \ldots, X_n$, la loi des grands nombres affirme, entre autres et sous des conditions faibles,

que la moyenne $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ converge en probabilité vers l'espérance $\mathbb{E}(X_1)$. Le théorème central limite permet quant à lui de connaître la distribution de l'écart, parfois appelé « erreur », entre l'estimateur $\overline{X}_n$ et la quantité qu'il estime $\mathbb{E}(X_1)$, via le résultat $\sqrt{n}\Big(\overline{X}_n - \mathbb{E}(X_1)\Big) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \mathrm{Var}(X_1)\Big)$. En particulier, il indique la « vitesse de convergence » de l'estimateur $\overline{X}_n$ vers $\mathbb{E}(X_1)$, le fameux $\sqrt{n}$.

Ces deux résultats fondamentaux de la statistique inférentielle ne sont cependant pas toujours suffisants. C'est pourquoi les deux sous-sections suivantes présentent quelques résultats de convergence complémentaires qui nous seront utiles.

II.5.2 U-statistique et projection de Hájek

Introduit par Hoeffding (1948), les U-statistiques (« U » pour « Unbiased ») peuvent être considérées comme des généralisations des moyennes. Le plus souvent elles sont asymptotiquement gaussiennes. Pour le montrer et pour calculer leur variance, un outil très puissant est le principe de projection de Hájek (van der Vaart, 1998, Sec. 11.3).

Notons qu'il existe plusieurs types de U-statistiques : celles à 1, 2 ou plus généralement $K > 2$ échantillons. Dans tous les cas, leur étude est très similaire, puisqu'elle est essentiellement basée sur le même principe de projection. Seuls les résultats de celles à un échantillon nécessaires dans la suite sont ici présentés.

Soient $X_1, X_2, \ldots, X_n$ des vecteurs ou des variables aléatoires i.i.d issus d'une distribution quelconque et $m < n$. Étant donnée une fonction h connue, on considère l'estimation d'un paramètre quelconque θ, défini comme :

$$\theta = \mathbb{E}h(X_1, \ldots, X_m).$$

La fonction h est appelée noyau et est supposée symétrique en ses m arguments. Une U-statistique, de noyau h, est définie par

$$U = \frac{1}{\binom{n}{m}} \sum_{i:combi} h(X_{i_1}, \ldots, X_{i_m}),$$

avec $\sum_{i:combi}$ la somme sur les $\binom{n}{m}$ combinaisons $(i_1, \ldots, i_m)$ possibles de m éléments distincts de $\{1, \ldots, n\}$.

Exemple II.1. *La « signed rank statistic » a pour noyau $h(x_1, x_2) = \mathbb{1}_{(x_1+x_2>0)}$ et est définie par*

$$U = \frac{1}{\binom{n}{2}} \sum_{i<j}\sum \mathbb{1}_{(X_i+X_j>0)}.$$

Elle est souvent utilisée pour tester la non-symétrie d'une distribution autour de zéro.

Par linéarité de l'espérance, on note que U est un estimateur sans biais de θ. Cependant, comme U est une somme de quantités non indépendantes, sa distribution asymptotique ne peut pas immédiatement être dérivée du théorème central limite. Le théorème suivant est donc très pratique : il permet d'approximer U par une somme de termes i.i.d quand $n \to \infty$, et par conséquent d'en dériver sa distribution asymptotique, via le théorème central limite et le lemme de Slutsky.

Théorème II.2 (Décomposition i.i.d et loi asymptotique d'une U-statistique)**.** *Supposons que $\mathbb{E}h(X_1, \ldots, X_m) < \infty$ et définissons la projection de Hájek $\widehat{U}$ par*

$$\widehat{U} = \sum_{i=1}^{n} \mathbb{E}(U - \theta | X_i).$$

Alors,

1. $\widehat{U} = \frac{m}{n}\sum_{i=1}^{n} h_1(X_i)$*, avec* $h_1(x) = \mathbb{E}\left[h(X_1, X_2, \ldots, X_m) | X_1 = x\right] - \theta$*, aussi souvent noté* $h_1(x) = \mathbb{E}h(x, X_2, \ldots, X_m) - \theta$*.*
2. $\sqrt{n}(U - \theta - \widehat{U}) = o_p(1)$*,*
3. $\sqrt{n}(U - \theta)$ *est asymptotiquement gaussien, de moyenne 0 et de variance* $m^2 Var\left[h_1(X_1)\right]$*.*

Démonstration. Une preuve est donnée par van der Vaart (1998, Section 12.3). □

Exemple II.1 (suite)**.** *Pour la « signed rank statistic », on a*

$$h_1(x) = \mathbb{E}\mathbb{1}_{(x+X_2>0)} - \theta = 1 - F(-x) - \theta,$$

avec $\theta = \mathbb{E}U = 1 - \mathbb{E}F(-X_i)$*, ou* F *désigne la fonction de répartition de* X*. Ainsi,*

$$\widehat{U} = -\frac{2}{n}\sum_{i=1}^{n}\Big\{F(-X_i) - \mathbb{E}F(-X_i)\Big\},$$

et sous l'hypothèse nulle de symétrie autour de zéro, i.e.,

$$\mathcal{H}_0 : \forall x \quad F(x) + F(-x) = 1,$$

alors

$$\theta = \mathbb{E}\Big[F(-X_j)\Big] = \mathbb{E}\Big[1 - F(X_j)\Big] = \mathbb{E}\Big(\mathcal{U}[0,1]\Big) = 1/2$$

et

$$\textit{Var}\Big[F(-X_i)\Big] = \textit{Var}\Big(\mathcal{U}[0,1]\Big) = 1/12.$$

Finalement, sous $\mathcal{H}_0$ *on a alors*

$$\sqrt{n}\,(U - 1/2) \xrightarrow{\mathcal{D}} \mathcal{N}\,(0, 1/3)\,,$$

qui permet de tester l'hypothèse de non-symétrie autour de zéro.

Les termes i.i.d du théorème II.2, dont la somme $\widehat{U}$ approxime U, s'interprètent simplement comme les projections de la statistique centrée $U - \theta$ sur l'espace engendré par l'information de chaque sujet. C'est la projection de Hájek. Comme noté par Serfling (1980), les décompositions i.i.d des U-statistiques peuvent aussi être vues comme des cas particuliers de résultats plus généraux et plus forts, dérivés du « calcul de van Mises », de la « delta-méthode fonctionnelle » et des notions de « fonctions d'influence » (Serfling, 1980, Chap. 6) (van der Vaart, 1998, Chap. 20).

Remarque II.2 (Symétriser un noyau). *Lorsqu'un noyau* $h(x_1, \ldots, x_m)$ *est non symétrique, on peut toujours considérer un noyau symétrique associé, défini par*

$$g(x_1, \ldots, x_m) = \frac{1}{m!}\sum_{i:permut} h(x_{i_1}, \ldots, x_{i_m})$$

où $\sum_{i:permut}$ *désigne la somme sur les* $m!$ *permutations* $(i_1, \ldots, i_m)$ *de* $\{1, \ldots, m\}$*.*

Remarque II.3 (Projection de Hájek et noyau non symétrique). *Soit $h(x_1, \ldots, x_m)$ un noyau non symétrique et une statistique T définie par*

$$T = \frac{1}{(n)_m} \sum_{i:listes} h(x_{i_1}, \ldots, x_{i_m}), \tag{II.19}$$

avec $(n)_m = n(n-1)\ldots(n-m+1)$ et $\sum_{i:listes}$ la sommes sur les $(n)_m$ listes ordonnées d'éléments distincts de $\{1, \ldots, n\}$ (arrangements). Comme $\binom{n}{m} = (n)_m/m!$, on a alors

$$\begin{aligned} T = \frac{1}{(n)_m} \sum_{i:listes} h(x_{i_1}, \ldots, x_{i_m}) &= \frac{1}{\binom{n}{m}} \sum_{j:combi} \left\{ \frac{1}{m!} \sum_{i:permut} h(x_{i_{j_1}}, \ldots, x_{i_{j_m}}) \right\} \\ &= \frac{1}{\binom{n}{m}} \sum_{j:combi} g(x_{j_1}, \ldots, x_{j_m}) \end{aligned}$$

avec la notation $g(x_1, \ldots, x_m) = \frac{1}{m!} \sum_{i:permut} h(x_{i_1}, \ldots, x_{i_m})$. Supposons maintenant que, comme pour le résultat du théorème II.2, on a une décomposition telle que

$$U = \frac{1}{\binom{n}{m}} \sum_{i:combi} g(x_{i_1}, \ldots, x_{i_m}) = \frac{m}{n} \sum_{i=1}^{n} g_1(X_i) + o_p(n^{-1/2})$$

avec $g_1(x) = \mathbb{E}g(x, X_2, \ldots, X_m)$. Par linéarité de l'espérance conditionnelle, et comme pour toutes variables i.i.d $X_1, \ldots, X_m, X'_1, \ldots, X'_m$, on a

$$\begin{array}{ccc} \mathbb{E}\Big(h(X_1, X_2, \ldots, X_m)\Big|X_1\Big) & = & \mathbb{E}\Big(h(X_1, X'_2, \ldots, X'_m)\Big|X_1\Big) \\ \vdots & \vdots & \vdots \\ \mathbb{E}\Big(h(X_1, \ldots, X_{m-1}, X_m)\Big|X_m\Big) & = & \mathbb{E}\Big(h(X'_1, \ldots, X'_{m-1}, X_m)\Big|X_m\Big) \end{array}$$

alors

$$\begin{aligned} g_1(x) &= \mathbb{E}\Big[g(X_1, X_2, \ldots, X_m)\Big|X_1 = x\Big] \\ &= \frac{1}{m!} \sum_{j:permut} \mathbb{E}\Big[h(X_{j_1}, \ldots, X_{j_m})\Big|X_1 = x\Big] = \frac{1}{m!} \times (m-1)! \times \sum_{k=1}^{m} h_k(x), \end{aligned}$$

avec

$$\begin{aligned} h_k(x) &= \mathbb{E}\Big[h(X_1, \ldots, X_{k-1}, X_k, X_{k+1}, \ldots, X_m)\Big|X_k = x\Big] \\ &= \mathbb{E}h(X_1, \ldots, X_{k-1}, x, X_{k+1}, \ldots, X_m). \end{aligned}$$

Et finalement, on obtient :

$$\begin{aligned}
T = \frac{1}{(n)_m} \sum_{i:listes} h(x_{i_1}, \ldots, x_{i_m}) &= \frac{1}{\binom{n}{m}} \sum_{j:combi} g(x_{j_1}, \ldots, x_{j_m}) \\
&= \frac{m}{n} \sum_{i=1}^{n} g_1(X_i) + o_p(n^{-1/2}) \\
&= \frac{m}{n} \sum_{i=1}^{n} \left\{ \left(\frac{1}{m!} \times (m-1)! \right) \sum_{k=1}^{m} h_k(X_i) \right\} + o_p(n^{-1/2}) \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{m} h_k(X_i) + o_p(n^{-1/2}).
\end{aligned}$$

La remarque II.3 nous sera utile dans la suite, car nos estimateurs des Chapitres IV et V seront tous présentés, pour les développements asymptotiques, sous la forme de l'équation (II.19) (à un terme négligeable près), avec un noyau non-symétrique et une somme sur toutes les listes ordonnées, plutôt que sous la forme plus usuelle supposée par le théorème II.2, avec la somme sur toutes les combinaisons et un noyau symétrique.

II.5.3 Représentation martingale de l'estimateur de Kaplan-Meier

On suppose qu'on observe le n-échantillon i.i.d $\{(\widetilde{T}_1, \Delta_1), \ldots, (\widetilde{T}_n, \Delta_n)\}$, avec $\widetilde{T}_i = \min(T_i, C_i)$ et $\Delta_i = \mathbb{1}_{(T_i \leqslant C_i)}$, où C est une censure indépendante. Soit $\tau < \sup\{u : S_{\widetilde{T}}(u) > 0\}$. L'intervalle $[0, \tau]$ représente ainsi un intervalle de temps dans lequel la probabilité d'observer quelqu'un à risque est non nulle. On a alors le résultat suivant pour l'estimateur de Kaplan-Meier précédement défini à l'équation (II.17) page 58.

Proposition II.3 (Représentation martingale de l'estimateur de Kaplan-Meier).

$$\sup_{t \in [0,\tau]} \left| \sqrt{n} \left(\widehat{S}(t) - S(t) \right) + \frac{S(t)}{\sqrt{n}} \sum_{i=1}^{n} \int_0^t \frac{dM_i(u)}{S_{\widetilde{T}}(u)} \right| = o_p(1) \tag{II.20}$$

où $M_i(t) = \mathbb{1}_{(\Delta_i = 1, \widetilde{T}_i \leqslant t)} - \int_0^t \mathbb{1}_{(\widetilde{T}_i \geqslant t)} d\Lambda(u)$ *est la martingale usuelle associée au processus de comptage* $\mathbb{1}_{(\Delta_i = 1, \widetilde{T}_i \leqslant t)}$*, et* $\Lambda(\cdot)$ *est le risque cumulé de la variable aléatoire* T*.*

Démonstration. Voir par exemple Gill (1994, Chapitre 6). □

Plusieurs propriétés de $\widehat{S}(t)$ peuvent être dérivées à partir de cette représentation de l'estimateur de Kaplan-Meier comme une moyenne d'intégrales stochastiques par rapport à des martingales. Notamment, on peut en déduire la normalité asymptotique de $\widehat{S}(t)$.

Combinée à la proposition II.2 et à la remarque II.3, un autre intérêt pratique de la proposition II.3 est qu'elle permet d'étudier le comportement d'estimateurs IPCW (« Inverse Probability of Censoring Weighting ») basés sur l'estimateur $\widehat{S}(t)$. Cette approche a déja été utilisée par Hung et Chiang (2010) et Datta *et al.* (2010), entre autres, et nous la ré-utiliserons aux Chapitres IV et V.

II.5.4 « Conditional multiplier central limit theorem »

Informellement, le résultat est le suivant. Soient $X_1, \ldots, X_n$ des variables (ou des vecteurs) aléatoires réelles i.i.d. et $G_1, \ldots, G_n$ des variables aléatoires i.i.d indépendantes de loi $\mathcal{N}(0,1)$ et supposons qu'il existe une variable aléatoire U telle que

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathcal{D}} U.$$

Alors, sous certaines conditions, le théorème affirme que

$$\frac{1}{n}\sum_{i=1}^{n} G_i X_i \xrightarrow{\mathcal{D}} U.$$

En pratique, l'utilité de ce résultat vient aussi du fait que, si $\widehat{X}_i$ est une « bonne approximation » de X_i, alors on a aussi souvent

$$\frac{1}{n}\sum_{i=1}^{n} G_i \widehat{X}_i \xrightarrow{\mathcal{D}} U.$$

Ce résultat a permis à Lin *et al.* (1994) de proposer des méthodes de simulation pour approximer les distributions de processus complexes, et d'en dériver des régions de confiances simultanées, appelées « bandes de confiances », ainsi que des tests. Depuis, l'approche a été extensivement reprise, notamment par Martinussen et Scheike (2006) qui l'ont étudiée en détails pour plusieurs modèles de survie, et plus récemment par Beyersmann *et al.* (2012), qui l'assimilent d'ailleurs à du « wild bootstrap » (Wu, 1986).

III. Courbe ROC dépendant du temps et censure dépendante de l'outil prédictif à évaluer

III.1 Publication dans le *Biometrical Journal*

Résumé :

Dans ce premier article, on propose une revue de la littérature des estimateurs de la courbe ROC dépendant du temps pour les définitions *Cumulative/dynamique* des cas et des contrôles. On rappelle les hypothèses supposées par chacun des estimateurs, et on discute leurs similitudes et différences, ainsi que leurs principales propriétés. Notamment, on rappelle les hypothèses qu'ils supposent sur le mécanisme de censure. On propose également un estimateur alternatif, qui est l'adaptation des estimateurs IPCW de Uno *et al.* (2007) et de Hung et Chiang (2010) au cas d'une censure qui dépend du marqueur étudié. On compare ensuite les différents estimateurs par simulation, en présence d'une censure indépendante du temps d'événement, ou d'une censure seulement indépendante du temps d'événement conditionnellement au marqueur étudié. Enfin, on utilise les données de la cohorte Paquid pour les comparer sur des données réelles.

Les résultats suggèrent qu'une censure dépendant du marqueur peut biaiser les estimateurs qui n'y sont pas adaptés, et que l'estimateur que nous proposons performe aussi bien que le « nearest neighbor estimator » de Heagerty *et al.* (2000).

Biometrical Journal **00** (2013) 00, 1–18 DOI: 10.1002/bimj.201200045 1

Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring

Paul Blanche*[1,2], **Jean-François Dartigues**[1,2], and **Hélène Jacqmin-Gadda**[1,2]

[1] Université Bordeaux, ISPED, Centre INSERM U897-Epidemiologie-Biostatistique, F-33000 Bordeaux, France
[2] INSERM, ISPED, Centre INSERM U897-Epidemiologie-Biostatistique, F-33000 Bordeaux, France

Received 24 February 2012; revised 18 February 2013; accepted 17 April 2013

To quantify the ability of a marker to predict the onset of a clinical outcome in the future, time-dependent estimators of sensitivity, specificity, and ROC curve have been proposed accounting for censoring of the outcome. In this paper, we review these estimators, recall their assumptions about the censoring mechanism and highlight their relationships and properties. A simulation study shows that marker-dependent censoring can lead to important biases for the ROC estimators not adapted to this case. A slight modification of the inverse probability of censoring weighting estimators proposed by Uno et al. (2007) and Hung and Chiang (2010a) performs as well as the nearest neighbor estimator of Heagerty et al. (2000) in the simulation study and has interesting practical properties. Finally, the estimators were used to evaluate abilities of a marker combining age and a cognitive test to predict dementia in the elderly. Data were obtained from the French PAQUID cohort. The censoring appears clearly marker-dependent leading to appreciable differences between ROC curves estimated with the different methods.

Keywords: AUC; IPCW; Prediction; ROC curve; Survival analysis.

1 Introduction

For many diseases, it is useful to identify a marker or a combination of markers that enables the identification of subjects at high and low risk of the disease in the future. In prostate cancer, the value or the change in prostate-specific antigen level is frequently used to predict cancer recurrence after the initial therapy and then to decide whether or not to undergo a secondary therapy (Proust-Lima and Taylor, 2009). In Alzheimer's disease (AD), the decline in cognitive functions begins a long time before all the criteria for the clinical diagnosis are reached. To ensure safety and care of these declining patients, it would be useful to identify them as early as possible. In particular, AD treatments given after the clinical diagnosis have been shown to have modest effects and research is currently focussing on preventive treatment given in the prediagnosis phase (Vellas et al., 2006). To ensure sufficient power for this preventive trials and then to apply the preventive treatment if its efficacy is demonstrated, validated markers for detecting subjects at high risk of AD in next years are required.

The diagnostic accuracy of a quantitative marker is often evaluated by the ROC curve that displays the sensitivity (probability that the marker X be above the cutpoint c for a diseased subject) versus 1-specificity (where the specificity is the probability that X be below c for a healthy subject) for all the possible cutpoints c. The diagnostic accuracy is often summarized by the area under the ROC curve

*Corresponding author: e-mail: Paul.Blanche@isped.u-bordeaux2.fr, Phone: +33-5-57-57-95-76

(AUC) that may be interpreted as the probability that the marker value of a randomly chosen case is above the marker value of a randomly chosen healthy subject (Pepe, 2003). In a diagnostic study, the marker and disease are measured at the same time and are known for all participants. In prognostic studies, the marker is measured at a given time (considered as time 0 in the following) while the disease may occur at any time thereafter. Thus, sensitivity, specificity, and ROC curve are time-dependent and may be computed for different time durations t (window of prediction).

Heagerty and Zheng (2005) proposed several definitions of the ROC curve for survival time that are also discussed in Pepe et al. (2008) and Cai et al. (2006). In this paper, we focus on the *cumulative/dynamic* definition (Heagerty and Zheng, 2005) where a case is a subject diagnosed before time t and a healthy subject (a control) is a subject free of the disease at time t. From our point of view, this definition is the most relevant as clinicians often want to predict disease onset in a period of time rather than at a specific time t (as in *incident* sensitivity) and want to distinguish healthy subjects at the end of the same period rather than at a later prespecified time τ (as in *static* specificity).

Without censoring, one could easily estimate such quantities by empirical proportions. Sensitivity could be estimated by the observed true-positive fraction and specificity could be estimated by the true-negative fraction (Pepe, 2003). However, in practice, there is often loss to follow-up before the time point t. Therefore, for some subjects, it is impossible to know if the outcome occurred before the time point t. To deal with this censoring, several approaches have been proposed to estimate the cumulative sensitivity, the dynamic specificity, and the associated ROC curve. First, Heagerty et al. (2000) proposed estimators based on Bayes' theorem and the Kaplan–Meier estimator (denoted by $\mathrm{KM}_{\mathrm{HLP}}$ in the following). As this approach does not guarantee the monotonicity of estimators, they also proposed estimators based on the nearest neighbor estimator of the bivariate distribution of the marker and the time-to-event (denoted by NNE). Then, Chambless and Diao (2006) proposed two alternative methods. The first one deals with censoring by conditioning on observed event times as in the Kaplan–Meier estimator (denoted by $\mathrm{KM}_{\mathrm{CD}}$). The second method, studied in detail by Song and Zhou (2008), uses a model for $S(t|X)$, the conditional survival probability of the outcome at time t given the marker X. Finally, Uno et al. (2007) and Hung and Chiang (2010b) have independently proposed an inverse probability of censoring weighting method (denoted by IPCW).

Among these estimators, only the NNE (Heagerty et al., 2000) and the model-based one (Chambless and Diao, 2006) allow the censoring to depend on the marker, whereas this is often the case in epidemiology. For instance, in cohorts of elderly people, it is known that poor cognitive level at entry is associated with both AD risk and study dropout (Jacqmin-Gadda et al., 1997). More generally, the marginal independence assumption between censoring time and event time required by most of these estimators is much less tenable than the conditional independence assumption given the marker. However, the weighting approach of the IPCW estimator may be slightly modified to obtain a nonparametric estimator robust to marker-dependent censoring. The resulting estimator is denoted CIPCW in the following (for conditional IPCW). On the other hand, the model-based approach has two drawbacks: it does not preserve invariance to increasing transformation of the marker and can lead to biases when the model is misspecified. To our knowledge, all of these estimators have never been compared together; the most exhaustive comparisons by simulation were performed under the independence assumption (Viallon and Latouche, 2011) or under weak dependence structure (Chiang and Hung, 2010).

The goal of this paper is to review the estimators proposed in the literature, to point out their similarities and differences and to compare their behaviors when the censoring time depends on the marker.

The "Naive" estimator, the estimators assuming independent censoring and the estimators robust to marker-dependent censoring are reviewed in Sections 2, 3, and 4, respectively, and then compared for different censoring scenarios in a simulation study in Section 5. Section 6 presents an application to the PAQUID cohort of elderly people (Jacqmin-Gadda et al., 1997; Amieva et al., 2005) to evaluate a predictive marker for dementia based on a cognitive test. Section 7 discusses limitations and possible extensions.

2 Naive estimator of time-dependent ROC curve

2.1 Notations and definitions

Let X_i denote a quantitative marker, T_i a time-to-event, C_i a censoring time, $\delta_i = I(T_i \leq C_i)$ the indicator of event, and $T_i^* = \min(T_i, C_i)$ the observed time for a subject i. We observe a sample of n independent subjects i: $\{(T_i^*, \delta_i, X_i), i = 1, \ldots, n\}$. To simplify the presentation of the estimators, we assume there are no ties in either the observed times $\{T_i^*, i = 1, \ldots, n\}$ or in the quantitative marker values $\{X_i, i = 1, \ldots, n\}$. We discuss adaptation of estimators for ties at the end of Section 4. Hereafter, we denote $S(t)$ the survival probability $\mathrm{P}(T > t)$ and $S(t|X)$ the conditional survival probability $\mathrm{P}(T > t|X)$. $I(\cdot)$ denotes the indicator function.

For a threshold $c \in \mathbb{R}$ and a given time t, Heagerty and Zheng (2005) defined the *cumulative* sensitivity $Se(c, t)$ and the *dynamic* specificity $Sp(c, t)$ by

$$Se(c, t) = \mathrm{P}(X > c|T \leq t) \quad \text{and} \quad Sp(c, t) = \mathrm{P}(X \leq c|T > t).$$

The corresponding time-dependent ROC curve, denoted by $ROC(t)$, is defined as the plot of $Se(c, t)$ versus $1 - Sp(c, t)$ for all possible values of c, i.e.,

$$ROC(t) = \{(1 - Sp(c, t), Se(c, t)), c \in \mathbb{R}\}.$$

The area under the $ROC(t)$ curve, denoted by $AUC(t)$, is therefore equal to $\mathrm{P}(X_i > X_j|T_i \leq t, T_j > t)$ with i and j the indexes of two independent subjects. In the following, given any estimators of sensitivity and specificity $\widehat{Se}$ and $\widehat{Sp}$, we denote $\widehat{ROC}(t) = \{(1 - \widehat{Sp}(c, t), \widehat{Se}(c, t)), c \in \mathbb{R}\}$, and $\widehat{AUC}(t)$ the area under the $\widehat{ROC}(t)$ curve.

2.2 "Naive" estimator

When there is no censoring, $Se(c, t)$ can be estimated as the proportion of subjects with $X_i > c$ among subjects diagnosed before t and $Sp(c, t)$ can be estimated as the proportion of subjects with $X_i \leq c$ among subjects free of disease at time t.

With censored data, the "Naive" estimator is computed by removing all subjects censored before time point t. Thus, for each subject kept in this subsample, we know if the event occurred before time t or not. "Naive" estimators of sensitivity and specificity can then be defined by observed true-positive and true-negative fractions in this subsample,

$$\widehat{Se}(c, t) = \frac{\sum_{i=1}^{n} \delta_i I(X_i > c, T_i^* \leq t)}{\sum_{i=1}^{n} \delta_i I(T_i^* \leq t)}, \qquad \widehat{Sp}(c, t) = \frac{\sum_{i=1}^{n} I(X_i \leq c, T_i^* > t)}{\sum_{i=1}^{n} I(T_i^* > t)}. \tag{1}$$

As a consequence, the area under the resulting step $\widehat{ROC}(t)$ curve is equal to

$$\widehat{AUC}(t) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i I(T_i^* \leq t, T_j^* > t) I(X_i > X_j)}{\sum_{i=1}^{n} \delta_i I(T_i^* \leq t) \sum_{j=1}^{n} I(T_j^* > t)}.$$

These estimators induce a loss of information and are often biased depending on the censoring distribution. We show in the Appendix that as $n \to \infty$, we obtain:

$$\widehat{Se}(c,t) \xrightarrow{a.s} Se(c,t) \times \frac{\mathrm{P}\big(T \le C | X > c, T \le t\big)}{\mathrm{P}\big(T \le C | T \le t\big)},$$

$$\widehat{Sp}(c,t) \xrightarrow{a.s} Sp(c,t) \times \frac{\mathrm{P}\big(C > t | X \le c, T > t\big)}{\mathrm{P}\big(C > t | T > t\big)},$$

$$\text{and} \quad \widehat{AUC}(t) \xrightarrow{a.s} AUC(t) \times \frac{\mathrm{P}(T_i \le C_i, C_j > t | T_i \le t, T_j > t, X_i > X_j)}{\mathrm{P}(T_i \le C_i, C_j > t | T_i \le t, T_j > t)},$$

with i and j the indexes of two independent subjects. These results show that the "Naive" specificity estimator is consistent if censoring is independent of X and T while the "Naive" sensitivity and AUC estimators may be biased in this case since T depends on X. For instance, with independent uniform censoring over $[0, t]$, $\mathrm{P}(T \le C | X > c, T \le t) \neq \mathrm{P}(T \le C | T \le t)$ for some c as soon as T depends on X. However, we will see in Section 5 that this bias can be small in practice due to conditioning on $T \le t$.

3 Estimators assuming independent censoring

These estimators were proposed to account for the information brought by censored subjects with unknown status at time t.

3.1 Kaplan–Meier estimator of Heagerty et al. (2000) ($\mathbf{KM_{HLP}}$)

To deal with censored data, Heagerty et al. (2000) first proposed to use Bayes' theorem to rewrite sensitivity and specificity as functions of easy computable terms in presence of censoring. They proposed the estimators

$$\widehat{Se}(c,t) = \frac{[1 - \hat{S}(t|X > c)](1 - \hat{F}_X(c))}{1 - \hat{S}(t)} \quad \text{and} \quad \widehat{Sp}(c,t) = \frac{\hat{S}(t|X \le c)\hat{F}_X(c)}{\hat{S}(t)} \tag{2}$$

with $\hat{F}_X(\cdot)$ the empirical distribution function of the marker X, $\hat{S}(\cdot)$ the Kaplan–Meier estimator of the marginal survival function of the time-to-event T (Kaplan and Meier, 1958), and $\hat{S}(\cdot|X > c)$ and $\hat{S}(\cdot|X \le c)$ the Kaplan–Meier estimators computed on the subset of subjects with marker values such as $X > c$ and $X \le c$, respectively.

As the Kaplan–Meier estimator requires independence between the censoring time and the event time, these estimators are not robust to marker-dependent censoring. Another problem is that $\widehat{Se}(c,t)$ and $\widehat{Sp}(c,t)$ are not necessarily monotone in c nor bounded in $[0, 1]$. Therefore, the corresponding $\widehat{ROC}(t)$ curve is neither monotone nor necessarily included in the square $[0, 1] \times [0, 1]$. This is due to the fact that the conditional survival functions are estimated on different subsamples when c varies.

Heagerty et al. (2000) proposed a Bootstrap approach for computing the variances and the confidence intervals of these estimates.

3.2 Kaplan–Meier-like estimator of Chambless and Diao (2006) ($\mathbf{KM_{CD}}$)

Chambless and Diao (2006) proposed a Kaplan–Meier-like estimator that consists in a recursive computation using the risk sets at each event time. Let us consider the ordered observed event times

$s_0 = 0 < s_1 < s_2 < \cdots < s_{m(t)}$, with s_k the k-th observed event time and $s_{m(t)}$ the last observed event time before time point t. Chambless and Diao (2006) proposed the following estimators, which we present with a slightly different formulation in order to make the comparisons with the other estimators easier:

$$\widehat{Se}(c,t) = \frac{\sum_{k=1}^{m(t)} I(X_{d(k)} > c)(\hat{S}(s_{k-1}) - \hat{S}(s_k))}{1 - \hat{S}(s_{m(t)})}$$

$$\text{and} \quad \widehat{Sp}(c,t) = \frac{\hat{F}_X(c) - \sum_{k=1}^{m(t)} I(X_{d(k)} \le c)(\hat{S}(s_{k-1}) - \hat{S}(s_k))}{\hat{S}(s_{m(t)})}, \tag{3}$$

where $d(k)$ is the index of the subject who experiences the event at time t_k. Here again, $\hat{S}(\cdot)$ denotes the Kaplan–Meier estimator of the survival function of the time-to-event T. The indicator $I(X_{d(k)} > c)$ estimates $\mathrm{P}(X > c | s_{k-1} < T \le s_k)$, the indicator $I(X_{d(k)} \le c)$ estimates $\mathrm{P}(X \le c | s_{k-1} < T \le s_k)$, and the difference $\hat{S}(s_{k-1}) - \hat{S}(s_k)$ estimates $\mathrm{P}(s_{k-1} < T \le s_k)$.

By contrast to $\mathrm{KM}_{\mathrm{HLP}}$, this sensitivity estimator is monotone from 0 to 1. However, the specificity estimator is not monotone and is not necessarily bounded in [0, 1]. Therefore, the corresponding $\widehat{ROC}(t)$ curve is neither monotone nor necessarily included in the square $[0, 1] \times [0, 1]$. Indeed, if we order subjects according to X, the change of specificity between two thresholds corresponding to two successive observed values $X_{(i)} < X_{(i+1)}$ of the marker X is negative when $\delta_{(i+1)} = 1$ and $T^*_{(i+1)} < t$. This is due to the fact that the change of the Kaplan–Meier estimator between two successive observed times s_{k-1} and s_k is always greater than or equal to $1/n$.

Chambless and Diao (2006) suggested to use bootstrapping for computing variances of these estimators and their confidence intervals.

3.3 Inverse probability of censoring weighting

Uno et al. (2007) and Hung and Chiang (2010b) separately proposed to correct the "Naive" estimator by weighting the observations kept in the subsample of uncensored subjects before time t by their probability of being kept in the subsample, that is their probability of being uncensored; i.e., they proposed:

$$\widehat{Se}(c,t) = \frac{\sum_{i=1}^{n} I(T_i^* \le t, X_i > c) \frac{\delta_i}{n\hat{S}_C(T_i^*)}}{\sum_{i=1}^{n} I(T_i^* \le t) \frac{\delta_i}{n\hat{S}_C(T_i^*)}} \quad \text{and} \quad \widehat{Sp}(c,t) = \frac{\sum_{i=1}^{n} I(T_i^* > t, X_i \le c)}{\sum_{i=1}^{n} I(T_i^* > t)}, \tag{4}$$

where $\hat{S}_C(\cdot)$ is the Kaplan–Meier estimator of the survival function of the censoring time C. $\hat{S}_C(T_i^*)$ estimates the probability of being uncensored at the observed time T_i^*.

This specificity estimator is the same as the "Naive" one because weights are all equal to $1/(n\hat{S}_C(t))$, which allows simplification of the formula. Although not mentioned by the authors, this sensitivity estimator is identical to the $\mathrm{KM}_{\mathrm{CD}}$ estimator given at formula 3. Indeed, Satten and Datta (2001) showed that, if we order subjects according to T^*, the change of the Kaplan–Meier estimator between two successive observed times $T^*_{(i-1)}$ and $T^*_{(i)}$ is $\hat{S}(T^*_{(i-1)}) - \hat{S}(T^*_{(i)}) = \delta_{(i)}/(n\hat{S}_C(T^*_{(i)}))$ and, as a consequence, we also have $\sum_{i=1}^{n} I(T_i^* \le t) \frac{\delta_i}{n\hat{S}_C(T_i^*)} = 1 - \hat{S}(t)$.

Interestingly, IPCW sensitivity and specificity estimators and the corresponding $\widehat{ROC}(t)$ curve are monotone and bounded in [0, 1]. Given that $\frac{1}{n}\sum_{i=1}^{n} I(T_i^* > t) = \hat{S}_C(t)\hat{S}(t)$, some simple algebra lead to the following formula for the area under the resulting step $\widehat{ROC}(t)$ curve (Hung and Chiang, 2010a):

$$\widehat{AUC}(t) = \frac{\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n} I(T_i^* \le t, T_j^* > t) I(X_i > X_j) \frac{\delta_i}{n^2 \hat{S}_C(T_i^*)\hat{S}_C(t)}}{\hat{S}(t)(1-\hat{S}(t))}.$$

The usual $\sqrt{n}$-consistency and asymptotic normality of these estimators have been established by Uno et al. (2007) and Hung and Chiang (2010a, b), and resampling techniques such as bootstrapping can be used to estimate the variances of the estimators.

4 Estimators for marker-dependent censoring

Using the Kaplan–Meier estimator, previous approaches assume independence between censoring time and time-to-event. However, in epidemiology, censoring often depends on the marker. Thus, time-to-event and censoring cannot be assumed independent. They are more likely independent conditionally on the marker. The three following approaches allow censoring to depend on the marker and only assume the conditional independence assumption between censoring time and time-to-event given the marker.

4.1 Model-based approach

Chambless and Diao (2006) and Song and Zhou (2008) proposed to use a model-based estimator for the conditional survival probability $S(t|X)$. Modeling the probability of each subject i to be a case by $1 - S(t|X_i)$ or a control by $S(t|X_i)$ to deal with censoring, they proposed the estimators:

$$\widehat{Se}(c,t) = \frac{\displaystyle\sum_{i=1}^{n}(1-\hat{S}(t|X_i))I(X_i > c)}{\displaystyle\sum_{i=1}^{n} 1-\hat{S}(t|X_i)} \quad \text{and} \quad \widehat{Sp}(c,t) = \frac{\displaystyle\sum_{i=1}^{n}\hat{S}(t|X_i)I(X_i \le c)}{\displaystyle\sum_{i=1}^{n}\hat{S}(t|X_i)}. \tag{5}$$

Consequently, the area under the resulting $\widehat{ROC}(t)$ curve is

$$\widehat{AUC}(t) = \frac{\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{S}(t|X_j)(1-\hat{S}(t|X_i))I(X_i > X_j)}{\displaystyle\sum_{j=1}^{n}\sum_{i=1}^{n}\hat{S}(t|X_i)(1-\hat{S}(t|X_j))}.$$

Chambless and Diao (2006) suggested a Cox proportional hazards model $\hat{S}(t|X) = \hat{S}_0(t)^{\exp\hat{\beta}X}$; Song and Zhou (2008) have studied asymptotic properties in this case and have extended these estimators to include covariate adjustment. Usual $\sqrt{n}$-consistency and asymptotic normality have been established and resampling techniques such as Bootstrap can be used to estimate the variances. Recently, Foucher et al. (2010) proposed a similar model-based approach to deal with competing risks. This method is

appealing because sensitivity and specificity are monotone, bounded in [0, 1], and allow censoring to depend on the marker. Moreover, some simulations have shown that this approach is more efficient than the nonparametric ones (Chambless and Diao, 2006; Song and Zhou, 2008). However, this model-based approach can easily lead to biases if the survival model is misspecified as it was shown in a simulation study by Viallon and Latouche (2011). Moreover, this estimator does not preserve the invariance to an increasing transformation of the marker X, which is a desirable property of ROC curve estimator (Pepe, 2003, p. 125). For these reasons, we only focused on nonparametric approaches and chose not to compare this method to the others in our simulation study in Section 5.

4.2 NNE of Heagerty et al. (2000)

Heagerty et al. (2000) suggested to use the NNE of the bivariate distribution of (X, T) introduced by Akritas (1994). Although the presentation of their estimators is sligthly different in the original paper, we can define their estimators by formula 5 but replacing the model-based estimator of $S(t|X)$ by the nonparametric estimator proposed by Akritas (1994):

$$\hat{S}(t|X_i) = \prod_{j:T_j^* \leq t} \left(1 - \frac{W_{\lambda_n}(X_i, X_j)}{\sum_l W_{\lambda_n}(X_i, X_l) I(T_l^* \geq T_j^*)}\right)^{\delta_j} \tag{6}$$

with $W_{\lambda_n}(X_i, X_j) = I(|\hat{F}_X(X_i) - \hat{F}_X(X_j)| < \lambda_n)$, and $\lambda_n \in (0, 1]$ a bandwidth. Thus, $\hat{S}(t|X_i)$ is the standard Kaplan–Meier estimator computed with the subset of the $2\lambda_n$ percent of subjects who have the nearest values of the marker X_i (except at the tails). This method was recently extended to the competing risks setting (Saha and Heagerty, 2010). One important feature of the NNE $\hat{S}(t|X_i)$ of Akritas (1994) compared to the more standard Beran (1981) estimator is that the resulting $\widehat{ROC}(t)$ curve is invariant to any monotone transformation of the marker. With this method, sensitivity and specificity estimators are monotone and bounded in [0, 1]. Moreover, this method allows the censoring to depend on the marker X and is nonparametric. Based on results of Akritas (1994), asymptotic properties of sensitivity, specificity, AUC, and partial AUC estimators have been studied by Cai et al. (2011), Hung and Chiang (2011), and Hung and Chiang (2010b). Usual $\sqrt{n}$-consistency and asymptotic normality have been established and resampling techniques such as Bootstrap can be used to estimate the variances. However, to our knowledge, no optimal rule has been proposed to choose the value of λ_n in practice, although results on real data set with finite sample size are sensitive to this choice. The theory only tells us that choosing $\lambda_n = \mathcal{O}(n^{-1/3})$ works when n is large enough (Heagerty et al., 2000).

Finally, let us note that due to smoothing, when applied to uncensored data, the NNE ROC curve estimator remains different from the usual empirical estimator based on empirical true- and false-positive fractions. By contrast, when applied to uncensored data, $\mathrm{KM}_{\mathrm{HLP}}$, $\mathrm{KM}_{\mathrm{CD}}$, and IPCW, all lead to the usual empirical estimator.

4.3 Conditional inverse probability of censoring weighting

Following the idea used by Gerds and Schumacher (2006) to extend the Brier score estimator of Graf et al. (1999), the IPCW estimator may be modified to be robust to marker-dependent censoring. The change consists in weighting the observations of uncensored subjects at time t by the conditional probability of being uncensored given the marker, instead of weighting by the marginal probability of

being uncensored. We suggest the following estimators, denoted CIPCW:

$$\widehat{Se}(c,t) = \frac{\sum_{i=1}^{n} I(X_i > c, T_i^* \le t)\dfrac{\delta_i}{n\widehat{S}_C(T_i^*|X_i)}}{\sum_{i=1}^{n} I(T_i^* \le t)\dfrac{\delta_i}{n\widehat{S}_C(T_i^*|X_i)}},$$

$$\text{and} \quad \widehat{Sp}(c,t) = \frac{\sum_{i=1}^{n} I(X_i \le c, T_i^* > t)\dfrac{1}{n\widehat{S}_C(t|X_i)}}{\sum_{i=1}^{n} I(T_i^* > t)\dfrac{1}{n\widehat{S}_C(t|X_i)}}. \tag{7}$$

The censoring survival probability $S_C(t|X) = \mathrm{P}(C > t|X)$ may be estimated using a Cox model or any other model. However, to avoid any parametric assumption, we propose to use the nonparametric estimator of Akritas (1994). Therefore, $\widehat{S}_C(t|X_i)$ corresponds to the right term of formula (6) where δ_j is replaced by $1 - \delta_j$. Let us remark that, by definition, $\widehat{S}_C(T_i^*|X_i)$ cannot be equal to zero when $\delta_i = 1$ and $\widehat{S}_C(t|X_i)$ cannot be equal to zero when $T_i^* > t$. Then, with the convention $0/0 = 0$, the estimators are always correctly defined.

The area under the resulting step $\widehat{ROC}(t)$ curve is

$$\widehat{AUC}(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I(X_i > X_j) I(T_i^* \le t, T_j^* > t)\dfrac{\delta_i}{n^2\widehat{S}_C(T_i^*|X_i)\widehat{S}_C(t|X_j)}}{\left(\sum_{i=1}^{n} I(T_i^* \le t)\dfrac{\delta_i}{n\widehat{S}_C(T_i^*|X_i)}\right)\left(\sum_{j=1}^{n} I(T_j^* > t)\dfrac{1}{n\widehat{S}_C(t|X_j)}\right)}$$

and corresponds to an estimator already suggested by Hung and Chiang (2010a).

When $\lambda_n < 1/n$, the CIPCW estimators are equal to the "Naive" ones; when $\lambda_n = 1$, for all subject i and all time t, $\widehat{S}_C(t|X_i)$ is equal to the marginal Kaplan–Meier estimator $\widehat{S}_C(t)$, and then CIPCW estimators are equal to the IPCWs. Thus, whatever the value of λ_n, CIPCW has a sensible interpretation. By contrast, choosing $\lambda_n = 1$ yields NNE estimators to $\widehat{Se}(c,t) = 1 - \widehat{Sp}(c,t)$ and so the $\widehat{ROC}(t)$ curve is equal to the first bisector, whereas choosing $\lambda_n < 1/n$ leads to $\widehat{S}(t|X_i) = 1 - I(T_i^* < t)\delta_i$ and so the NNE sensitivity estimator is equal to the "Naive" one but NNE specificity has no meaning in this case.

The CIPCW estimators are nonparametric, monotone, bounded in [0, 1], and robust to marker-dependent censoring. As for NNE, when using the Akritas' conditional Kaplan–Meier estimator, the $\widehat{ROC}(t)$ curve is invariant to any monotone increasing transformation of the marker X, and the bandwidth λ_n does not depend on the scale of X.

By contrast to NNE, when applied to uncensored data, CIPCW estimators are equal to usual empirical estimators, since $\widehat{S}_C(t|X_i)$ is equal to 1 for all time t and all subjects i in this case.

4.4 Adaptation of estimators with ties

Although the marker is assumed quantitative, ties among marker values are frequent in real data sets. In this case, $I(X_i > X_j)$ has to be replaced by $I(X_i > X_j) + \frac{1}{2}I(X_i = X_j)$ in all formulae for $\widehat{AUC}(t)$.

Some issues regarding the interpretation and the validity of ROC curves for ordinal data are discussed in Section 4.5 of Pepe (2003).

If there are ties among the observed times $\{T_i^*, i = 1, \ldots, n\}$, we need to adapt the definition of the conditional Kaplan–Meier estimator, and formula (6) becomes

$$\widehat{S}(t|X_i) = \prod_{s \in \mathcal{T}_n, s \leq t} \left(1 - \frac{\sum_{j=1}^{n} W_{\lambda_n}(X_i, X_j) I(T_j^* = s)\delta_j}{\sum_l W_{\lambda_n}(X_i, X_l) I(T_l^* \geq s)} \right),$$

with $\mathcal{T}_n$ the set of unique values of T_i^*. Moreover, in $\widehat{Se}(t, c)$ and $\widehat{Sp}(t, c)$ in Subsection 3.2, we need to replace $I(X_{d(k)} > c)$ by $\frac{\sum_{i=1}^{n} \delta_i I(T_i^* = t_k) I(X_i > c)}{\sum_{i=1}^{n} \delta_i I(T_i^* = t_k)}$ and $I(X_{d(k)} \leq c)$ by $\frac{\sum_{i=1}^{n} \delta_i I(T_i^* = t_k) I(X_i \leq c)}{\sum_{i=1}^{n} \delta_i I(T_i^* = t_k)}$.

5 Simulation study

The aim of the simulation study was to compare the behavior of the nonparametric estimators of the area under the $ROC(t)$ curve with both independent and marker-dependent censoring.

5.1 Simulation scenarios

Several scenarios were generated. For each one, we generated $N = 500$ samples $\{(T_i^*, X_i, \delta_i), i = 1, \ldots, n\}$ including $n = 300$ subjects. The marker X was generated from the standard normal distribution $\mathcal{N}(0, 1)$. The time-to-event T and the censoring time C were generated from proportional hazards models with a Weibull baseline hazard function: $\lambda_T(t; X) = \frac{\beta t^{\beta-1}}{\eta^\beta} \exp(\alpha X)$ and $\lambda_C(t; X) = \frac{\nu t^{\nu-1}}{\theta^\nu} \exp(\gamma X)$.

In each scenario, we set the proportion of censored data before time point $t = 1$ to be $\mathrm{P}(T_i^* \leq 1, \delta_i = 0) = 50\%$. We simulated 12 scenarios corresponding to two different true AUCs at time point $t = 1$ ($AUC(1) = 0.75$ and $AUC(1) = 0.85$, generated from the corresponding hazard ratios for one standard deviation increase of the marker $\exp(\alpha) \approx 2.30$ and $\exp(\alpha) \approx 4.15$, respectively), two different risks of event ($\mathrm{P}(T > 1) \approx 0.55$ or $\mathrm{P}(T > 1) \approx 0.73$), and three different hazard ratios for one standard deviation increase of the marker $HR_C = \exp(\gamma)$, for the association between the censoring time C and the marker X: independent censoring ($HR_C = 1$), moderate dependence ($HR_C = 1.35$), and strong dependence ($HR_C = 2.40$). The risk of censoring and the risk of event both increased with time in all scenarios ($\nu = 2$ and $\beta > 1$). Values of parameters η, β, and θ were chosen to control the previous values.

The proportion of censoring before $t = 1$ was equal to 50%, and the value $\mathrm{P}(T > 1) \approx 0.73$ was chosen to mimic the real data set of the PAQUID cohort at time $t = 10$ years. The hazard ratios HR_C have the same magnitude than the hazard ratio fitted on the PAQUID data presented in Section 6.

5.2 Results

For each scenario, we computed the bias and the root mean-squared error (RMSE) of the different estimators of $AUC(t)$ at time point $t = 1$. For NNE and CIPCW, we present results for $\lambda_n = 2.5\%$, 5%, and 10%. The true AUC was estimated as the mean over the 500 data sets of the AUC estimates computed by the usual empirical estimate with uncensored data $\{(T_i, X_i), i = 1, \ldots, n\}$. The biases

Table 1 Results of the simulation study : Average bias ($\widehat{AUC}(1) - AUC(1)$) multiplied by 100. Results from NNE and CIPCW estimates are given for $\lambda_n = 2.5\%$, 5%, and 10%.

						NNE			CIPCW		
AUC(1)	P($T > 1$)	"Naive"	KM_{HLP}	KM_{CD}	IPCW	2.5%	5%	10%	2.5%	5%	10%
$HR_C = 1$											
0.75	0.73	0.50	0.03	0.04	−0.06	−0.21	−0.43	−1.21	0.01	−0.01	−0.04
	0.53	0.94	0.02	0.10	−0.07	−0.32	−0.34	−0.99	0.34	0.12	−0.01
0.85	0.72	0.83	0.02	0.04	−0.03	−0.24	−0.38	−1.24	0.15	0.04	−0.03
	0.55	1.70	0.02	0.06	0.04	−0.70	−0.47	−0.97	0.58	0.23	0.04
$HR_C = 1.35$											
0.75	0.73	3.02	2.59	−4.26	2.20	−0.48	−0.61	−1.44	0.26	0.04	−0.07
	0.53	3.05	3.15	−7.96	1.81	−1.00	−0.60	−1.12	0.79	0.35	0.12
0.85	0.72	2.74	4.05	−3.39	1.84	−0.47	−0.48	−1.35	0.44	0.23	0.02
	0.55	3.53	3.90	−5.24	1.98	−1.25	−0.74	−1.12	1.01	0.49	0.33
$HR_C = 2.40$											
0.75	0.73	5.24	5.14	−13.82	3.92	−3.31	−2.44	−2.72	1.49	0.59	−0.06
	0.53	3.40	5.23	−23.11	1.85	−6.13	−3.87	−3.13	1.41	0.80	0.06
0.85	0.72	4.31	9.76	−11.67	3.21	−2.79	−2.00	−2.53	1.17	0.63	0.05
	0.55	4.52	8.27	−15.37	2.90	−4.11	−2.80	−2.67	1.60	1.02	0.66

and RMSE computed with reference to this ideal AUC estimate are displayed in Tables 1 and 2, respectively.

When censoring is independent of the marker ($HR_C = 1$), all estimators are consistent with similar RMSE. It is interesting to note that the bias for the "Naive" estimator is very small in these scenarios. As expected, simulations show that "Naive," KM_{HLP}, KM_{CD}, and IPCW are biased when censoring depends on the marker, and bias and RMSE increase with an increasing association between the marker and the censoring time. Whereas "Naive," KM_{HLP}, and IPCW overestimate the AUC, KM_{CD} underestimates it in these simulations. Even if IPCW assumes independent censoring like KM_{HLP} and KM_{CD}, it appears to be more robust to dependent censoring than KM_{HLP} and KM_{CD}. Interestingly, one can note that KM_{HLP} and KM_{CD} may be more biased than the "Naive" estimator for strongly dependent censoring.

NNE and CIPCW are both suitable when censoring depends on the marker. With respect to both bias and RMSE, these estimators perform as well when censoring is independent ($HR_C = 1$) or moderately dependent on the marker ($HR_C = 1.35$). However, when censoring strongly depends on the marker ($HR_C = 2.4$), NNE tends to be more biased than CIPCW. In these scenarios with $HR_C = 2.4$, there are neighborhoods of some X_i that include only subjects that either met the event before time t or are censored before time t. For these neighborhoods, the conditional Kaplan–Meier estimator of the time-to-event survival function $S(t|X_i)$ performs poorly. This is probably the main reason of the biases in NNE with highly dependent censoring, and this could explain why the bias increases when survival probability of time-to-event is lower (≈ 0.53). This also explains that NNE behavior depends more strongly on the size of the neighborhood λ_n when $HR_C = 2.4$. By contrast, CIPCW only requires the computation of the censoring survival function $S_C(t|X_i)$ for X_i such that $T_i^* \geq t$. Thus, there is always at least one subject at risk in each of these neighborhoods. Finally, the behavior of NNE appears more dependent on the choice of the bandwidth λ_n than the behavior of CIPCW. This is confirmed using other bandwidths $\lambda_n = 7.5\%$, 12.5%, and 15% (results not shown).

Table 2 Results of the simulation study: Root mean-squared error of $\widehat{AUC}(1)$ multiplied by 100. Results from NNE and CIPCW estimates are given for $\lambda_n = 2.5\%$, 5%, and 10%.

						NNE			CIPCW		
AUC(1)	P($T > 1$)	"Naive"	KM_{HLP}	KM_{CD}	IPCW	2.5%	5%	10%	2.5%	5%	10%
$HR_C = 1$											
0.75	0.73	4.12	4.27	4.27	4.25	4.20	4.18	4.31	4.27	4.13	4.08
	0.53	4.09	4.25	4.69	4.19	4.00	4.02	4.07	4.24	4.03	3.93
0.85	0.72	3.10	3.42	3.38	3.24	3.08	3.08	3.34	3.16	3.03	3.00
	0.55	3.34	3.44	3.57	3.31	3.03	3.02	3.16	3.27	3.09	3.01
$HR_C = 1.35$											
0.75	0.73	4.88	5.60	6.13	4.59	4.33	4.29	4.45	4.37	4.23	4.15
	0.53	4.82	5.85	9.26	4.35	4.29	4.14	4.27	4.32	4.21	4.11
0.85	0.72	3.91	5.78	4.93	3.57	3.15	3.17	3.48	3.16	3.10	3.13
	0.55	4.32	5.49	6.33	3.48	3.17	3.04	3.23	3.22	3.06	3.06
$HR_C = 2.40$											
0.75	0.73	6.57	9.02	14.62	5.76	6.22	6.01	6.06	5.30	5.46	5.54
	0.53	4.95	8.86	23.60	4.24	8.21	6.92	6.59	5.03	5.22	5.45
0.85	0.72	5.22	12.03	12.41	4.62	4.89	4.63	5.00	3.90	4.00	4.05
	0.55	5.08	9.76	15.86	3.95	5.41	4.51	4.47	3.49	3.33	3.42

6 Illustration on prediction of dementia

6.1 Objective

The objective of this analysis was to evaluate the predictive performance of a cognitive marker to predict dementia onset over a period of time of 5, 10, 15, or 17 years. The marker was previously defined as a linear combination of age at baseline (the main risk factor for dementia) and Digit Symbol Substitution Test (DSST) score (Wechsler, 1981) at baseline using a Cox model. The DSST explores attention and psychomotor speed and is highly associated with cognitive ageing and risk of dementia. It consists in filling in, in a 90-s interval, blank squares with the symbol that is paired to the digit displayed above the square, according to a code table of associated digits and symbols. Such a combined marker could be useful for preventive clinical trials to select population at high risk of dementia over the duration of the trial.

6.2 The PAQUID sample

The PAQUID cohort is a French prospective study on cognitive ageing including 3777 subjects aged 65 years and older and living at home at baseline. Subjects were initially interviewed at home in 1988 and 1, 3, 5, 8, 10, 13, 15, 17, and 20 years later. Their cognition was evaluated at each visit with several psychometric tests such as the DSST. In addition, dementia diagnosis was assessed at each visit by the investigating psychologist and then confirmed through a clinical examination by a neurologist if the subjects were screened positive by the psychologist.

The sample for this analysis included 2516 subjects nondemented at the initial visit, without missing value for the DSST at baseline and who were visited at least once thereafter. The baseline time for the analysis is the follow-up time from the baseline visit. Time to dementia onset is computed as the mean between the time at the visit of diagnosis and the time at the previous visit. Censoring time is the time at the last follow-up visit for the subject.

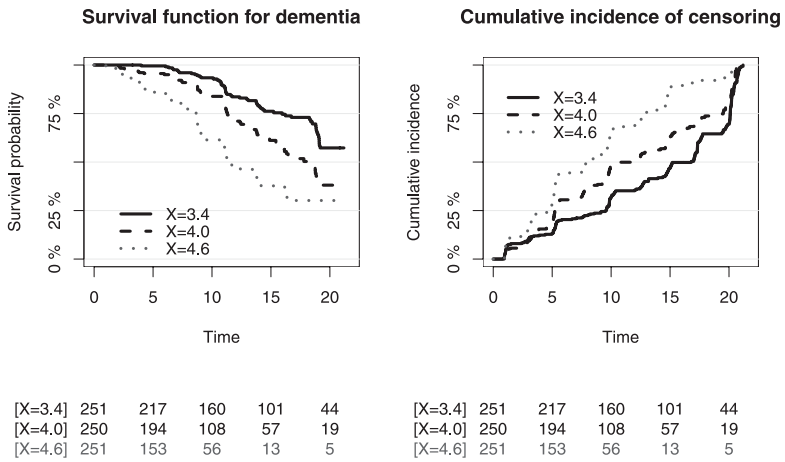


Figure 1 Survival function for dementia and cumulative incidence function for censoring. These curves are computed by conditional Kapan–Meier (formula (6) with $\lambda_n = 0.05$) for the first quartile ($X = 3.4$), the median ($X = 4$), and the third quartile ($X = 4.6$) of the distribution of the marker X. Numbers of subjects at risk at time $t = 0, 5, 10, 15, 20$ years are indicated below the graph.

Table 3 Proportion of censored, demented, and nondemented subjects for each time t ($n = 2516$ subjects).

Time t	Censored before t	Dementia before t	Free of dementia at t
5	566 (22.5%)	128 (5.1%)	1822 (72.4%)
10	1132 (45.0%)	326 (13.0%)	1058 (42.1%)
15	1389 (55.2%)	533 (21.2%)	594 (23.6%)
17	1444 (57.4%)	591 (23.5%)	481 (19.1%)

The mean age at enrollment for this sample was 74.1 (standard deviation = 6.2) and the mean DSST score was 27.5 (standard deviation = 11.5). The marker was defined by $X = 0.073 \times \text{AGE} - 0.052 \times \text{DSST}$ with AGE and DSST measured at enrollment and the coefficients 0.073 and 0.052 are the estimates from a Cox model for time to dementia. The mean of X was 4.00 (standard deviation = 0.89) with a minimum of 1.18 and a maximum of 6.79.

6.3 Results

The association between the marker and the risk of dementia and between the marker and the risk of censoring are illustrated in Fig. 1. A high value of the marker is associated with high risks of both dementia and censoring. For instance, the probability of being free of dementia is estimated at 0.93 and the probability of censoring at 0.31 at time $t = 10$ for the first quartile of the marker distribution, whereas they were, respectively, estimated at 0.63 and 0.61 for the third quartile. To illustrate the association between the censoring time and the marker, a Cox regression model was also fitted. The hazard ratio for one standard deviation increase of the marker value was estimated to $\widehat{HR}_C = 1.65$ (95% confidence interval [1.56, 1.74]). Thus, the censoring time clearly depends on the marker. We computed the ROC curves and the AUCs at time points $t = 5$, 10, 15, and 17 years. Table 3 displays the proportions of censored, demented, and healthy subjects before each time point.

Table 4 $\widehat{AUC}(t)$ of the $\widehat{ROC}(t)$ curve evaluating predictive ability of the cognitive marker ($X = 0.073 \times \text{AGE} - 0.052 \times \text{DSST}$) for dementia (PAQUID, $n = 2561$).

Time t	"Naive"	KM_{HLP}	KM_{CD}	IPCW	NNE	CIPCW
5	0.828	0.832	0.788	0.827	0.802	0.803
10	0.828	0.843	0.724	0.819	0.782	0.798
15	0.807	0.806	0.622	0.784	0.762	0.775
17	0.816	0.813	0.593	0.790	0.759	0.773

We present the $\widehat{ROC}(t)$ curve in Fig. 2 and the corresponding $\widehat{AUC}(t)$ in Table 4. For NNE and CIPCW estimators, due to the large sample size, we chose $\lambda_n = 0.01$ and therefore there were about $2 \times \lambda_n \times n \approx 50$ subjects included in each neighborhood (except for the boundaries) to compute the conditional Kaplan–Meier estimators. Comparison of AUC estimates leads to similar results to those of the simulation study where censoring depended moderately on the marker ($HR_C = 1.35$). For the two estimates that deal with marker-dependent censoring, NNE and CIPCW, results are close (the largest difference is 0.016) although the NNE is always slightly smaller than CIPCW. The "Naive," KM_{HLP}, and IPCW estimates are higher than those of NNE and CIPCW, while the KM_{CD} estimate is smaller than the NNE and CIPCW ones. As expected, the differences between estimates that handle independent censoring and the others increase with increasing time point t and so the percentage of censoring (Tables 3 and 4).

Finally, NNE and CIPCW estimates show that this cognitive marker exhibits good predictive power for the identification of dementia cases. As expected, $\widehat{AUC}(t)$ tends to decrease with increasing window of prediction t, reflecting poorer long-term predictive ability. However, this decline with time appears small, and the predictive ability remains quite good for a window of prediction as long as 17 years. This marker could probably be improved by including several cognitive tests and should be evaluated on a validation sample.

7 Discussion

In this paper, we have reviewed the estimators of *cumulative/dynamic* sensitivity, specificity, and ROC curve previously proposed for censored event-time, and detailed their properties and relationships. We then focused on the behavior of the nonparametric estimators when the censoring time depends on the marker, and thus is only conditionally independent from the event time given the marker. When the censoring depends on the marker, the simulation study has shown that some of these AUC estimators handling censoring may have poorer behavior than the "Naive" one that removes censored subjects. Only the NNE (Heagerty et al., 2000) and the CIPCW are robust in this case. As in Chambless and Diao (2006), our simulations have shown slight biases for the "Naive" estimator when censoring does not depend on the marker, and as in Hung and Chiang (2010a) they have confirmed that IPCW is quite robust to censoring that depends moderately on the marker. Based on the expression of the biases presented in Section 2.2, it is possible to simulate scenarios with independent censoring and larger biases for the "Naive" AUC. However, we failed to find simple realistic scenarios with independent censoring leading to large biases for the "Naive" estimators.

NNE and CIPCW both rely on the nearest neighbor Kaplan–Meier estimator proposed by Akritas (1994) to estimate the conditional survival function either for the event (NNE) or for the censoring (CIPCW). This partly drives their properties. First, NNE may have trouble in circumstances where

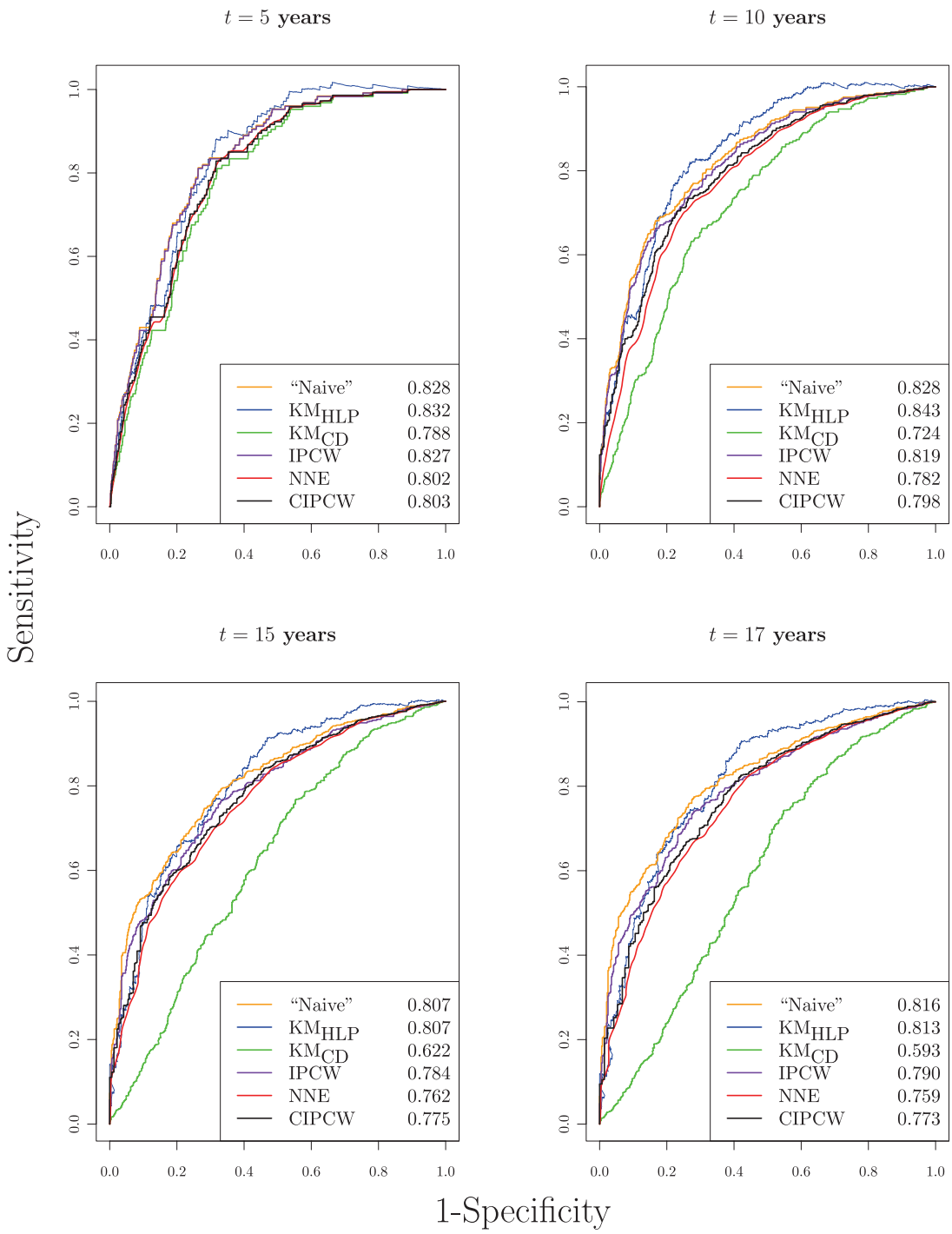


Figure 2 $\widehat{ROC}(t)$ curve and $\widehat{AUC}(t)$ for windows of prediction $t = 5$, 10, 15 and 17 years, estimated with the nonparametric estimators presented in Sections 2, 3, and 4. PAQUID, $n = 2516$ subjects.

there is no subject at risk at the time t of interest in some neighborhoods. However, this problem may only occur with small sample sizes, heavy censoring, and high association between marker and censoring. Performances of NNE and CIPCW should be equivalent with bigger sample sizes. Second, both estimators require the choice of a smoothing parameter λ_n (neighborhood size) but the two extreme values of λ_n correspond to sensible estimators for CIPCW ("Naive" or IPCW), whereas when λ_n tends to 1, NNE tends to a degenerated estimate such as $Se(c, t) = 1 - Sp(c, t)$. Besides, simulation results suggest that CIPCW could be less affected by the choice of λ_n. As most of the ROC curve estimators, CIPCW estimator is also a step function that better highlights the information available in the data compared to a smooth function.

CIPCW can be adapted to other informative censoring than marker-dependent censoring. For instance, if it is more realistic to assume that censoring does not depend on the marker X but on another measured marker Z, we just need to replace the estimator $\widehat{S}_C(\cdot|X_i)$ by $\widehat{S}_C(\cdot|Z_i)$ in the weights of CIPCW to ensure consistency.

CIPCW may also be extended for more complicated diagnostic rules involving several markers or repeated measures of one marker. For instance, Rizopoulos (2011) suggested to consider as positive, subjects with both the baseline measure of the marker X_0 greater than a threshold c and the measure one year later X_1 greater than $0.5c$. Although Rizopoulos (2011) used parametric sensitivity and specificity estimators, CIPCW could be easily extended to this case to propose an alternative nonparametric estimator. Assuming censoring depends only on the last value of the marker computation of CIPCW would only require the estimate of $\widehat{S}_C(t|X_1)$.

Saha and Heagerty (2010) extended NNE for competing risks using cumulative incidence functions, instead of conditional survival functions, to compute sensitivity and specificity. The estimators IPCW and CIPCW may also be generalized for competing risks since cumulative incidence functions can be estimated by IPCW (Scheike et al., 2008). However, this cannot be directly applied to PAQUID data to account for competing risk due to death. Indeed, the censoring times are different for the two events. Dementia may be assessed only if the subject completes the follow-up visit while the vital status is known for every subject. The exact date of death may be collected for all the subjects who died before the final visit (20 years) while the dementia time is interval censored. To our knowledge, only parametric multistate models (Joly et al., 2002) can account for this kind of interval censoring and thus defining a nonparametric ROC curve estimator for this kind of design seems not feasible.

To conclude, we showed that marker-dependent censoring may bias dramatically some ROC curve estimators for independently censored data. Thus, we recommend to use estimators for marker-dependent censoring (CIPCW or NNE) or IPCW that appears quite robust.

Softwares We used the `survivalROC` R package to compute KM_{HLP} and NNE. R code for KM_{CD}, IPCW, CIPCW is available on request from the corresponding author.

Acknowledgments This work was partly funded by a grant from France Alzheimer awarded to Hélène Jacqmin-Gadda in 2009. The PAQUID study, managed by Jean-François Dartigues, is funded by IPSEN and Novartis laboratories. We thank the two anonymous referees and the associate editor for their helpful comments.

Conflict of interest

The authors have declared no conflict of interest.

Appendix: Bias of "Naive" estimators

For all subsets $\mathcal{A}, \mathcal{B}, \mathcal{C}$ with non null probability, we have the equality:

$$\begin{aligned}\mathrm{P}(\mathcal{A}|\mathcal{B}\cap\mathcal{C}) &= \frac{\mathrm{P}(\mathcal{A}\cap\mathcal{B}\cap\mathcal{C})}{\mathrm{P}(\mathcal{B}\cap\mathcal{C})} = \frac{\mathrm{P}(\mathcal{A}\cap\mathcal{B}\cap\mathcal{C})}{\mathrm{P}(\mathcal{A}\cap\mathcal{C})} \times \frac{\mathrm{P}(\mathcal{A}\cap\mathcal{C})}{\mathrm{P}(\mathcal{C})} \times \frac{\mathrm{P}(\mathcal{C})}{\mathrm{P}(\mathcal{B}\cap\mathcal{C})} \\ &= \mathrm{P}(\mathcal{A}|\mathcal{C}) \times \frac{\mathrm{P}(\mathcal{B}|\mathcal{A}\cap\mathcal{C})}{\mathrm{P}(\mathcal{B}|\mathcal{C})}. \qquad \star\end{aligned}$$

Due to the law of large numbers, as $n \to \infty$ we have for the "Naive" sensitivity estimator:

$$\begin{aligned}\widehat{Se}(c,t) = \frac{\frac{1}{n}\sum_{i=1}^{n}\delta_i I(X_i > c) I(T_i^* \le t)}{\frac{1}{n}\sum_{i=1}^{n}\delta_i I(T_i^* \le t)} &\xrightarrow{a.s} \frac{\mathrm{P}\,(X > c, T \le t, T \le C)}{\mathrm{P}\,(T \le t, T \le C)} \\ &= \mathrm{P}\,(X > c | T \le t, T \le C) \\ &= \mathrm{P}\,(X > c | T \le t) \times \frac{\mathrm{P}\,(T \le C | X > c, T \le t)}{\mathrm{P}\,(T \le C | T \le t)} \\ &= Se(c,t) \times \frac{\mathrm{P}\,(T \le C | X > c, T \le t)}{\mathrm{P}\,(T \le C | T \le t)}\end{aligned}$$

by the formula (*) with $\mathcal{A} = \{X > c\}$, $\mathcal{B} = \{T \le C\}$, and $\mathcal{C} = \{T \le t\}$. Similarly, the formula $(*)$ with $\mathcal{A} = \{X \le c\}$, $\mathcal{B} = \{C > t\}$, and $\mathcal{C} = \{T > t\}$ yields to the "Naive" specificity bias:

$$\begin{aligned}\widehat{Sp}(c,t) = \frac{\sum_{i=1}^{n} I(X_i \le c) I(T_i^* > t)}{\sum_{i=1}^{n} I(T_i^* > t)} &\xrightarrow{a.s} \frac{\mathrm{P}(X \le c, T > t, C > t)}{\mathrm{P}(T > t, C > t)} \\ &= Sp(c,t) \times \frac{\mathrm{P}(C > t | X \le c, T > t)}{\mathrm{P}(C > t | T > t)}.\end{aligned}$$

For the "Naive" AUC estimator, due to theory of U-statistics, when $n \to \infty$ we have

$$\begin{aligned}\widehat{AUC}(t) &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i I(T_i^* \le t) I(T_j^* > t) I(X_i > X_j)}{\sum_{i=1}^{n}\delta_i I(T_i^* \le t)\sum_{j=1}^{n} I(T_j^* > t)} \\ &\xrightarrow{a.s} \frac{\mathrm{P}(X_i > X_j, T_i \le C_i, T_i \le t, T_j > t, C_j > t)}{\mathrm{P}(T_i \le C_i, T_i \le t, T_j > t, C_j > t)} \\ &= \mathrm{P}(X_i > X_j | T_i \le C_i, T_i \le t, T_j > t, C_j > t)\end{aligned}$$

$$= \mathrm{P}(X_i > X_j | T_i \le t, T_j > t) \times \frac{\mathrm{P}(T_i \le C_i, C_j > t | T_i \le t, T_j > t, X_i > X_j)}{\mathrm{P}(T_i \le C_i, C_j > t | T_i \le t, T_j > t)}$$

by the formula (*) with $\mathcal{A} = \{X_i > X_j\}$, $\mathcal{B} = \{T_i \le C_i, C_j > t\}$, and $\mathcal{C} = \{T_i \le t, T_j > t\}$.

References

Akritas, M. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics* **22**, 1299–1327.

Amieva, H., Jacqmin-Gadda, H., Orgogozo, J., Le Carret, N., Helmer, C., Letenneur, L., Barberger-Gateau, P., Fabrigoule, C. and Dartigues, J. (2005). The 9 year cognitive decline before dementia of the Alzheimer type: a prospective population-based study. *Brain* **128**, 1093–1101.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Unpublished technical report, University of California, Berkeley.

Cai, T., Gerds, T., Zheng, Y. and Chen, J. (2011). Robust prediction of t-year survival with data from multiple studies. *Biometrics* **67**, 436–444.

Cai, T., Pepe, M., Zheng, Y., Lumley, T. and Jenny, N. (2006). The sensitivity and specificity of markers for event times. *Biostatistics* **7**, 182–197.

Chambless, L. and Diao, G. (2006). Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in Medicine* **25**, 3474–3486.

Chiang, C. and Hung, H. (2010). Non-parametric estimation for time-dependent AUC. *Journal of Statistical Planning and Inference* **140**, 1162–1174.

Foucher, Y., Giral, M., Soulillou, J. and Daures, J. (2010). Time-dependent ROC analysis for a three-class prognostic with application to kidney transplantation. *Statistics in Medicine* **29**, 3079–3087.

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

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.

Heagerty, P., Lumley, T. and Pepe, M. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344.

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

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

Hung, H. and Chiang, C. (2010b). Optimal composite markers for time-dependent receiver operating characteristic curves with censored survival data. *Scandinavian Journal of Statistics* **37**, 664–679.

Hung, H. and Chiang, C.-T. (2011). Nonparametric methodology for the time-dependent partial area under the ROC curve. *Journal of Statistical Planning and Inference* **141**, 3829–3838.

Jacqmin-Gadda, H., Fabrigoule, C., Commenges, D., and Dartigues, J. (1997). A 5-year longitudinal study of the Mini-Mental State Examination in normal aging. *American Journal of Epidemiology* **145**, 498–506.

Joly, P., Commenges, D., Helmer, C. and Letenneur, L. (2002). A penalized likelihood approach for an illness–death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics* **3**, 433–443.

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

Pepe, M. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, Oxford.

Pepe, M., Zheng, Y., Jin, Y., Huang, Y., Parikh, C. and Levy, W. (2008). Evaluating the ROC performance of markers for future events. *Lifetime Data Analysis* **14**, 86–113.

Proust-Lima, C. and Taylor, J. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. *Biostatistics* **10**, 535–549.

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

Saha, P. and Heagerty, P. (2010). Time-dependent predictive accuracy in the presence of competing risks. *Biometrics* **66**, 999–1011.

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

Scheike, T., Zhang, M. and Gerds, T. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika* **95**, 205–220.

Song, X. and Zhou, X. (2008). A semiparametric approach for the covariate-specific ROC curve with survival outcome. *Statistica Sinica* **18**, 947–965.

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

Vellas, B., Andrieu, S., Ousset, P., Ouzid, M. and Mathiex-Fortunet, H. (2006). The guidage study: methodological issues. A 5-year double-blind randomized trial of the efficacy of egb 761 (r) for prevention of alzheimer disease in patients over 70 with a memory complaint. *Neurology* **67**(Suppl 3), S6.

Viallon, V. and Latouche, A. (2011). Discrimination measures for survival outcomes: connection between the AUC and the predictiveness curve. *Biometrical Journal* **53**, 217–236.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale* (rev. ed.). Psychological Corporation, New York.

III.2 Discussion complémentaire

III.2.1 À propos des estimateurs CIPCW et de propriétés asymptotiques

Intuitivement, la pondération $\delta_i/\widehat{S}_C(T_i^*|X_i)$ de l'estimateur CIPCW de la sensibilité de l'article précédent se justifie par

$$\begin{aligned}
\mathbb{E}\left\{\mathbb{1}_{(T_i^*\leqslant t)}\frac{\delta_i}{S_C(T_i^*|X_i)}\right\} &= \mathbb{E}\left\{\mathbb{E}\left[\mathbb{1}_{(T_i\leqslant t)}\frac{\mathbb{1}_{(T_i\leqslant C_i)}}{S_C(T_i|X_i)}\middle| T_i, X_i\right]\right\} \qquad (\text{car} \quad \delta_i = 1 \Rightarrow T_i^* = T_i) \\
&= \mathbb{E}\left\{\mathbb{1}_{(T_i\leqslant t)}\frac{\mathbb{E}\left[\mathbb{1}_{(T_i\leqslant C_i)}\middle| T_i, X_i\right]}{S_C(T_i|X_i)}\right\} \\
&= \mathbb{E}\left\{\mathbb{1}_{(T_i\leqslant t)}\frac{S_C(T_i|X_i)}{S_C(T_i|X_i)}\right\} = \mathbb{P}\left(T \leqslant t\right),
\end{aligned}$$

avec les notations de l'article $T_i^* = \min(T_i, C_i)$, $\delta_i = \mathbb{1}_{(T_i\leqslant C_i)}$ et pour tout u, $S_C(u|X_i) = \mathbb{P}(C_i > u|X_i)$. Dans ce raisonnement, l'hypothèse d'une censure C_i indépendante de T_i conditionnellement à X_i est utilisée pour avoir $\mathbb{E}\left[\mathbb{1}_{(T_i\leqslant C_i)}\middle| T_i, X_i\right] = S_C(T_i|X_i)$. La pondération $1/\widehat{S}_C(t|X_i)$ se justifie similairement.

Cette argumentation est en fait étroitement liée aux idées des « intégrales de Kaplan-Meier » en présence de censure dépendant du marqueur, récemment proposées par Lopez (2011). D'ailleurs, en suivant le même raisonnement que la preuve du théorème 3.1 de Lopez (2011), on peut montrer rigoureusement que, sous des conditions faibles, les estimateurs proposés sont convergents. Cependant, deux légères différences existent entre les estimateurs CIPCW et les « intégrales de Kaplan-Meier » récemment proposées par Lopez (2011). Premièrement, l'estimateur de Kaplan-Meier conditionnel utilisé n'est pas le même. Nous utilisons celui de Akritas (1994) et non celui de Beran (1981). Par ailleurs, pour la spécificité, nous utilisons le poids $1/\widehat{S}_C(t|X_i)$ pour tenir compte spécifiquement de l'horizon de prédiction t. Pour ces deux raisons, on ne peut donc pas directement appliquer les résultats de Lopez (2011) pour justifier, entre autres, la normalité asymptotique des estimateurs CIPCW via des décompositions i.i.d (Lopez, 2011, théorème 3.3). Pour autant, en reprenant les raisonnements de Lopez (2011), des résultats similaires pourraient probablement être montrés. Nous n'avons pas cherché à les montrer dans cette thèse car, en plus de la difficulté technique non négligeable que

cela représenterait, ils auraient probablement été d'un intérêt assez limité, en pratique, pour nos applications. En effet, contrairement aux décompositions i.i.d des estimateurs des prochains chapitres, les décompositions i.i.d d'estimateurs basés sur des poids du type $\delta_i/\widehat{S}_C(T_i^*|X_i)$ sont souvent compliquées à estimer. En pratique, on préfère donc généralement utiliser des méthodes de Bootstrap, plus usuelles et plus informelles, pour calculer des variances et faire de l'inférence.

III.2.2 À propos de la censure dépendante et des décès

Dans l'article précédent, l'application aux données de la cohorte Paquid illustre bien qu'une censure dépendant fortement de l'outil pronostique étudié peut induire des différences sensibles entre les estimations des différents estimateurs.

Sur ces données, la censure apparaît cependant dépendante de l'outil pronostique principalement car certains sujets sont considérés comme censurés parce que non revus à cause de leur décès. L'outil prédictif étant une combinaison linéaire de l'âge et d'un test psychométrique, il est naturellement associé au décès, et par conséquent à la censure dans ce cas.

Or, comme on l'a vu en Section II.1.3 et à la Figure II.3, un sujet décédé (sans démence) ne devrait pas être considéré comme censuré. Il devrait plutôt être considéré comme ayant subi un risque concurrent, car suite à son décès un individu n'est plus à risque de démence.

Cependant, comme on l'a rappelé dans la discussion de l'article, il n'est pas, à notre connaissance, possible de définir des estimateurs non paramétriques de courbes ROC dépendant du temps « parfaitement » adaptés aux données censurées par intervalle de la cohorte Paquid. Néanmoins, en considérant une règle d'imputation « brutale » (mais acceptable), comme on le reverra au chapitre suivant, on peut se ramener au cas d'une situation de risques concurrents plus usuelle, sans censure par intervalle, avec uniquement une censure à droite. D'où notre motivation pour le travail du chapitre suivant, qui porte sur les courbes ROC dépendant du temps en présence de risques concurrents.

IV. Courbe ROC dépendant du temps et risques concurrents

IV.1 Publication dans *Statistics in Medicine*

Résumé :

Motivé par le souhait de tenir compte du risque concurrent de décès sans démence pour nos applications aux données de la cohorte Paquid, dans ce second article on s'intéresse aux courbes ROC dépendant du temps en présence de risques concurrents. On s'intéresse en particulier à deux définitions cumulatives/dynamiques des cas et des contrôles, et aux définitions des aires sous la courbe ROC associées, pour lesquelles on propose des estimateurs IPCW (« Inverse Probability of Censoring Weighting »). Quelques propriétés asymptotiques sont étudiées, et des intervalles de confiance et des tests en sont dérivés. Une étude de simulations est ensuite présentée pour examiner le comportement des méthodes d'inférence proposées à taille d'échantillon finie.

Enfin, l'application des méthodes aux données de la cohorte Paquid clôt le manuscrit. Son objectif est la comparaison des capacités de deux tests psychométriques à prédire la démence chez les sujets âgés, en tenant compte du risque concurrent de décès sans démence.

Ci-après sont présentés le manuscrit principal et son *web-appendix*.

Special Issue Paper

Statistics in Medicine

Received 6 November 2012, Accepted 2 August 2013 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5958

Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks

Paul Blanche,[a,b*†] Jean-François Dartigues[a,b] and Hélène Jacqmin-Gadda[a,b]

The area under the time-dependent ROC curve (AUC) may be used to quantify the ability of a marker to predict the onset of a clinical outcome in the future. For survival analysis with competing risks, two alternative definitions of the specificity may be proposed depending of the way to deal with subjects who undergo the competing events. In this work, we propose nonparametric inverse probability of censoring weighting estimators of the AUC corresponding to these two definitions, and we study their asymptotic properties. We derive confidence intervals and test statistics for the equality of the AUCs obtained with two markers measured on the same subjects. A simulation study is performed to investigate the finite sample behaviour of the test and the confidence intervals. The method is applied to the French cohort PAQUID to compare the abilities of two psychometric tests to predict dementia onset in the elderly accounting for death without dementia competing risk. The 'timeROC' R package is provided to make the methodology easily usable. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: AUC; competing risks; discrimination; inverse probability of censoring weighting; prognosis; survival analysis

1. Introduction

It is often useful to identify markers that enable the discrimination between subjects at high and low risk of a disease in the future. Markers with high predictive accuracy help clinicians for early medical decisions and thus may reduce morbidity and mortality in the high-risk population. Identification of such markers are therefore a major concern in medical research. For instance, the level of prostate-specific antigen is often used to identify subjects at high risk of prostate cancer recurrence and to guide treatment management. In Alzheimer's disease, the most frequent dementia in the elderly, some studies suggest that the decline in cognitive functions begins long before all the criteria for the clinical diagnosis of dementia are reached [1]. These results suggest that cognitive tests could be useful for identification of subjects with high risk of Alzheimer's disease in the coming years. As treatment against Alzheimer's disease have only demonstrated a modest efficiency, current research is focusing on preventive treatments that could be administered in the pre-diagnosis phase to subjects at high risk of dementia [2]. In this context, relevant questions are as follows. Are psychometric tests good markers for discrimination of subjects with respect to their risk of Alzheimer's disease in the few years following a test result? Are some psychometric tests better than others for this task? Data from large prospective cohort studies can be used to answer these questions. To evaluate prognostic abilities on such data, statistical methodology

[a]*University Bordeaux, ISPED, Centre INSERM U897-Epidémiologie-Biostatistique, F-33000 Bordeaux, France*
[b]*INSERM, ISPED, Centre INSERM U897-Epidémiologie-Biostatistique, F-33000 Bordeaux, France*
**Correspondence to: Paul Blanche, INSERM U897 - Equipe de biostatistique, ISPED, Université Bordeaux Segalen, 146 rue Leo Saignat, 33076 Bordeaux cedex, France.*
[†]*E-mail: Paul.Blanche@isped.u-bordeaux2.fr*

have (i) to account for censoring due to lost of follow-up and (ii) to deal with the competing risk of death without dementia, which is a major issue in the elderly population.

Motivated by the prediction of Alzheimer's disease, the goal of this paper is to propose a method to estimate and compare the predictive accuracy of different rival markers using the ROC methodology, accounting for censoring and competing events.

The ROC methodology was originally introduced in medicine for the evaluation of diagnostic accuracy of quantitative markers. The ROC curve displays the sensitivity (true positive rate) versus 1-specificity (false positive rate) for all possible cutpoints that define a binary test, by dichotomizing a quantitative marker. The area under the ROC curve (AUC) is often used to summarize and compare diagnostic accuracy of several markers. The AUC may be interpreted as a concordance index. Indeed, the AUC is equal to the probability that the marker value of a randomly chosen diseased subject is above the marker value of a randomly chosen healthy subject.

For the evaluation of the predictive accuracy, Heagerty *et al.* [3, 4] introduced the time-dependent ROC curves and proposed several definitions of cases and controls. In particular, they introduced the so-called *cumulative/dynamic* definition, where a case is a subject diagnosed before a time point t and a healthy subject is free of the disease at time t. More recently, Saha and Heagerty [5] and Zheng *et al.* [6] extended this definition for the competing risks setting. As noticed by Zheng *et al.* [6], two definitions of the specificity can be considered depending of the way to deal with subjects who undergo the competing events. Such subjects can be considered as controls or not, depending on the clinical setting.

Without censoring, sensitivity and specificity can be estimated by empirical true positive and true negative fractions [7] and AUCs are often compared with the test of DeLong *et al.* [8]. These approaches assume that all subjects can be classified as cases or controls. However, when data are censored, the status of subjects lost of follow-up before time t is unknown. In the standard situation without competing risks, Uno *et al.* [9] and Hung and Chiang [10] proposed a nonparametric estimator of the time-dependent ROC curve using the inverse probability of censoring weighting (IPCW) approach, and Chiang and Hung [11] proposed a test for comparing AUCs. Besides, simulation studies have shown the good behaviour of the IPCW approach for various censoring scenarios [11, 12].

In the competing risks setting, only estimators that rely on parametric [13], semiparametric [6] or nonparametric smooth estimators that require a bandwidth selection [5] were proposed. In addition, no test was proposed to compare AUCs of different markers. In this paper, we first extend the IPCW estimators for ROC curves and AUCs to the competing risks setting for both definitions of specificity. By contrast to previous nonparametric proposed methods, this nonparametric approach does not require any bandwidth selection and has practical computational advantages, avoiding bootstrapping for making inference and being implemented in the `timeROC` R package. Moreover, whereas previous works were mainly focused on ROC curve estimation, this paper mainly focusses on providing and studying a test for comparing AUCs of two rival markers on censored data with competing events. We also propose some extensions such as estimators that properly account for marker dependent censoring when comparing AUCs or estimation of confidence bands.

The paper is organized as follows. Section 2 presents the notations and the definitions of the time-dependent ROC curves in presence of competing risks. Section 3 describes the proposed estimators, the asymptotic results and the inference procedures. The finite sample performances of the inference procedures are evaluated by simulations in Section 4. Section 5 presents the application of the proposed method to the comparison of two cognitive tests to predict dementia onset in the elderly. Finally, Section 6 concludes the paper contrasting our work with the previous ones and discussing some perspectives.

2. Receiver operating characteristic curves with competing risks

2.1. Notations

Let M denote a marker that is measured at baseline. Let T denote the event-time, C the censoring time and $\Delta = \mathbb{1}_{(T \leqslant C)}$ the censoring indicator, with $\mathbb{1}_{(\cdot)}$ denoting the indicator function. Let K denote the number of competing event and η the type of event. Let $\widetilde{T} = \min(T, C)$, the observed time and $\widetilde{\eta} = \Delta\eta$, which indicates either the type of event (when $\widetilde{\eta} \in \{1, \ldots, K\}$) or a censored observation (when $\widetilde{\eta} = 0$). Hereafter, for all time point t, we denote $S_T(t) = \mathbb{P}(T > t)$ and $G(t) = \mathbb{P}(C > t)$ and $\widehat{S}_T(t)$ and $\widehat{G}(t)$ the Kaplan–Meier estimators of $S_T(t)$ and $G(t)$, $S_{\widetilde{T}}(t) = \mathbb{P}(\widetilde{T} > t)$ and $\widehat{S}_{\widetilde{T}}(t)$ its empirical estimator. To make formulae easier to read, we also introduce the weight $W(t) = 1/G(t)$ and its estimator

$\widehat{W}(t) = 1/\widehat{G}(t)$.
We observe the independent and identically distributed (i.i.d.) sample of n subjects $\{(\widetilde{T}_i, \Delta_i, \widetilde{\eta}_i, M_i), i = 1, \ldots, n\}$. To simplify the presentation of the estimators, we assume that the marker M is measured on a continuous scale without ties. Adaptation for ties is discussed in Section 3.5.

2.2. *Definitions of receiver operating characteristic curves and area under the receiver operating characteristic curves*

With competing events, definition of cases is clear but for controls, Zheng *et al.* [6] considered two definitions leading to two different definitions of the time-dependent specificity.

For clarity, we suppose that we are interested in the assessment of the predictive accuracy for the first type of event (called main event) corresponding to $\eta = 1$. Cases at time t are defined as subjects who undergo the main event $\eta = 1$ before time t, that is, subjects i with $T_i \leqslant t, \eta_i = 1$. Without loss of generality, we assume that larger values of the marker M are associated with higher risks of events. Then, for a threshold $c \in \mathbb{R}$ the sensitivity at time t is defined by

$$Se(c,t) = \mathbb{P}\left(M > c | T \leqslant t, \eta = 1\right).$$

Controls at time t were originally defined as event-free subjects at time t, that is, subjects j with $T_j > t$ [5]. In the following, we denote '*control**' (with an asterix) the controls for this definition that leads to a specificity at time t defined by

$$Sp^*(c,t) = \mathbb{P}\left(M \leqslant c | T > t\right). \tag{1}$$

According to this definition, subjects who experience another type of event before time t, that is, subjects j with $T_j \leqslant t, \eta_j \neq 1$, are neither cases nor controls.
Alternatively, a control may be defined as a subject who is not a case, that is, a subject j with $T_j > t$ or $T_j \leqslant t, \eta_j \neq 1$ [6]. In the following, we denote '*control*' (without an asterix) the controls for this definition, leading to the specificity at time t:

$$Sp(c,t) = \mathbb{P}\left(M \leqslant c | \{T > t\} \cup \{T \leqslant t, \eta \neq 1\}\right). \tag{2}$$

Two different time-dependent ROC curves can be obtained plotting $Se(c,t)$ versus either $1 - Sp^*(c,t)$ or $1 - Sp(c,t)$. As with usual ROC curve, the AUC can be shown to be the probability that the marker of a case is greater than the marker of a control [7]. As a consequence, the AUC at time t for both definitions are

$$AUC^*(t) = \mathbb{P}\left(M_i > M_j | T_i \leqslant t, \eta_i = 1, T_j > t\right) \tag{3}$$

$$\text{and} \quad AUC(t) = \mathbb{P}\left(M_i > M_j \middle| T_i \leqslant t, \eta_i = 1, \{T_j > t\} \cup \{T_j \leqslant t, \eta_j \neq 1\}\right) \tag{4}$$

with i and j the indexes of two independent subjects. In the following, we will use AUC for the generic term and $AUC^*(t)$ or $AUC(t)$ for definitions (3) or (4).

Thus, subjects who meet one of the competing events before time t do not contribute to $AUC^*(t)$, although they are considered as control in $AUC(t)$. Applied to the example of dementia (as main event) and death without dementia (as competing event), this means $AUC^*(t)$ evaluates discrimination between demented subjects and subjects alive and non-demented at the time point t of interest, whereas $AUC(t)$ evaluates discrimination between demented subjects and subjects non-demented at the time point t or dead without dementia in the window of prediction. Given that many predictors of dementia are also predictors of death, the two measures may give very different results.

3. Inverse probability of censoring weighting estimators and inference procedures

3.1. *Inverse probability of censoring weighting estimators*

Without competing risks, Uno *et al.* [9] and Hung and Chiang [10] proposed IPCW estimators for the ROC curve and the AUC with censored data. A more general theory about IPCW can be found in [14] or [15]. The rational of the IPCW approach is to mainly use the observed cases and controls and to weight them by their probability of being observed. In the competing risk setting, *observed cases* are subjects i with $\widetilde{T}_i \leqslant t$ and $\widetilde{\eta}_i = 1$; *observed controls** are the uncensored event-free subjects at t, that is, subjects

j with $\widetilde{T}_j > t$, and *observed controls* are either uncensored and event-free or subjects who met a competing event before t, that is, subjects j with $\widetilde{T}_j > t$ or $\widetilde{T}_j \leqslant t, \widetilde{\eta}_j \notin \{0,1\}$. Subjects censored before t, that is, subjects i with $\widetilde{T}_i \leqslant t, \widetilde{\eta}_i = 0$, are only used to estimate the weights.

Assuming that the censoring C is independent of (T, η, M), we propose to estimate sensitivity $Se(c,t)$ by

$$\widehat{Se}(c,t) = \frac{\sum_{i=1}^{n} \mathbb{1}_{(M_i > c)} \mathbb{1}_{(\widetilde{T}_i \leqslant t, \widetilde{\eta}_i = 1)} \widehat{W}\left(\widetilde{T}_i\right)}{\sum_{i=1}^{n} \mathbb{1}_{(\widetilde{T}_i \leqslant t, \widetilde{\eta}_i = 1)} \widehat{W}\left(\widetilde{T}_i\right)}. \tag{5}$$

Heuristically, the weighting $\widehat{W}\left(\widetilde{T}_i\right) = 1/\widehat{G}\left(\widetilde{T}_i\right)$ is justified by the fact that as $n \to \infty$, the value of $n^{-1}\times$ numerator of $\widehat{Se}(c,t)$ converges to

$$\begin{aligned}
\mathbb{E}\left\{\frac{\mathbb{1}_{(M_i > c)} \mathbb{1}_{(\widetilde{T}_i \leqslant t, \widetilde{\eta}_i = 1)}}{G\left(\widetilde{T}_i\right)}\right\} &= \mathbb{E}\left\{\mathbb{E}\left(\frac{\mathbb{1}_{(M_i > c)} \mathbb{1}_{(T_i \leqslant t)} \mathbb{1}_{(\eta_i = 1)} \mathbb{1}_{(T_i \leqslant C_i)}}{G(T_i)} \middle| T_i, \eta_i, M_i\right)\right\} \\
&= \mathbb{E}\left\{\frac{\mathbb{1}_{(M_i > c)} \mathbb{1}_{(T_i \leqslant t)} \mathbb{1}_{(\eta_i = 1)}}{G(T_i)} \mathbb{E}\left(\mathbb{1}_{(T_i \leqslant C_i)} \middle| T_i, \eta_i, M_i\right)\right\} \\
&= \mathbb{E}\left\{\frac{\mathbb{1}_{(M_i > c)} \mathbb{1}_{(T_i \leqslant t)} \mathbb{1}_{(\eta_i = 1)}}{G(T_i)} G(T_i)\right\} = \mathbb{P}\left(M > c, T \leqslant t, \eta = 1\right)
\end{aligned}$$

and similarly the value of $n^{-1}\times$ denominator of $\widehat{Se}(c,t)$ converges to $\mathbb{P}\left(T \leqslant t, \eta = 1\right)$, both together leading to a consistent estimator of sensitivity.
By analogy, specificity $Sp^*(c,t)$ is estimated by

$$\widehat{Sp}^*(c,t) = \frac{\sum_{i=1}^{n} \mathbb{1}_{(M_i \leqslant c)} \mathbb{1}_{(\widetilde{T}_i > t)} \widehat{W}(t)}{\sum_{i=1}^{n} \mathbb{1}_{(\widetilde{T}_i > t)} \widehat{W}(t)} = \frac{\sum_{i=1}^{n} \mathbb{1}_{(M_i \leqslant c)} \mathbb{1}_{(\widetilde{T}_i > t)}}{\sum_{i=1}^{n} \mathbb{1}_{(\widetilde{T}_i > t)}} \tag{6}$$

and specificity $Sp(c,t)$ by

$$\widehat{Sp}(c,t) = \frac{\sum_{i=1}^{n} \mathbb{1}_{(M_i \leqslant c)} \left(\mathbb{1}_{(\widetilde{T}_i > t)} \widehat{W}(t) + \mathbb{1}_{(\widetilde{T}_i \leqslant t, \widetilde{\eta}_i \notin \{0,1\})} \widehat{W}(\widetilde{T}_i)\right)}{\sum_{i=1}^{n} \left\{\mathbb{1}_{(\widetilde{T}_i > t)} \widehat{W}(t) + \mathbb{1}_{(\widetilde{T}_i \leqslant t, \widetilde{\eta}_i \notin \{0,1\})} \widehat{W}(\widetilde{T}_i)\right\}}. \tag{7}$$

The resulting estimated ROC curves are increasing step functions. As with usual empirical ROC curve for uncensored data [7, p. 103], summing rectangular areas of widths and heights corresponding to increases in specificity and sensitivity, it can easily be shown that the area under the estimated ROC curves corresponding to the two specificity estimators (6) and (7) are given by

$$\widehat{AUC}^*(t) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}_{(\widetilde{T}_i \leqslant t, \widetilde{\eta}_i = 1)} \widehat{W}(\widetilde{T}_i) \mathbb{1}_{(\widetilde{T}_j > t)} \widehat{W}(t) \mathbb{1}_{(M_i > M_j)}}{\left(\sum_{i=1}^{n} \mathbb{1}_{(\widetilde{T}_i \leqslant t, \widetilde{\eta}_i = 1)} \widehat{W}(\widetilde{T}_i)\right)\left(\sum_{j=1}^{n} \mathbb{1}_{(\widetilde{T}_j > t)} \widehat{W}(t)\right)} \tag{8}$$

and

$$\widehat{AUC}(t) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}_{(\widetilde{T}_i \leqslant t, \widetilde{\eta}_i = 1)} \widehat{W}(\widetilde{T}_i) \left(\mathbb{1}_{(\widetilde{T}_j > t)} \widehat{W}(t) + \mathbb{1}_{(\widetilde{T}_j \leqslant t, \widetilde{\eta}_j \notin \{0,1\})} \widehat{W}(\widetilde{T}_j)\right) \mathbb{1}_{(M_i > M_j)}}{\left(\sum_{i=1}^{n} \mathbb{1}_{(\widetilde{T}_i \leqslant t, \widetilde{\eta}_i = 1)} \widehat{W}(\widetilde{T}_i)\right)\left(\sum_{j=1}^{n} \left\{\mathbb{1}_{(\widetilde{T}_j > t)} \widehat{W}(t) + \mathbb{1}_{(\widetilde{T}_j \leqslant t, \widetilde{\eta}_j \notin \{0,1\})} \widehat{W}\left(\widetilde{T}_j\right)\right\}\right)} \tag{9}$$

Let us recall that the IPCW estimator of the cumulative incidence function $F_1(t) = \mathbb{P}\left(T \leqslant t, \eta = 1\right)$ defined by

$$\widehat{F}_1(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(\widetilde{T}_i \leqslant t, \widetilde{\eta}_i = 1)} \widehat{W}\left(\widetilde{T}_i\right) \tag{10}$$

is equal to the usual nonparametric maximum likelihood estimator, defined by $\widehat{F}_1(t) = \sum_{i=1}^{n} \widehat{\lambda}_1\left(\widetilde{T}_i\right) \widehat{S}_T\left(\widetilde{T}_i\right)$ where $\widehat{\lambda}_1(t) = \frac{\sum_{j=1}^{n} \mathbb{1}_{(\widetilde{\eta}_j = 1, \widetilde{T}_j = t)}}{\sum_{j=1}^{n} \mathbb{1}_{(\widetilde{T}_j \geqslant t)}}$ [16, 17]. Besides, by the equality $\widehat{S}_T(t) =$

$1-\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{(\widetilde{T}_i\leqslant t)}\Delta_i\widehat{W}\left(\widetilde{T}_i\right)$ that represents the Kaplan–Meier estimator as an IPCW estimator [18], then $1-\widehat{F}_1(t)=\frac{1}{n}\sum_{j=1}^{n}\left\{\mathbb{1}_{(\widetilde{T}_j>t)}\widehat{W}(t)+\mathbb{1}_{(\widetilde{T}_j\leqslant t,\widetilde{\eta}_j\notin\{0,1\})}\widehat{W}\left(\widetilde{T}_j\right)\right\}$. As a consequence, the denominator of formulae (8) and (9) are respectively equal to the more compact form $\widehat{S}_T(t)\widehat{F}_1(t)$ and $\left(1-\widehat{F}_1(t)\right)\widehat{F}_1(t)$.
In both formulae (8) and (9), the AUC estimator is the ratio of the estimated probability of observing a pair of a case and a control with ordered markers over the estimated probability of observing a pair with a case and a control.

3.2. Large sample properties

Only results about $\widehat{AUC}(t)$ are presented in the following. Results about $\widehat{AUC}^*(t)$, based on lemma 2 provided in the web Supporting Information[‡] B, are similar.
Assume that $\tau_1>\inf\{u:F_1(u)>0\}$ and $\tau_2<\sup\left\{u:S_{\widetilde{T}}(u)>0\right\}$. Thus, $[\tau_1,\tau_2]$ represents a period of times t in which there is both a non-null probability of observing a main event before time t and a non-null probability of observing someone at risk at time t.

Lemma 1
Let us assume that the censoring time C is independent of (T,η,M), then for all time t in $[\tau_1,\tau_2]$:

$$\sqrt{n}\left(\widehat{AUC}(t)-AUC(t)\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{IF}\left(\widetilde{T}_i,\widetilde{\eta}_i,M_i,t\right)+o_p(1)$$

where $\mathbb{E}\left[\mathrm{IF}\left(\widetilde{T},\widetilde{\eta},M,t\right)\right]=0$. The influence function of the estimator $\mathrm{IF}(\cdot)$ is detailed in the appendix.

Proof
The proof is the adaptation to the competing risks setting of the proof of Theorem 1 of [10] and is given in the web Supporting Information A. □

From the decomposition of the estimator $\widehat{AUC}(t)$ as a sum of i.i.d. terms in lemma 1, it follows that

$$\sqrt{n}\left(\widehat{AUC}(t)-AUC(t)\right)\xrightarrow{\mathcal{D}}\mathcal{N}\left(0,\sigma_t^2\right).$$

The variance of the influence function σ_t^2 can be consistently estimated by the empirical estimator

$$\widehat{\sigma}_t^2=\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathrm{IF}}\left(\widetilde{T}_i,\widetilde{\eta}_i,M_i,t\right)^2 \tag{11}$$

where $\widehat{\mathrm{IF}}\left(\widetilde{T},\widetilde{\eta},M,t\right)$ is a simple plug-in estimator that is detailed in the Appendix. Therefore, we obtain the asymptotic $(1-\alpha)$-level confidence interval (CI)

$$\left[\widehat{AUC}(t)-z_{1-\alpha/2}\frac{\widehat{\sigma}_t}{\sqrt{n}}\,,\ \widehat{AUC}(t)+z_{1-\alpha/2}\frac{\widehat{\sigma}_t}{\sqrt{n}}\right]$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the univariate standard normal distribution.

3.3. Test for comparing the area under the receiver operating characterisitc curves

Let M_1 and M_2 denote two rival markers measured on the same subject, $AUC_{M_1}(t)$ and $AUC_{M_2}(t)$ the areas under their time-dependent ROC curve and $\widehat{AUC}_{M_1}(t)$ and $\widehat{AUC}_{M_2}(t)$ their estimators. Assuming that the censoring C is independent of (T,η,M_1,M_2), under the null hypothesis $\mathcal{H}_0:AUC_{M_1}(t)=AUC_{M_2}(t)$, it follows from lemma 1 that

$$\frac{\sqrt{n}}{\widehat{\sigma}_{t12}}\left(\widehat{AUC}_{M_1}(t)-\widehat{AUC}_{M_2}(t)\right)\xrightarrow{\mathcal{D}}\mathcal{N}(0,1)$$

[‡]*Supporting information may be found in the online version of this article.*

where

$$\widehat{\sigma}^2_{t12} = \frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{\mathrm{IF}}\left(\widetilde{T}_i, \widetilde{\eta}_i, M_{1i}, t\right) - \widehat{\mathrm{IF}}\left(\widetilde{T}_i, \widetilde{\eta}_i, M_{2i}, t\right)\right\}^2. \tag{12}$$

Alternatively, σ^2_{t12} may also be estimated by bootstrapping.

3.4. Extension for marker-dependent censoring

The previous methodology assumes that the censoring time C is independent of (T, η, M). It is sometimes more realistic to assume that C is only independent of (T, η) given the marker M. Indeed, the marker may be associated with the risk of censoring. The IPCW estimators can be adapted to deal with this lighter assumption by replacing in the weights, the marginal Kaplan–Meier estimator $\widehat{G}(t)$ by any consistent estimator of the conditional survival function $G(t|M) = \mathbb{P}(C > t|M)$. Nonparametric estimators proposed by Beran [19] and Akritas [20] can be used for this task.

However, to build a test for comparing two or more rival markers measured on the same subject, it is necessary to account for possible dependency of censoring on all the markers being compared. Thereby, in presence of marker-dependent censoring, assuming that the censoring time C is independent of (T, η) given $\mathbf{M}$, where $\mathbf{M}$ is the vector of rival markers ($\mathbf{M} = (M_1, M_2)^t$ for two rival markers), the weights should be based on an estimator $\widehat{G}(t|\mathbf{M})$ of $G(t|\mathbf{M}) = \mathbb{P}(C > t|\mathbf{M})$. As nonparametric estimators are often not efficient with moderate sample size and several explanatory variables, semiparametric estimators are therefore favoured.

Estimators of sensitivity, specificity and AUCs can be adapted replacing the marginal Kaplan–Meier $\widehat{G}(\cdot)$ by any semiparametric estimator of $G(\cdot|\mathbf{M})$ in the weights of formulae (5) to (9). Note that the weights no longer cancel in formula (6) in this case, because for $i \neq j$ then $G(\cdot|\mathbf{M}_i) \neq G(\cdot|\mathbf{M}_j)$ in general. Under the assumption that the censoring mechanism is correctly specified, these estimators are unbiased. Decompositions of these estimators as a sum of i.i.d terms as presented in lemma 1 can also be obtained for most of usual semiparametric estimators of $G(\cdot|\mathbf{M})$, including those derived from proportional or additive hazards models. However, the influence function of these estimators of $G(\cdot|\mathbf{M})$ is complex [21] and so are the influence functions of the AUC estimators. Therefore, a bootstrap resampling method is required to estimate the variances.

3.5. Adaptation for ties and markers with ordered discrete results

In practice, samples often include ties in the marker values, especially when the marker is measured on a discrete scale. Definitions and estimators of sensitivity and specificities are still valid in this case. However, in all formulae of estimators of $AUC(t)$ and $AUC^*(t)$, the term $\mathbb{1}_{(M_i > M_j)}$ need to be replaced by $\mathbb{1}_{(M_i > M_j)} + \frac{1}{2}\mathbb{1}_{(M_i = M_j)}$. This leads to a consistent estimator of the area under the ROC curve defined by the linear interpolation between two points. In this case, the AUC has a slightly different interpretation: this is the probability that the marker of a case is greater than the marker of a control, plus half the probability that the marker of a case is equal to the marker of a control. We refer to Section 4.5 of [7] for a more thorough discussion about the ROC curve for ordinal marker.

3.6. Adjusted tests for multiple comparisons

In prognostic studies, several time points t_l, $l = 1, \ldots, L$ are often of interest for clinicians who aim at evaluating the predictive accuracy of biomarkers in more or less long term. Thus, ROC curves may be estimated for different windows of prediction t_l, and the corresponding AUC of two rival markers may be compared at all these different times t_l. When several tests are performed, the p-values must be adjusted for multiple testing. Using the same approach as in the proof of lemma 1, under the null hypotheses $\mathcal{H}_0^l : AUC_{M_1}(t_l) = AUC_{M_2}(t_l), \quad l = 1, \ldots, L$, it can be shown that the vector of the L test statistics, which is the vector of L standardized differences of AUC estimated at time points t_l, as in Section 3.3, converges to a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Sigma_L)$. The covariance terms of the covariance-matrix Σ_L may be estimated similarly to the variance terms in Sections 3.2 and 3.3, either by empirical estimator and estimated influence functions $\widehat{\mathrm{IF}}(\cdot)$ or by Bootstrap. These results may be used either for performing a multivariate test or for computing asymptotically exact adjusted p-value for multiple univariate tests. The asymptotically exact adjusted p-value for testing $\mathcal{H}_0^l : AUC_{M_1}(t_l) = AUC_{M_2}(t_l)$ is

computed as

$$p\text{-value}(t_l) = \mathbb{P}\left(\max|\mathbf{Z}| > |z(t_l)|\right)$$

where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma_L)$ and $z(t_l)$ is the value of the realization of the test statistic at time point t_l [22].

3.7. Simultaneous confidence bands

When an interval of time points is of interest, say $[t_{\min}, t_{\max}]$, it may be desirable to compute a confidence band. A so-called $(1-\alpha)$-simultaneous confidence band for the curve $\{(t, AUC(t)), t \in [t_{\min}, t_{\max}]\}$ is a region containing this curve with probability level $1-\alpha$. Note that the simultaneous confidence band is by definition larger that the band of pointwise CIs. We propose to compute an asymptotic $(1-\alpha)$-simultaneous confidence band by

$$\left[\widehat{AUC}(t) - \widehat{q}_{1-\alpha}\frac{\widehat{\sigma}_t}{\sqrt{n}},\ \widehat{AUC}(t) + \widehat{q}_{1-\alpha}\frac{\widehat{\sigma}_t}{\sqrt{n}}\right], \quad t \in [t_{\min}, t_{\max}]$$

where $\widehat{q}_{1-\alpha}$ is estimated by the following simulation technique, paralleling the approaches of [10, 21, 23, 24] among others, and making use of lemma 1:

Step 1: Generate a random sample $\left(\omega_1^b, \ldots, \omega_n^b\right)$ from n independent standard normal distributions.
Step 2: Using step 1 and the estimator $\widehat{\mathrm{IF}}(\cdot)$ detailed in the appendix, compute

$$\Theta^b = \sup_{t_{\min} \leqslant t \leqslant t_{\max}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \omega_i^b \frac{\widehat{\mathrm{IF}}\left(\widetilde{T}_i, \widetilde{\eta}_i, M_i, t\right)}{\widehat{\sigma}_t} \right|$$

Step 3: Repeat steps 1 and 2 a large number of times, say $B = 2000$ times for instance, and compute $\widehat{q}_{1-\alpha}$ as the $100(1-\alpha)$th percentile of $\left\{\Theta^1, \ldots, \Theta^B\right\}$.

Similarly, we are able to compute confidence band for the curve $\left\{\left(t, AUC_{M_1}(t) - AUC_{M_2}(t)\right), t \in [t_{\min}, t_{\max}]\right\}$ replacing $\widehat{AUC}(t)$ by $\widehat{AUC}_{M_1}(t) - \widehat{AUC}_{M_2}(t)$, and $\frac{\widehat{\mathrm{IF}}(\widetilde{T}_i, \widetilde{\eta}_i, M_i, t)}{\widehat{\sigma}_t}$ by $\frac{\widehat{\mathrm{IF}}(\widetilde{T}_i, \widetilde{\eta}_i, M_{1i}, t) - \widehat{\mathrm{IF}}(\widetilde{T}_i, \widetilde{\eta}_i, M_{2i}, t)}{\widehat{\sigma}_{t12}}$ in the previous expressions. This latter confidence band is of particular interest for testing by observing whether or not the zero function is contained within the band.

3.8. Inference for points on receiver operating characteristic curve

Although we mainly focus on AUC in this paper, inference procedures for points on ROC curve are also desirable. In particular, CIs for sensitivity and specificity at a clinically relevant threshold c are useful. We therefore provide lemma 3 in the web Supporting Material C that gives large sample results for sensitivity and specificity estimators, similarly to lemma 1 for the $AUC(t)$ estimator. Then, CI or other confidence regions may be derived using the approach described in Section 3.2.

4. Simulation studies

We conducted numerical investigations to assess the performances of the proposed estimators and of the test of comparison.

4.1. Data generation

We generated two rival markers M_1 and M_2 through standard normal distribution $\mathcal{N}(0, 1)$ with correlation 0.5. We generated two competing events with proportional cause-specific hazards models

$$\begin{aligned}\alpha_j(t) &= \lim_{dt \to 0} P\left(T \in [t, t+dt), \eta = j \,|\, T \geqslant t\right)/dt \\ &= \alpha_{0j} \exp\left(\beta_{j1} M_1 + \beta_{j2} M_2\right),\end{aligned}$$

for $j = 1, 2$, following the algorithm described in [25]. We generated the censoring time with a Cox proportional hazards model $\lambda_C(t) = \lambda_{0C} \exp\left(\gamma_1 M_1 + \gamma_2 M_2\right)$. We considered several scenarios: two sample sizes $n = 200$ or $n = 400$, a censoring that does not depend on any markers ($HR_{C_1} = \exp(\gamma_1) = 1$

and $HR_{C_2} = \exp(\gamma_2) = 1$) or that depends on the marker M_1 only with hazard ratio $HR_{C_1} = 1.35$ (and $HR_{C_2} = 1$), and four values 0, 0.22, 0.45 or 0.68 for parameter Δ_{β_1}, where $\beta_{11} = 1 + \Delta_{\beta_1}$ and $\beta_{12} = 1 - \Delta_{\beta_1}$, to make the differences of AUCs of M_1 and M_2 at time $t = 1$ equal to 0, 0.05, 0.10 or 0.15. Both markers were also associated with the competing event ($\beta_{21} = \beta_{22} = 0.2$). To more thoroughly investigate the behaviour of the test when censoring depends on the markers, we also performed simulations under the null hypothesis where $\left(HR_{C_1}, HR_{C_2}\right) = (2, 1)$ and $(1.35, 0.6)$. Other parameters were chosen to have approximately 33%, 17% and 38% of *observed cases, controls* and controls* (as defined in Section 3.1) and 28% of censored subjects at time $t = 1$. These frequencies can be 1% to 3% greater or smaller depending on the scenario generated.
The choice of the hazard ratio $HR_{C_1} = 1.35$, the correlation between M_1 and M_2 equal to 0.5 and the proportion of censored subjects equal to 27.5% was guided by the PAQUID cohort data.

For each data set, we computed the IPCW estimators of $AUC(t)$ and $AUC^*(t)$ for the first event-type, the 95%-level CIs and the tests of comparison with α-level equal to 5%. We used the two weighting procedures: assuming C is independent of (T, η, M_1, M_2), estimating $G(t)$ by Kaplan–Meier (KM-weights), or relaxing this assumption, assuming that C is independent of (T, η) given (M_1, M_2), estimating $G\left(t|M_1, M_2\right)$ by a Cox model (Cox-weights). We also computed the nonparametric nearest neighbour estimators (NNE) of Saha and Heagerty [5] for comparison with our approach. As no optimal rule exists, the choice of a bandwidth $\lambda_n = O\left(n^{-1/3}\right)$ required for NNE was set equal to $\lambda_n = 0.33 \times n^{-1/3}$, that leads to approximately $2\lambda_n = 11.3\%$ and 8.9% of subjects included in each neighbourhood, when $n = 200$ and $n = 400$ respectively.

The standard error estimators based on the influence function defined at equations (11) and (12) were computed for the KM-weights estimator, and for all estimators, standard errors were also computed by bootstrapping (400 replications).

4.2. Simulation results

Tables I–III summarize the main simulation results on the basis of 1000 replications for each scenario including $n = 400$ subjects. Similar tables of results with $n = 200$ are provided in the web Supporting Information D. First, with the KM-weights, the bootstrap and influence function estimators of the standard errors lead to very similar results.

When the censoring does not depend on the markers (Table I), estimators based on both weightings perform as well. As expected, they are unbiased, and the coverage probabilities of the CIs are close to the nominal value 95%. Under the null hypothesis that AUCs are equal for both markers, estimated α-level is close to the nominal value 5%, and the power of the test tends to one when the difference between AUCs increases. The two weighting procedures also lead to similar results in terms of efficiency.

When the censoring depends moderately on the marker M_1 (Table II), the AUC estimates for M_1 are slightly biased, and the coverage probabilities of their CIs decrease when C is assumed independent of M_1 (KM-weights). Because of the positive correlation between the two markers, the bias for the KM-estimates of AUC for marker M_2 is in the same direction but much smaller (with no impact on the coverage rates). As a consequence, the test of comparison remains valid in these simulation scenarios even if the weights are not well adapted to the censoring.

However, Table III shows that when the dependency between C and M_1 is stronger ($HR_{C_1} = 2$, $HR_{C_2} = 1$), or when there is a positive association between C and M_1 and a negative association between C and M_2 ($HR_{C_1} = 1.35$, $HR_{C_2} = 0.6$), the test is biased. AUC estimators and all inference procedures perform still very well in any cases when weights are computed from a Cox model.

The NNE performs also well for independent censoring with only slight underestimations, probably because of smoothing (Table II). With moderate marker-dependent censoring (Table II), the bias for NNE increases little, and this has no consequences on the validity of inference procedures even when censoring depends on marker M_1, whereas the ROC curve is estimated for M_2, setting which is not handled by NNE. The main reason is that slight underestimation due to smoothing compensates slight overestimation due to marker-dependent censoring. Even when the association between censoring and marker becomes stronger, or when censoring is associated with both markers with opposite directions, NNE appears quite robust (Table III).

Finally, these simulations illustrate the difference between the two definitions of the specificity. Defining all subjects who met the competing event as controls decreases the AUC and increases the power of the test of comparison. The decrease of AUC is due to the positive dependence of both types of event on the markers. Indeed, discrimination between subjects who will experience the main event and subjects

Table I. Simulation results with sample size $n = 400$ and independent censoring, that is, $\mathrm{HR}_{C_1}=\mathrm{HR}_{C_2}=1$ (1000 replications).

			AUC(t)				Coverage probability					
			True		Bias		M_1		M_2		Type I error or power	
	Method	ΔAUC(t)	M_1	M_2	M_1	M_2	IF	B	IF	B	IF	B
$AUC^*(t)$	IPCW KM	0.0	85.6	85.6	0.0	0.1	93.2	93.3	94.5	94.7	5.0	4.8
	IPCW Cox				0.0	0.1	—	94.0	—	95.5	—	5.0
	NNE				−0.3	−0.2	—	95.0	—	95.2	—	4.7
$AUC(t)$	IPCW KM	0.0	79.8	79.8	0.0	0.0	95.7	95.7	95.1	94.9	4.7	4.6
	IPCW Cox				0.0	0.0	—	95.1	—	94.9	—	4.4
	NNE				−0.3	−0.3	—	94.4	—	95.3	—	4.1
$AUC^*(t)$	IPCW KM	5.0	87.9	82.9	0.1	0.1	93.7	93.8	95.1	95.6	26.7	25.2
	IPCW Cox				0.1	0.0	—	93.3	—	95.9	—	26.7
	NNE				−0.2	−0.2	—	94.6	—	95.6	—	30.2
$AUC(t)$	IPCW KM	5.1	82.2	77.1	0.1	0.0	94.4	94.4	96.8	96.5	34.8	34.3
	IPCW Cox				0.1	0.0	—	94.8	—	96.2	—	35.7
	NNE				−0.2	−0.2	—	94.9	—	95.4	—	36.0
$AUC^*(t)$	IPCW KM	10.0	89.9	79.9	0.1	0.1	93.5	93.7	94.8	94.8	77.7	78.1
	IPCW Cox				0.1	0.1	—	94.4	—	95.2	—	79.3
	NNE				−0.2	−0.2	—	94.5	—	95.3	—	82.4
$AUC(t)$	IPCW KM	10.3	84.4	74.1	0.1	0.1	94.9	94.9	95.9	95.8	89.8	90.0
	IPCW Cox				0.1	0.1	—	94.8	—	95.9	—	90.8
	NNE				−0.2	−0.2	—	95.5	—	95.2	—	90.3
$AUC^*(t)$	IPCW KM	14.7	91.4	76.8	0.1	0.1	94.3	94.3	94.8	95.1	98.4	98.4
	IPCW Cox				0.1	0.1	—	94.7	—	96.0	—	98.6
	NNE				−0.2	−0.2	—	94.7	—	95.3	—	99.3
$AUC(t)$	IPCW KM	15.1	86.0	71.0	0.1	0.0	94.0	94.6	95.3	95.8	99.7	99.8
	IPCW Cox				0.1	0.0	—	94.4	—	96.3	—	99.7
	NNE				−0.2	−0.2	—	94.5	—	95.8	—	99.7

Bias of estimates for two markers M_1 and M_2 (multiplied by 100), empirical coverage probabilities of 95% confidence intervals and type I errors and powers of the test of $\mathcal{H}_0 : \mathrm{AUC}_{M_1}(t) = \mathrm{AUC}_{M_2}(t)$, depending on $\Delta\mathrm{AUC}(t) = \mathrm{AUC}_{M_1}(t) - \mathrm{AUC}_{M_2}(t)$. Inverse probability of censoring weighting (IPCW) estimators using the Kaplan–Meier estimator (IPCW KM) or a Cox proportional hazards model with covariates M_1 and M_2 (IPCW Cox) for weighting, and the nearest neighbour estimator (NNE). Variances are computed from the estimated influence function (IF) for IPCW KM, or by bootstrapping 400 times (B).

who will undergo the competing event is more difficult than discrimination between subjects who will undergo the main event and subjects who will experience none of the event. As it will be illustrated in Section 5, it is frequent that both events are associated with the markers. The increase of the power is because subjects who met the competing event are more informative for estimating $AUC(t)$, for which they are controls, than for estimating $AUC^*(t)$ for which they are only used for estimating the weights.

The good behaviour of the inference procedures for points on the ROC curve discussed in Subsection 3.8 was also assessed, and some simulation results are provided in the web Supporting Information Table 4.

5. Application to dementia prediction

5.1. Objective

The objective of this analysis is the estimation and the comparison of the abilities of two cognitive tests to predict the risk of dementia within the 3, 5 and 10 years following the test result, in the elderly population accounting for death without dementia competing risk. A window of prediction of 3 or 5 years could correspond to the duration of a preventive clinical trial, and thus after validation, these cognitive tests could be used to select the population at risk to include in a trial. A window of 10 years is probably too long for a preventive trial but may be of interest for general practitioner for reassuring worried patients. The two tests compared are the mini-mental state examination (MMSE) [26] and the

Table II. Simulation results with sample size $n = 400$ and moderate M_1-dependent censoring, that is, HR_{C_1}=1.35, HR_{C_2}=1 (1000 replications).

			AUC(t)				Coverage probability					
			True		Bias		M_1		M_2		Type I error or power	
	Method	ΔAUC(t)	M_1	M_2	M_1	M_2	IF	B	IF	B	IF	B
$AUC^*(t)$	IPCW KM	0.0	85.6	85.6	1.5	0.6	87.6	87.8	93.7	93.7	5.2	5.6
	IPCW Cox				0.0	0.2	—	94.1	—	95.0	—	4.8
	NNE				−0.5	−0.2	—	94.2	—	95.9	—	5.5
$AUC(t)$	IPCW KM	0.0	79.8	79.8	1.1	0.6	92.0	92.0	92.9	93.0	5.2	5.4
	IPCW Cox				0.0	0.1	—	95.6	—	94.3	—	5.4
	NNE				−0.4	0.0	—	94.9	—	94.3	—	5.4
$AUC^*(t)$	IPCW KM	5.0	87.9	82.9	1.3	0.6	86.2	86.0	94.1	94.4	36.7	36.2
	IPCW Cox				0.1	0.1	—	93.4	—	95.8	—	26.5
	NNE				−0.4	−0.2	—	94.3	—	95.5	—	29.5
$AUC(t)$	IPCW KM	5.1	82.2	77.1	1.0	0.7	91.5	91.2	94.6	94.3	41.2	41.2
	IPCW Cox				0.1	0.0	—	94.2	—	95.6	—	33.3
	NNE				−0.3	0.0	—	94.9	—	95.5	—	33.8
$AUC^*(t)$	IPCW KM	10.0	89.9	79.9	1.1	0.8	88.0	87.9	94.1	93.7	86.7	86.2
	IPCW Cox				0.1	0.1	—	94.2	—	95.3	—	78.8
	NNE				−0.4	−0.2	—	95.5	—	95.3	—	81.8
$AUC(t)$	IPCW KM	10.3	84.4	74.1	0.9	0.8	92.1	92.1	94.5	94.6	92.7	92.5
	IPCW Cox				0.1	0.1	—	95.4	—	95.6	—	89.5
	NNE				−0.3	0.0	—	94.9	—	95.1	—	90.3
$AUC^*(t)$	IPCW KM	14.7	91.4	76.8	0.9	0.9	87.8	88.5	93.4	93.0	99.5	99.4
	IPCW Cox				0.1	0.1	—	93.7	—	94.8	—	99.0
	NNE				−0.4	−0.2	—	95.5	—	94.9	—	99.0
$AUC(t)$	IPCW KM	15.1	86.0	71.0	0.8	0.8	91.2	90.9	93.8	93.6	99.9	99.8
	IPCW Cox				0.1	0.0	—	94.6	—	95.0	—	99.8
	NNE				−0.2	0.0	—	94.9	—	95.0	—	99.8

Bias of estimates for two markers M_1 and M_2 (multiplied by 100), empirical coverage probabilities of 95% confidence intervals and type I errors and powers of the test of $\mathcal{H}_0 : \mathrm{AUC}_{M_1}(t) = \mathrm{AUC}_{M_2}(t)$, depending on $\Delta\mathrm{AUC}(t) = \mathrm{AUC}_{M_1}(t) - \mathrm{AUC}_{M_2}(t)$. Inverse probability of censoring weighting (IPCW) estimators using the Kaplan–Meier estimator (IPCW KM) or a Cox proportional hazards model with covariates M_1 and M_2 (IPCW Cox) for weighting, and the nearest neighbour estimator (NNE). Variances are computed from the estimated influence function (IF) for IPCW KM, or by bootstrapping 400 times (B).

digit symbol substitution test (DSST) [27]. The MMSE is a sum-score evaluating various dimensions of cognition (memory, calculation, orientation in space and time, language and word recognition), which is often used as an index of global cognitive performance and for screening of dementia. MMSE score ranges from 0 to 30. The DSST explores attention and psychomotor speed. Given a code table displaying the correspondence between pairs of digits and symbols, the subjects have to fill in blank squares with the symbol, which is paired to the digit displayed above the square. The subjects have to fill in as many squares as possible in 90 s. The maximum value is 90.

As larger values of DSST and MMSE are associated with lower risks of dementia, let us note that, following the previous notations, ROC analyses were performed for minus DSST and minus MMSE to reverse the associations.

5.2. *The PAQUID data*

Paquid is a population-based study including 3777 subjects aged 65 years and older, living at home in the south-west of France at enrollment in 1988. Individuals were seen at home by psychologists trained in home interviews at the initial visit and at 1, 3, 5, 8, 10, 13, 15, 17 and 20 years later. Each visit included a neuropsychological evaluation through a battery of psychometric tests and a standardized diagnosis of dementia [28].

We used MMSE and DSST collected at baseline to predict dementia diagnosis over the first 3, 5 or 10 years of follow-up. We excluded from the sample subjects demented, blind, deaf or confined to bed

Table III. Simulation results with sample size $n = 400$ and strong dependent censoring (1000 replications).

			AUC(t)				Coverage probability					
			True		Bias		M_1		M_2		Type I error or power	
	Method	ΔAUC(t)	M_1	M_2	M_1	M_2	IF	B	IF	B	IF	B
HR$_{C_1}$=2, HR$_{C_2}$=1												
$AUC^*(t)$	IPCW KM	0.0	85.6	85.6	2.6	0.8	74.1	74.8	91.8	91.6	9.2	9.3
	IPCW Cox				0.1	0.2	—	94.5	—	94.2	—	5.0
	NNE				−1.0	−0.3	—	94.2	—	95.3	—	5.8
$AUC(t)$	IPCW KM	0.0	79.8	79.8	2.0	1.0	84.4	84.3	90.3	90.5	6.7	6.6
	IPCW Cox				0.0	0.1	—	95.5	—	93.8	—	5.6
	NNE				−0.6	0.0	—	95.3	—	93.9	—	6.1
HR$_{C_1}$=1.35, HR$_{C_2}$=0.6												
$AUC^*(t)$	IPCW KM	0.0	85.6	85.6	1.2	−2.5	90.3	90.7	91.7	91.5	15.8	14.9
	IPCW Cox				0.1	0.1	—	92.5	—	94.2	—	4.8
	NNE				−0.2	0.2	—	94.8	—	94.6	—	5.6
$AUC(t)$	IPCW KM	0.0	79.8	79.8	0.4	−1.7	94.0	94.1	92.5	92.3	9.7	9.9
	IPCW Cox				0.0	0.0	—	94.9	—	94.0	—	5.1
	NNE				−0.5	0.3	—	94.7	—	93.6	—	6.2

Bias of estimates for two markers M_1 and M_2 (multiplied by 100), empirical coverage probabilities of 95% confidence intervals and type I errors and powers of the test of $\mathcal{H}_0 : \text{AUC}_{M_1}(t) = \text{AUC}_{M_2}(t)$, depending on $\Delta\text{AUC}(t) = \text{AUC}_{M_1}(t) - \text{AUC}_{M_2}(t)$. Inverse probability of censoring weighting (IPCW) estimators using the Kaplan–Meier estimator (IPCW KM) or a Cox proportional hazards model with covariates M_1 and M_2 (IPCW Cox) for weighting, and the nearest neighbour estimator (NNE). Variances are computed from the estimated influence function (IF) for IPCW KM, or by bootstrapping 400 times (B).

Table IV. Comparison of $AUC(t)$ of digit symbol substitution test and mini-mental state examination at times $t = 3$, 5 and 10 years.

		$t = 3$	$t = 5$	$t = 10$
Demented		70(2.7%)	122(4.8%)	318(12.4%)
Died without dementia		194(7.6%)	313(12.2%)	545(21.3%)
Censored		180(7.0%)	292(11.4%)	591(23.1%)
KM weights				
$AUC(t)$	DSST	79.9 [74.9,84.9]	77.8 [74.0,81.6]	72.2 [69.2,75.2]
	MMSE	74.7 [68.7,80.8]	72.0 [67.2,76.8]	66.9 [63.6,70.2]
	p-value	0.03	0.01	< 0.01
	Adjusted p-value	0.09	0.02	< 0.01
$AUC^*(t)$	DSST	80.9 [76.0,85.8]	79.7 [76.0,83.5]	76.7 [73.8,79.7]
	MMSE	75.4 [69.4,81.4]	73.2 [68.4,78.0]	69.9 [66.6,73.3]
	p-value	0.02	< 0.01	< 0.01
	Adjusted p-value	0.06	0.01	< 0.01
Cox weights				
$AUC(t)$	DSST	79.8 [74.8,84.8]	77.5 [73.7,81.4]	71.7 [68.8,74.7]
	MMSE	74.7 [68.6,80.7]	71.9 [67.0,76.7]	66.6 [63.3,69.9]
	p-value	0.04	0.01	< 0.01
	Adjusted p-value	0.10	0.03	< 0.01
$AUC^*(t)$	DSST	80.8 [75.9,85.7]	79.4 [75.6,83.2]	76.1 [73.2,79.1]
	MMSE	75.2 [69.3,81.4]	73.0 [68.2,77.8]	69.6 [66.2,73.0]
	p-value	0.03	< 0.01	< 0.01
	Adjusted p-value	0.07	0.01	< 0.01

Estimates (multiplied by 100), confidence intervals and p-values of the test of $\mathcal{H}_0 : \text{AUC}_{DSST}(t) = \text{AUC}_{MMSE}(t)$. PAQUID cohort, $n = 2561$.
DSST, digit symbol substitution test; MMSE, mini-mental state examination; AUC, area under the receiver operating characterisitic; KM weights, Kaplan–Meier weights.

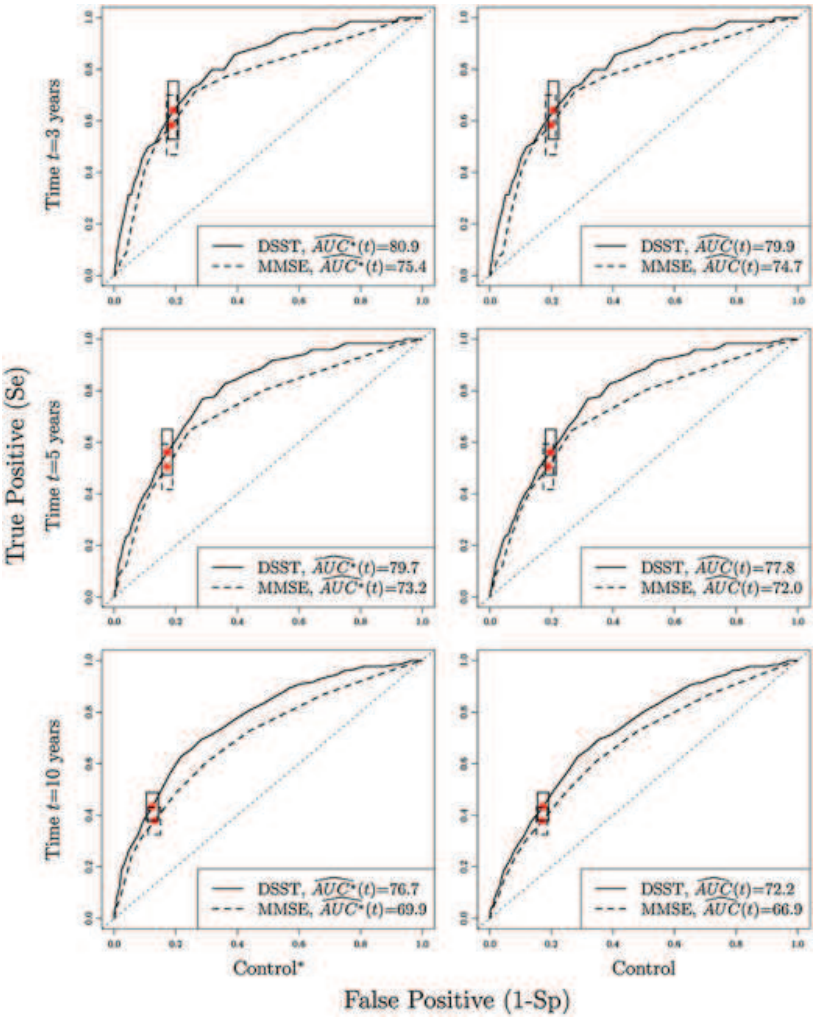


Figure 1. Time-dependent receiver operating characteristic curves estimated by inverse probability of censoring weighting estimators (KM-weights) at $t = 3, 5$ and 10 years for digit symbol substitution test and mini-mental state examination with the two definitions of the specificity. PAQUID, $n = 2561$ subjects. Points and rectangles display estimates and 95% confidence intervals for sensitivity and one minus specificity estimates for a threshold value equal to the third quartile of the marker distribution (19 for digit symbol substitution test, 25 for mini-mental state examination).

at the initial visit, subjects with missing values for the MMSE and the DSST and subjects who dropped out just after the baseline visit. The final sample included $n = 2561$ subjects. The time-to-dementia was computed as the mid-point between the time of the visit of diagnosis and the time of the last visit without dementia. Subjects who died without a dementia diagnosis were considered as free of dementia at their death if the last visit was less than 2 years before the death and were considered as censored at the last visit if duration between the last visit and death was longer. Using a threshold of 3 years instead of 2 years led to the same conclusions (results not shown).

5.3. *Results*

The mean DSST score was 27.4 (standard deviation: 11.5, inter quartile range: 19–35), and the mean MMSE was 26.4 (standard deviation: 2.9, inter quartile range: 25–28). The correlation between the two markers was estimated at 0.55.

We fitted a multivariate Cox model in order to explore the dependence of the censoring time on both psychometric tests. The estimated hazard ratios were 1.019 for DDST (95% CI: (1.010, 1.028)) and 1.023 for MMSE (95% CI :(0.989, 1.057)). As in our simulation study, only one marker is significantly

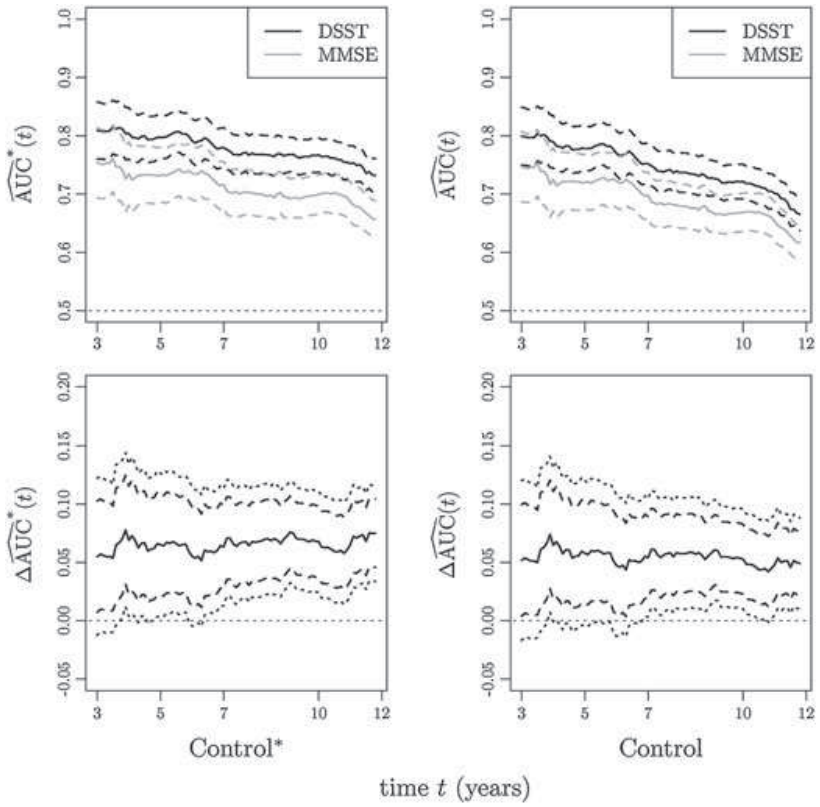


Figure 2. Comparisons of the evolutions of $\widehat{AUC}(t)$ and $\widehat{AUC}^*(t)$ over time for digit symbol substitution test and mini-mental state examination. PAQUID cohort, $n = 2561$. $\Delta\widehat{AUC}(t)$ and $\Delta\widehat{AUC}^*(t)$ mean the differences of estimates between digit symbol substitution test and mini-mental state examination. Dashed and dotted lines display respectively 95% pointwise confidence intervals and 95% simultaneous confidence bands.

associated with the censoring time. The hazard ratio for the increase of one standard error of the DSST was 1.25, smaller than the one in the simulation study.

The frequencies of *observed cases, controls* and controls* (as defined in Section 3.1) and censored subjects before each time point t, for $t = 3, 5, 10$ years, are given in Table IV. Although the proportions of *observed cases* at time $t = 3, 5, 10$ are small, respectively 2.7%, 4.8% and 12.4%, the frequencies are big enough, respectively 70, 122 and 328 because of the large sample size ($n = 2561$). Respectively, 7.0%, 11.4% and 23.1% of subjects are lost of follow-up at time 3, 5 and 10 years.

Figure 1 displays the ROC curves estimated at time $t = 3, 5, 10$ years for the two definitions of specificity, assuming independence between censoring and marker (KM-weights). For both definitions, the ROC curves go down and become closer to the diagonal as time increases, confirming that the discrimination decreases as the window of prediction enlarges. This is illustrated on Figure 2 that displays the estimates of $AUC(t)$ and $AUC^*(t)$ for time t from 3 to 12 years and their 95% pointwise CIs. With both definition of controls, Figure 1 shows that the discrimination ability of DSST appears always better than the one of MMSE, and more specifically, the three ROC curves of DSST are always above the ROC curves of MMSE. Consequently, whatever the time $t = 3, 5, 10$ years and whatever the cutpoint chosen for the marker (or equivalently the value of the specificity), the sensitivity is always estimated greater for DSST. As an illustration, Figure 1 also depicts CIs for sensitivity and specificity estimates, for a threshold value equal to the third quartile of the marker distribution. Table IV displays AUC estimates with 95% CI and tests for comparison between MMSE and DSST for the two definitions of controls and the two weightings (with KM-weights, estimated influence functions were used to compute the variances; with Cox-weights, variances were estimated by bootstrapping 5000 times). According to the unadjusted p-values, the differences between the AUC were significant for the three windows of prediction; adjusting for multiple testing (one test for each time point), only the differences at 5 and 10 years were still significant. The 95% pointwise CIs of the differences between AUCs of MMSE and DSST displayed in

Figure 2 show that the differences are significant for all times t from 3 to 12 years. However, the 95% simultaneous confidence bands of these differences intersect the zero line. As the association between censoring and marker is weak, results obtained with the Cox-weights and with the NNE of Saha and Heagerty [5] were nearly identical, as expected (results not shown).

Figure 2 and Table IV also show that the discrimination is better when subjects died without dementia are not defined as controls ($AUC^*(t)$ versus $AUC(t)$). However, the differences between DSST and MMSE are similar for both definitions, and the test results are concordant. The great difference between $AUC^*(t)$ and $AUC(t)$ is due to the association between the two markers and the competing event: the hazard ratio for the association between DSST or MMSE and death without dementia are 1.037 for DSST (95% CI : $(1.030, 1.044)$) and 1.080 for MMSE (95% CI : $(1.055, 1.107)$). Thus, it is easier to discriminate future demented subjects from subjects alive and non-demented rather than from subjects alive and non-demented or died without dementia.

To conclude, although MMSE is largely used as a screening test for dementia, all these analyses found that the DSST has better performances to detect future demented subjects in a period of time of 3, 5 or 10 years compared with the MMSE. Moreover, whatever the specificity, the sensitivity of DSST is always estimated better than the one of MMSE.

6. Discussion

In this manuscript, we used the IPCW approach to estimate time-dependent ROC curve and AUC for censored events with competing risks. Large sample theory of the estimators was established, and CIs, simultaneous confidence band and tests of comparison were derived. Simulation results show that the proposed procedures work well for moderate sample sizes. The practical interest of the procedure for estimating and comparing psychometric tests for dementia onset was illustrated with data from the PAQUID cohort. Significant results suggest that the DSST is better than the MMSE to discriminate subjects at high risk of Alzheimer's disease in the next few years.

We proposed estimators for the two definitions of specificity because both definitions may be useful depending the clinical setting. For instance, if the prognostic marker is used as inclusion criteria in a clinical trial for Alzheimer's disease, it is essential to be able to discriminate subjects at high risk of dementia from all subjects and especially from subjects who have a high risk of death without dementia during the trial duration. Similarly, in prostate cancer treatment of very old patients, the main criteria for treatment is the risk of progression of the disease before death from other causes. Thus, for both setting the marker must be validated with $AUC(t)$ defined at equation (4). On the other hand, from the point of view of a patient worried about his cognitive decline, the main question of interest may be : If I survive 10 years, am I at high risk of becoming demented during this period? Then, $AUC^*(t)$ defined at equation (3) is more relevant. In any case, as the estimated value may be very different, it is essential in the clinical application to clearly state which definition is used for the specificity.

As pointed out by one referee, our analyses of the predictiveness of the cognitive tests on the PAQUID data do not account for the age of the subjects. Consequently, the predictiveness of the cognitive tests has to be carefully interpreted as a marginal predictiveness among the population of subjects aged 65 years and older enrolled in the PAQUID study. Thus, the estimated predictive accuracy is of interest for clinicians who are interested in quantifying how informative can be the result of a cognitive test to evaluate the risk of dementia of a patient in the t-years following a test result.

To account for age in the estimation of the ROC curves and AUCs, we could have fitted semiparametric models for ROC curves or AUC given age, using semiparametric methods proposed by Zheng *et al.* [6] or adapting a method of Hung and Chiang [10]. Alternatively, our nonparametric methods could also have been computed on several age groups, performing stratified analyses. Another possible approach would have been to fit a multivariate prediction model accounting for age, using a binomial regression model [29] for example, to then estimate the ROC curve and AUC of the resulting prediction tool. For this latter approach, a careful correction for overfitting and optimism bias is required [30], which is outside the scope of this manuscript. Kerr and Pepe [31] have recently discussed and contrasted these two strategies in detail.

To deal with interval censoring of dementia in the PAQUID data, we used a simple imputation rule of the status of subjects deceased without dementia diagnosis. Changing this rule (3 years instead of 2 years between the last visit and the death) did not change the results. To our knowledge, only parametric estimators based on an illness-death model can account for this particular kind of interval-censoring [32]. In an ongoing work, we developed such model-based estimators. However, simulation results suggest

that the bias for the IPCW estimators are small and may be smaller than the bias of the model-based estimators when the model is misspecified.

Compared with earlier papers that focused on estimation, this work mainly focused on tests for comparing AUCs. The test of comparison that we proposed is the extension to the competing risks setting with censored data of the popular test proposed by DeLong *et al.* [8]. Indeed, without censoring, weights are all equal to one.

Assuming that censoring does not depend on the markers, the estimation and the test do not require any parametric assumptions by contrast to the approach of Foucher *et al.* [13] or Zheng *et al.* [6], nor bandwidth selection due to kernel smoothing as in Saha and Heagerty [5] or as in the smooth method of Zheng *et al.* [6].

Moreover, nonparametric approaches previously proposed cannot properly deal with censoring depending on several markers. Indeed, with the method of Saha and Heagerty [5] or the smooth method of Zheng *et al.* [6], the cause specific hazards are estimated only conditionally on the marker under study, by non-parametric kernel-based methods. Therefore, dependence between censoring and other markers are not taken into account, whereas when modelling hazard in survival analysis, the assumption is that the censoring mechanism can only depend on covariates included in the model [33, Section 2.2.8]. Although they could be extended to deal with censoring depending on several markers, using multidimensional kernels, their practical interest would be limited by the so-called curse of dimensionality phenomenon [34]. To our mind, they are therefore not suitable for comparing several markers and making tests with marker dependent censoring, even though they appeared robust in our realistic simulation scenarii with two correlated markers and moderate associations between censoring and markers. By contrast, when censoring depend on the markers, the IPCW approach enables such tests of comparison under the additional assumption that the censoring mechanism is well specified. Nevertheless, when the dependence between censoring and markers is light, our simulations suggest that the estimator using KM-weights can be robust.

The proposed methodology is implemented in the `timeROC` package, written in R [35], that is publicly available on the Comprehensive R Archive Network (CRAN) site. The PAQUID data presented in Section 5 is attached to the package, and most of the analyses presented in Section 5 are presented as examples, making them easily reproducible.

Appendix A

Definition of IF$(\cdot)$

Let $M_{C_i}(t) = \mathbb{1}_{(\widetilde{\eta}_i=0,\widetilde{T}_i\leqslant t)} - \int_0^t \mathbb{1}_{(\widetilde{T}_i\geqslant t)} d\Lambda_C(u)$, where $\Lambda_C(\cdot)$ is the cumulative hazard function of the censoring variable C, $h_{tij,1} = \frac{\mathbb{1}_{(\widetilde{T}_i\leqslant t,\widetilde{\eta}_i=1)}\mathbb{1}_{(\widetilde{T}_j>t)}}{G(\widetilde{T}_i)G(t)}\mathbb{1}_{(M_i>M_j)}$, $h_{tij,2} = \frac{\mathbb{1}_{(\widetilde{T}_i\leqslant t,\widetilde{\eta}_i=1)}\mathbb{1}_{(\widetilde{T}_j\leqslant t,\widetilde{\eta}_j\notin\{0,1\})}}{G(\widetilde{T}_i)G(\widetilde{T}_j)}\mathbb{1}_{(M_i>M_j)}$, $f_{i1t} = \frac{\mathbb{1}_{(\widetilde{T}_i\leqslant t,\widetilde{\eta}_i=1)}}{G(\widetilde{T}_i)}$ and $h_t = \mathbb{E}\left[h_{tij,1} + h_{tij,2}\right]$. Then,

$$\mathrm{IF}\left(\widetilde{T}_i, \widetilde{\eta}_i, M_i, t\right) = \frac{\mathbb{E}\left(\Psi_{ijkl}(t) + \Psi_{jikl}(t) + \Psi_{jkil}(t) + \Psi_{jkli}(t) \middle| \left(\widetilde{T}_i, \widetilde{\eta}_i, M_i\right)\right)}{F_1(t)\left(1 - F_1(t)\right)}$$

where

$$\begin{aligned}\Psi_{ijkl}(t) =& h_{tij,1}\left(1 + \int_0^{\widetilde{T}_i} \frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right)\left(1 + \int_0^t \frac{dM_{C_l}(u)}{S_{\widetilde{T}}(u)}\right)\\ &+ h_{tij,2}\left(1 + \int_0^{\widetilde{T}_i} \frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right)\left(1 + \int_0^{\widetilde{T}_j} \frac{dM_{C_l}(u)}{S_{\widetilde{T}}(u)}\right)\\ &- h_t - \frac{h_t\left(1 - 2F_1(t)\right)}{F_1(t)\left(1 - F_1(t)\right)}\left[f_{i1t}\left(1 + \int_0^{\widetilde{T}_i} \frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right) - F_1(t)\right]\end{aligned}$$

Estimator $\widehat{IF}(\cdot)$

Let $\widehat{M}_C(\cdot)$ be the estimator defined by plugging in the usual Nelson–Aalen estimator of the cumulative incidence function of the censoring C and $\widehat{h}_{tij,1}$, $\widehat{h}_{tij,2}$ and $\widehat{f}_{i1t}$ be defined by plugging in the

Kaplan–Meier estimator $\widehat{G}(\cdot)$. Let $\widehat{h}_t = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\widehat{h}_{tij,1} + \widehat{h}_{tij,2}$. Then,

$$\widehat{\mathrm{IF}}\left(\widetilde{T}_i, \widetilde{\eta}_i, M_i, t\right) = \frac{\frac{1}{n(n-1)(n-2)}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\widehat{\Psi}_{ijkl}(t) + \widehat{\Psi}_{jikl}(t) + \widehat{\Psi}_{jkil}(t) + \widehat{\Psi}_{jkli}(t)}{\widehat{F}_1(t)\left(1 - \widehat{F}_1(t)\right)}$$

where $\widehat{\Psi}_{ijkl}(t)$ is defined plugging in the estimators $\widehat{h}_t$, $\widehat{h}_{tij,1}$, $\widehat{h}_{tij,2}$, $\widehat{f}_{i1t}$, $\widehat{M}_C(\cdot)$, $\widehat{F}_1(t)$ and $\widehat{S}_{\widetilde{T}}(\cdot)$.

Supporting Information

Web Appendix referenced in Sections 3.2, 3.8 and 4.2 is available with this paper at the Statistics in Medicine website on Wiley Online Library. It contains details about large sample results and additional simulation results.

Acknowledgements

The MCIA (Mésocentre de Calcul Intensif Aquitain) of the Université de Bordeaux and of the Université de Pau et des Pays de l'Adour provided computer time for this study with its computing facilities. France Alzheimer partly funded this work by a grant awarded to Hélène Jacqmin-Gadda in 2009. IPSEN and Novartis laboratories funded the PAQUID study, managed by Jean-François Dartigues. We are also grateful to the associate editor and two anonymous referees for their helpful comments and suggestions.

References

1. Amieva H, Jacqmin-Gadda H, Orgogozo J, Le Carret N, Helmer C, Letenneur L, Barberger-Gateau P, Fabrigoule C, Dartigues J. The 9-year cognitive decline before dementia of the Alzheimer type: a prospective population-based study. *Brain* 2005; **128**(5):1093. DOI: 10.1093/brain/awh451.
2. Aisen P, Andrieu S, Sampaio C, Carrillo M, Khachaturian Z, Dubois B, Feldman H, Petersen R, Siemers E, Doody R, *et al.* Report of the task force on designing clinical trials in early (predementia) AD. *Neurology* 2011; **76**(3):280–286. DOI: 10.1212/WNL.0b013e318207b1b9.
3. Heagerty P, Lumley T, Pepe M. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000; **56**(2):337–344. DOI: 10.1111/j.0006-341X.2000.00337.x.
4. Heagerty P, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* 2005; **61**(1):92–105. DOI: 10.1111/j.0006-341X.2005.030814.x.
5. Saha P, Heagerty P. Time-dependent predictive accuracy in the presence of competing risks. *Biometrics* 2010; **66**(4):999–1011. DOI: 10.1111/j.1541-0420.2009.01375.x.
6. Zheng Y, Cai T, Jin Y, Feng Z. Evaluating prognostic accuracy of biomarkers under competing risk. *Biometrics* 2012; **68**(2):388–396. DOI: 10.1111/j.1541-0420.2011.01671.x.
7. Pepe M. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, 2003.
8. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves : a nonparametric approach. *Biometrics* 1988; **44**(3):837–845.
9. Uno H, Cai T, Tian L, Wei LJ. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association* 2007; **102**(478):527–537. DOI: 10.1198/016214507000000149.
10. Hung H, Chiang CT. Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics* 2010; **38**(1):8–26. DOI: 10.1002/cjs.10046.
11. Chiang C, Hung H. Non-parametric estimation for time-dependent AUC. *Journal of Statistical Planning and Inference* 2010; **140**(5):1162–1174. DOI: 10.1016/j.jspi.2009.10.012.
12. Blanche P, Dartigues JF, Jacqmin-Gadda H. Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal* 2013. DOI: 10.1002/bimj.201200045. in press.
13. Foucher Y, Giral M, Soulillou JP, Daures JP. Time-dependent ROC analysis for a three-class prognostic with application to kidney transplantation. *Statistics in Medicine* 2010; **29**(30):3079–3087. DOI: 10.1002/sim.4052.
14. Van der Laan MJ, Robins JM. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Verlag: New York, 2003.
15. Tsiatis AA. *Semiparametric Theory and Missing Data*. Springer Verlag: New York, 2006.
16. Jewell NP, Lei X, Ghani AC, Donnelly CA, Leung GM, Ho LM, Cowling BJ, Hedley AJ. Non-parametric estimation of the case fatality ratio with competing risks data: an application to severe acute respiratory syndrome (SARS). *Statistics in Medicine* 2007; **26**(9):1982–1998. DOI: 10.1002/sim.2691.
17. Antolini L, Biganzoli EM, Boracchi P. Crude cumulative incidence in the form of a Horvitz-Thompson like and Kaplan-Meier like estimator, (October 2006). COBRA Preprint Series. Working Paper 10. http://biostats.bepress.com/cobra/art10.
18. Satten GA, Datta S. The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician* 2001; **55**(3):207–210. DOI: 10.1198/000313001317098185.

19. Beran R. Nonparametric regression with randomly censored survival data. *Unpublished technical report*, University of California, Berkeley, 1981.
20. Akritas MG. Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics* 1994; **22**:1299–1327.
21. Martinussen T, Scheike TH. *Dynamic Regression Models for Survival Data*. Springer: New York, 2006.
22. Bretz F, Hothorn T, Westfall P. *Multiple Comparisons Using R*. CRC press, Boca Raton, 2010.
23. Cai T, Pepe MS. Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *Journal of the American Statistical Association* 2002; **97**(460):1099–1107. DOI: 10.1198/016214502388618915.
24. Lin DY, Fleming TR, Wei LJ. Confidence bands for survival curves under the proportional hazards model. *Biometrika* 1994; **81**(1):73–81.
25. Beyersmann J, Latouche A, Buchholz A, Schumacher M. Simulating competing risks data in survival analysis. *Statistics in Medicine* 2009; **28**(6):956–971. DOI: 10.1002/sim.3516.
26. Folstein MF, Folstein SE, McHugh PR, *et al.* "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* 1975; **12**(3):189–198.
27. Wechsler D. *Wechsler Adult Intelligence Scale (rev. ed.)* Psychological Corporation: New York, 1981.
28. Dartigues JF, Gagnon M, Barberger-Gateau P, Letenneur L, Commenges D, Sauvel C, Michel P, Salamon R. The PAQUID epidemiological program on brain ageing. *Neuroepidemiology* 1992; **11**(1):14–18.
29. Scheike T, Zhang M, Gerds T. Predicting cumulative incidence probability by direct binomial regression. *Biometrika* 2008; **95**:205–220. DOI: 10.1093/biomet/asm096.
30. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer: New York, 2009.
31. Kerr KF, Pepe MS. Joint modeling, covariate adjustment, and interaction: contrasting notions in risk prediction models and risk prediction performance. *Epidemiology* 2011; **22**(6):805–812. DOI: 10.1097/EDE.0b013e31823035fb.
32. Joly P, Commenges D, Helmer C, Letenneur L. A penalized likelihood approach for an illness–death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics* 2002; **3**(3):433. DOI: 10.1093/biostatistics/3.3.433.
33. Aalen O, Borgan Ø, Gjessing HK, Gjessing S. *Survival and Event History Analysis: A Process Point of View*. Springer, 2008.
34. Robins JM, Ritov Y. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine* 1997; **16**:285–319.
35. R Core Team. R: A language and environment for statistical computing, R foundation for statistical computing, Vienna, Austria, 2013. Available from: http://www.R-project.org.

Research Article

Statistics in Medicine

111

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Estimating and comparing time-dependent areas under ROC curves for censored event times with competing risks : Supplementary materials

Paul Blanche[ab]*, Jean-François Dartigues[ab] and Hélène Jacqmin-Gadda[ab]

Notations

For reader convenience, we briefly recall the main notations before introducing technical others and providing asymptotic results.

Main notations

Let M denote a marker that is measured at baseline. Let T denote the event-time, C the censoring time and $\Delta = \mathbb{1}_{(T \leq C)}$ the censoring indicator, with $\mathbb{1}_{(\cdot)}$ denoting the indicator function. Let K denote the number of competing event and η the type of event. Let $\widetilde{T} = \min(T, C)$, the observed time and $\widetilde{\eta} = \Delta\eta$, which indicates either the type of event (when $\widetilde{\eta} \in \{1, \ldots, K\}$) or a censored observation (when $\widetilde{\eta} = 0$). Hereafter, for all time point t, we denote $S_T(t) = \mathbb{P}(T > t)$ and $G(t) = \mathbb{P}(C > t)$ and $\widehat{S}_T(t)$ and $\widehat{G}(t)$ the Kaplan-Meier estimators of $S_T(t)$ and $G(t)$, $S_{\widetilde{T}}(t) = \mathbb{P}(\widetilde{T} > t)$ and $\widehat{S}_{\widetilde{T}}(t)$ its empirical estimator. In addition, let $F_1(t)$ denote the cumulative incidence $\mathbb{P}(T \leq t, \eta = 1)$ and $\widehat{F}_1(t) = \sum_{i=1}^{n} \frac{\mathbb{1}_{(\widetilde{T}_i \leq t, \widetilde{\eta}_i = 1)}}{n\widehat{G}(\widetilde{T}_i)}$ its estimator.
Assume that $\tau_1 > \inf\{u : F_1(u) > 0\}$ and $\tau_2 < \sup\{u : S_{\widetilde{T}}(u) > 0\}$. Thus, $[\tau_1, \tau_2]$ represents a period of times t in which there is both a non null probability of observing a main event before time t and a non null probability of observing someone at risk at time t.

Technical notations

Let $M_{C_i}(t) = \mathbb{1}_{(\widetilde{\eta}_i = 0, \widetilde{T}_i \leq t)} - \int_0^t \mathbb{1}_{(\widetilde{T}_i \geq t)} d\Lambda_C(u)$, where $\Lambda_C(\cdot)$ is the cumulative hazard function of the censoring variable C,

$$h_{tij,1} = \frac{\mathbb{1}_{(\widetilde{T}_i \leq t, \widetilde{\eta}_i = 1)} \mathbb{1}_{(\widetilde{T}_j > t)}}{G(\widetilde{T}_i)G(t)} \mathbb{1}_{(M_i > M_j)}, \quad h_{tij,2} = \frac{\mathbb{1}_{(\widetilde{T}_i \leq t, \widetilde{\eta}_i = 1)} \mathbb{1}_{(\widetilde{T}_j \leq t, \widetilde{\eta}_j \notin \{0,1\})}}{G(\widetilde{T}_i)G(\widetilde{T}_j)} \mathbb{1}_{(M_i > M_j)}, \quad f_{i1t} = \frac{\mathbb{1}_{(\widetilde{T}_i \leq t, \widetilde{\eta}_i = 1)}}{G(\widetilde{T}_i)}$$

and $h_t = \mathbb{E}\left[h_{tij,1} + h_{tij,2}\right]$.
Let $\widehat{M}_C(\cdot)$ be the estimator defined by plugging in the usual Nelson-Aalen estimator of the cumulative incidence function

[a] *Univ. Bordeaux, ISPED, Centre INSERM U897-Epidémiologie-Biostatistique, F-33000 Bordeaux, France*
[b] *INSERM, ISPED, Centre INSERM U897-Epidémiologie-Biostatistique, F-33000 Bordeaux, France*
* *Correspondence to: INSERM U897 - Equipe de biostatistique, ISPED, Université Bordeaux Segalen, 146 rue Leo Saignat, 33076 BORDEAUX CEDEX, France. E-mail: Paul.Blanche@isped.u-bordeaux2.fr*

of the censoring C and let $\widehat{h}_{tij,1}$, $\widehat{h}_{tij,2}$ and $\widehat{f}_{i1t}$ be defined by plugging in the Kaplan-Meier estimator $\widehat{G}(\cdot)$. Let $\widehat{h}_t = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \widehat{h}_{tij,1} + \widehat{h}_{tij,2}$.
Similarly, let us also define

$$h^*_{tij} = \frac{\mathbb{1}_{(\widetilde{T}_i \leq t, \widetilde{\eta}_i = 1)} \mathbb{1}_{(\widetilde{T}_j > t)}}{G(\widetilde{T}_i)} \mathbb{1}_{(M_i > M_j)} \quad \text{and} \quad \widehat{h}^*_{tij} = \frac{\mathbb{1}_{(\widetilde{T}_i \leq t, \widetilde{\eta}_i = 1)} \mathbb{1}_{(\widetilde{T}_j > t)}}{\widehat{G}(\widetilde{T}_i)} \mathbb{1}_{(M_i > M_j)},$$

and $h^*_t = \mathbb{E}\left[h^*_{tij}\right]$ and $\widehat{h}^*_t = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \widehat{h}^*_{tij}$.
Let $(n)_k = n(n-1)\ldots(n-k+1)$ and let $\sum_{i\neq j\neq k\neq l}$, $\sum_{i\neq j\neq k}$ and $\sum_{i\neq j}$ denote respectively the summations over the $\binom{n}{4}$, $\binom{n}{3}$ and $\binom{n}{2}$ distinct 4-element, 3-element and 2-element subsets formed from $\{1,\ldots,n\}$.

A : i.i.d representation for $\widehat{AUC}(t)$

LEMMA 1 : *Let's assume that the censoring time C is independent of (T, η, M), then for all time t in $[\tau_1, \tau_2]$:*

$$\sqrt{n}\left(\widehat{AUC}(t) - AUC(t)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathrm{IF}(\widetilde{T}_i, \widetilde{\eta}_i, M_i, t) + o_p(1)$$

where $\mathbb{E}\left[IF(\widetilde{T}, \widetilde{\eta}, M, t)\right] = 0$ and $IF(\cdot)$ is detailed in the proof.

Proof:

The proof is the adaptation to the competing risks setting of the proof of Theorem 1 of [1]. The martingale representation of the Kaplan-Meier estimator of the censoring survival function entails that

$$\sup_t \left|\sqrt{n}\left(\widehat{G}(t) - G(t)\right) + \frac{G(t)}{\sqrt{n}}\sum_{i=1}^n \int_0^t \frac{dM_{C_i}(u)}{S_{\widetilde{T}}(u)}\right| = o_p(1) \tag{1}$$

See for example [2] for more details. By Taylor expansions, it follows from (1) that

$$\sup_t \left|\sqrt{n}\left(\widehat{h}_t - h_t\right) - \left[\frac{\sqrt{n}}{(n)_4}\sum_{i\neq j\neq k\neq l} h_{tij,1}\left(1 + \int_0^{\widetilde{T}_i}\frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right)\left(1 + \int_0^t \frac{dM_{C_l}(u)}{S_{\widetilde{T}}(u)}\right)\right.\right.$$
$$\left.\left. + h_{tij,2}\left(1 + \int_0^{\widetilde{T}_i}\frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right)\left(1 + \int_0^{\widetilde{T}_j} \frac{dM_{C_l}(u)}{S_{\widetilde{T}}(u)}\right) - h_t\right]\right| = o_p(1) \tag{2}$$

and

$$\sup_t \left|\sqrt{n}\left(\widehat{F}_1(t) - F_1(t)\right) - \left[\frac{\sqrt{n}}{n(n-1)}\sum_{i\neq k}^n f_{i1t}\left(1 + \int_0^{\widetilde{T}_i}\frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right) - F_1(t)\right]\right| = o_p(1) \tag{3}$$

As, $\widehat{AUC}(t) = \widehat{h}_t \Big/ \left[\widehat{F}_1(t)(1 - \widehat{F}_1(t))\right]$, combining (3), (2) and a Taylor expansion of the function $f : (x, y) \mapsto \frac{x}{y(1-y)}$ at $(x, y) = (\widehat{h}_t, \widehat{F}_1(t))$, we further derive

$$\sup_t \left|\sqrt{n}\left(\widehat{AUC}(t) - AUC(t)\right) - \frac{\sqrt{n}}{(n)_4}\sum_{i\neq j\neq k\neq l} \Psi_{ijkl}(t)\right| = o_p(1).$$

where

$$\Psi_{ijkl}(t) = \Bigg\{ h_{tij,1}\left(1+\int_0^{\widetilde{T}_i}\frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right)\left(1+\int_0^{t}\frac{dM_{C_l}(u)}{S_{\widetilde{T}}(u)}\right) + h_{tij,2}\left(1+\int_0^{\widetilde{T}_i}\frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right)\left(1+\int_0^{\widetilde{T}_j}\frac{dM_{C_l}(u)}{S_{\widetilde{T}}(u)}\right) - h_t - \frac{h_t(1-2F_1(t))}{F_1(t)(1-F_1(t))}\left[f_{i1t}\left(1+\int_0^{\widetilde{T}_i}\frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right) - F_1(t)\right]\Bigg\} \Big/ \Big(F_1(t)(1-F_1(t))\Big)$$

By Hájek projection principle, the following Hoeffding's decomposition holds from U-statistic theory [3, 4]:

$$\frac{\sqrt{n}}{(n)_4}\sum_{i\neq j\neq k\neq l}\Psi_{ijkl}(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{IF}(\widetilde{T}_i,\widetilde{\eta}_i,M_i,t) + o_p(1) \tag{4}$$

where

$$\mathrm{IF}(\widetilde{T}_i,\widetilde{\eta}_i,M_i,t) = \mathbb{E}\Big(\Psi_{ijkl}(t)+\Psi_{jikl}(t)+\Psi_{jkil}(t)+\Psi_{jkli}(t)\Big|(\widetilde{T}_i,\widetilde{\eta}_i,M_i)\Big)$$

and $\mathbb{E}\left[\mathrm{IF}(\widetilde{T},\widetilde{\eta},M,t)\right] = 0$.

B : i.i.d representation for $\widehat{AUC}^*(t)$

LEMMA 2 : *Let's assume that the censoring C is independent of (T,η,M), then for all time t in $[\tau_1,\tau_2]$:*

$$\sqrt{n}\left(\widehat{AUC^*}(t) - AUC^*(t)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{IF}^*(\widetilde{T}_i,\widetilde{\eta}_i,M_i,t) + o_p(1)$$

where $\mathbb{E}\left[IF^*(\widetilde{T},\widetilde{\eta},M,t)\right] = 0$ and $IF^*(\cdot)$ is given in the proof.

Proof:

First, let remark that the estimator can be written as

$$\widehat{AUC}^*(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\mathbb{1}_{(\widetilde{T}_i\leq t,\widetilde{\eta}_i=1)}}{n^2\widehat{G}(\widetilde{T}_i)}\mathbb{1}_{(\widetilde{T}_j>t)}\mathbb{1}_{(M_i>M_j)}}{\widehat{S}_{\widetilde{T}}(t)\widehat{F}_1(t)}$$

By Taylor expansion, it follows from (1) that

$$\sup_t\left|\sqrt{n}\left(\widehat{h}_t^* - h_t^*\right) - \left(\frac{\sqrt{n}}{(n)_3}\sum_{i\neq j\neq k}h_{tij}^*\left(1+\int_0^{\widetilde{T}_i}\frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right) - h_t^*\right)\right| = o_p(1) \tag{5}$$

As, $\widehat{AUC}^*(t) = \widehat{h}_t^* \Big/ \left[\widehat{S}_{\widetilde{T}}(t)\widehat{F}_1(t)\right]$, by Taylor expansion of the function $f:(x,y,z)\mapsto\frac{x}{yz}$ at $(x,y,z) = (\widehat{h}_t^*,\widehat{S}_{\widetilde{T}}(t),\widehat{F}_1(t))$, combining (3) and (5), we further derive

$$\sup_t\left|\sqrt{n}\left(\widehat{AUC}^*(t) - AUC^*(t)\right) - \frac{\sqrt{n}}{(n)_3}\sum_{i\neq j\neq k}\Psi_{ijk}^*(t)\right| = o_p(1)$$

where

$$\Psi_{ijk}^*(t) = \Bigg\{ h_{tij}^*\left(1+\int_0^{\widetilde{T}_i}\frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right) - h_t^*\left[\frac{\mathbb{1}_{(\widetilde{T}_i\geq t)}}{S_{\widetilde{T}}(t)} + \frac{1}{F_1(t)}\left\{f_{i1t}\left(1+\int_0^{\widetilde{T}_i}\frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right) - F_1(t)\right\}\right]\Bigg\} \Big/ \Big(S_{\widetilde{T}}(t)F_1(t)\Big)$$

By Hájek projection principle, the following Hoeffding's decomposition holds from U-statistic theory [3, 4]:

$$\frac{\sqrt{n}}{(n)_3}\sum_{i\neq j\neq k}\Psi^*_{ijk}(t)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{IF}^*(\widetilde{T}_i,\widetilde{\eta}_i,M_i,t)+o_p\left(1\right)$$

where

$$\mathrm{IF}^*(\widetilde{T},\widetilde{\eta},M,t)=\mathbb{E}\Big(\Psi^*_{ijk}(t)+\Psi^*_{jik}(t)+\Psi^*_{jki}(t)\Big|(\widetilde{T}_i,\widetilde{\eta}_i,M_i)\Big)$$

and $\mathbb{E}\left[\mathrm{IF}^*(\widetilde{T},\widetilde{\eta},M,t)\right]=0$.

C : i.i.d representation for sensitivity and specificity estimators

Let the generic term A denote either Se, Sp or Sp^*.

LEMMA 3 : *Let's assume that the censoring time C is independent of (T,η,M), then for all time t in $[\tau_1,\tau_2]$, for all cutpoint c in $\{c:0<A(c,t)<1\}$:*

$$\sqrt{n}\left(\widehat{A}(c,t)-A(c,t)\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{IF}^A(\widetilde{T}_i,\widetilde{\eta}_i,M_i,t,c)+o_p\left(1\right)$$

where $\mathbb{E}\left[IF^A(\widetilde{T},\widetilde{\eta},M,t,c)\right]=0$ and $IF^A(\cdot)$ is detailed in the proof.

Proof:

- $A\equiv Sp^*$:
  By Taylor expansion of $f:(x,y)\mapsto x/y$, a little algebra directly leads to :

  $$\mathrm{IF}^{Sp^*}(\widetilde{T}_i,\widetilde{\eta}_i,M_i,t,c)=\frac{1}{S_{\widetilde{T}}(t)}\Big(\mathbb{1}_{(M_i<c,\widetilde{T}_i>t)}-Sp^*(c,t)\mathbb{1}_{(\widetilde{T}_i>t)}\Big)$$

  Consequently, note that a little algebra also leads to :

  $$\mathrm{Var}\left(\mathrm{IF}^{Sp^*}(\widetilde{T}_i,\widetilde{\eta}_i,M_i,t,c)\right)=\frac{Sp^*(c,t)\big(1-Sp^*(c,t)\big)}{S_{\widetilde{T}}(t)}$$

- $A\equiv Se$:
  Combining Taylor expansions and (1) it follows that

  $$\sup_t\left|\sqrt{n}\left(\widehat{Se}(c,t)-Se(c,t)\right)-\frac{\sqrt{n}}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi^{Se}_{ij}(c,t)\right|=o_p(1)\tag{6}$$

  where

  $$\psi^{Se}_{ij}(c,t)=\frac{f_{i1t}}{F_1(t)}\left(1+\int_0^{\widetilde{T}_i}\frac{dM_{C_j}(u)}{S_{\widetilde{T}}(u)}\right)\left[\mathbb{1}_{(M_i>c)}-Se(c,t)\right]$$

  By Hájek projection principle, the following Hoeffding's decomposition holds from U-statistic theory [3, 4]:

  $$\frac{\sqrt{n}}{(n)_3}\sum_{i\neq j}\Psi^{Se}_{ij}(c,t)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{IF}^{Se}(\widetilde{T}_i,\widetilde{\eta}_i,M_i,t,c)+o_p\left(1\right)$$

  where $\mathrm{IF}^{Se}(\widetilde{T}_i,\widetilde{\eta}_i,M_i,t,c)=\mathbb{E}\Big(\Psi^{Se}_{ij}(t)+\Psi^{Se}_{ji}(t)\Big|(\widetilde{T}_i,\widetilde{\eta}_i,M_i)\Big)$ and $\mathbb{E}\left[\mathrm{IF}^{Se}(\widetilde{T},\widetilde{\eta},M,t,c)\right]=0$.

- $A \equiv Sp$:
  Similarly, we first derive

$$\sup_t \left| \sqrt{n}\left(\widehat{Sp}(c,t) - Sp(c,t)\right) - \frac{\sqrt{n}}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi_{ij}^{Sp}(c,t)\right| = o_p(1) \tag{7}$$

where

$$\psi_{ij}^{Sp}(c,t) = \frac{\mathbb{1}_{(M_i \le c)} - Sp(c,t)}{1 - F_1(t)}\left\{\frac{\mathbb{1}_{(\widetilde{T}_i \le t, \widetilde{\eta}_i \notin \{0,1\})}}{G(\widetilde{T}_i)}\left(1 + \int_0^{\widetilde{T}_i}\frac{dM_{C_j}(u)}{S_{\widetilde{T}}(u)}\right) + \frac{\mathbb{1}_{(\widetilde{T}_i > t)}}{G(t)}\left(1 + \int_0^{t}\frac{dM_{C_j}(u)}{S_{\widetilde{T}}(u)}\right)\right\}$$

Again, applying the Hájek projection principle leads to $\mathrm{IF}^{Sp}(\widetilde{T}_i, \widetilde{\eta}_i, M_i, t, c) = \mathbb{E}\left(\Psi_{ij}^{Sp}(t) + \Psi_{ji}^{Sp}(t)\middle|(\widetilde{T}_i, \widetilde{\eta}_i, M_i)\right)$ and $\mathbb{E}\left[\mathrm{IF}^{Sp}(\widetilde{T}, \widetilde{\eta}, M, t, c)\right] = 0$.

References

1. Hung H, Chiang C. Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics* 2010; **38**(1):8–26.
2. Andersen PK, Borgan Ø, Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. New york : Springer Verlag: New York, 1993.
3. Van der Vaart A. *Asymptotic statistics*. Cambridge University Press, 1998.
4. Serfling RJ. *Approximation theorems of mathematical statistics*. New York : John Wiley & Sons, 1980.

D : additional simulation results

Table 1. Simulation results with sample size $n = 200$ and independent censoring, i.e HR_{C_1}=HR_{C_2}=1 (1000 replications). Bias of estimates for two markers M_1 and M_2 (multiplied by 100), empirical coverage probabilities of 95% confidence intervals and Type I errors and powers of the test of $\mathcal{H}_0 : \text{AUC}_{M_1}(t) = \text{AUC}_{M_2}(t)$, depending on $\Delta\text{AUC}(t) = \text{AUC}_{M_1}(t) - \text{AUC}_{M_2}(t)$. IPCW estimators using the Kaplan-Meier estimator (IPCW KM) or a Cox proportional hazards model with covariates M_1 and M_2 (IPCW Cox) for weighting, and the Nearest Neighbor Estimator (NNE). Variances are computed from the estimated influence function (IF) for IPCW KM, or by bootstrapping 400 times (B).

	Method	ΔAUC(t)	AUC(t) True		AUC(t) Bias		Coverage Probability M_1		Coverage Probability M_2		Type I error or Power	
			M_1	M_2	M_1	M_2	IF	B	IF	B	IF	B
$AUC^*(t)$	IPCW KM	0.0	85.6	85.6	0.0	-0.2	93.2	93.5	94.5	94.7	4.6	4.3
	IPCW Cox				0.0	-0.1	-	93.6	-	95.1	-	3.7
	NNE				-0.6	-0.6	-	94.1	-	94.2	-	4.1
$AUC(t)$	IPCW KM	0.0	79.8	79.8	0.1	-0.1	94.6	94.6	95.3	95.2	4.5	4.5
	IPCW Cox				0.1	-0.1	-	94.3	-	95.5	-	4.4
	NNE				-0.3	-0.5	-	94.6	-	95.2	-	4.7
$AUC^*(t)$	IPCW KM	5.0	87.9	82.9	0.0	-0.3	92.7	92.9	94.3	94.4	17.2	16.6
	IPCW Cox				0.1	-0.3	-	92.8	-	95.4	-	17.5
	NNE				-0.5	-0.8	-	94.1	-	94.9	-	18.7
$AUC(t)$	IPCW KM	5.1	82.2	77.1	0.1	-0.3	94.7	94.9	94.4	94.5	23.4	22.6
	IPCW Cox				0.1	-0.3	-	94.7	-	94.6	-	23.3
	NNE				-0.3	-0.7	-	95.3	-	94.4	-	24.2
$AUC^*(t)$	IPCW KM	10.0	89.9	79.9	0.0	-0.2	92.0	92.7	94.0	94.1	49.2	47.7
	IPCW Cox				0.0	-0.2	-	92.6	-	93.9	-	49.0
	NNE				-0.5	-0.7	-	93.9	-	93.9	-	52.9
$AUC(t)$	IPCW KM	10.3	84.4	74.1	0.1	-0.2	93.8	93.6	94.8	95.1	65.6	64.8
	IPCW Cox				0.1	-0.2	-	93.9	-	95.2	-	65.5
	NNE				-0.4	-0.5	-	94.2	-	95.0	-	66.5
$AUC^*(t)$	IPCW KM	14.7	91.4	76.8	0.0	-0.1	91.6	91.9	93.1	94.1	82.4	80.7
	IPCW Cox				0.0	0.0	-	92.4	-	94.2	-	82.4
	NNE				-0.5	-0.6	-	94.5	-	94.3	-	86.2
$AUC(t)$	IPCW KM	15.1	86.0	71.0	0.1	0.0	94.1	94.3	95.2	95.1	93.5	93.5
	IPCW Cox				0.1	0.0	-	94.1	-	95.4	-	93.1
	NNE				-0.3	-0.3	-	94.8	-	94.1	-	93.3

Table 2. Simulation results with sample size $n = 200$ and moderate M_1-dependent censoring, i.e HR_{C_1}=1.35, HR_{C_2}=1 (1000 replications). Bias of estimates for two markers M_1 and M_2 (multiplied by 100), empirical coverage probabilities of 95% confidence intervals and Type I errors and powers of the test of $\mathcal{H}_0 : \text{AUC}_{M_1}(t) = \text{AUC}_{M_2}(t)$, depending on $\Delta\text{AUC}(t) = \text{AUC}_{M_1}(t) - \text{AUC}_{M_2}(t)$. IPCW estimators using the Kaplan-Meier estimator (IPCW KM) or a Cox proportional hazards model with covariates M_1 and M_2 (IPCW Cox) for weighting, and the Nearest Neighbor Estimator (NNE). Variances are computed from the estimated influence function (IF) for IPCW KM, or by bootstrapping 400 times (B).

	Method	ΔAUC(t)	AUC(t) True		AUC(t) Bias		Coverage Probability M_1		Coverage Probability M_2		Type I error or Power	
			M_1	M_2	M_1	M_2	IF	B	IF	B	IF	B
$AUC^*(t)$	IPCW KM	0.0	85.6	85.6	1.6	0.3	86.8	87.6	93.5	93.6	4.4	4.2
	IPCW Cox				0.1	-0.1	-	91.9	-	94.3	-	4.2
	NNE				-0.8	-0.6	-	95.4	-	94.2	-	4.3
$AUC(t)$	IPCW KM	0.0	79.8	79.8	1.3	0.4	91.7	92.0	94.3	94.4	4.7	4.7
	IPCW Cox				0.2	-0.1	-	93.8	-	95.1	-	4.3
	NNE				-0.4	-0.3	-	94.8	-	95.9	-	5.1
$AUC^*(t)$	IPCW KM	5.0	87.9	82.9	1.3	0.3	88.5	89.1	93.0	93.0	23.1	22.4
	IPCW Cox				0.1	-0.2	-	93.0	-	94.5	-	18.2
	NNE				-0.8	-0.7	-	94.5	-	94.6	-	16.2
$AUC(t)$	IPCW KM	5.1	82.2	77.1	1.1	0.4	91.9	91.6	93.7	93.9	26.7	26.1
	IPCW Cox				0.2	-0.2	-	94.7	-	94.9	-	22.2
	NNE				-0.4	-0.4	-	95.1	-	94.4	-	21.6
$AUC^*(t)$	IPCW KM	10.0	89.9	79.9	1.0	0.4	88.5	88.5	92.2	92.6	59.5	57.6
	IPCW Cox				0.1	-0.2	-	92.2	-	94.1	-	51.4
	NNE				-0.7	-0.7	-	94.2	-	94.2	-	52.2
$AUC(t)$	IPCW KM	10.3	84.4	74.1	0.9	0.5	91.6	91.6	93.7	94.2	71.5	70.1
	IPCW Cox				0.1	-0.2	-	93.9	-	95.2	-	65.3
	NNE				-0.4	-0.3	-	94.5	-	95.0	-	64.9
$AUC^*(t)$	IPCW KM	14.7	91.4	76.8	0.9	0.7	87.6	87.7	92	93.1	87.6	86.4
	IPCW Cox				0.1	0.0	-	90.6	-	94.8	-	82.3
	NNE				-0.7	-0.4	-	94.0	-	94.9	-	86.3
$AUC(t)$	IPCW KM	15.1	86.0	71.0	0.8	0.8	91.6	91.5	94	93.6	94.4	94.2
	IPCW Cox				0.1	0.1	-	94.2	-	95.5	-	93.0
	NNE				-0.4	-0.1	-	95.7	-	95.1	-	93.3

Table 3. Simulation results with sample size $n = 200$ and strong dependent censoring (1000 replications). Bias of estimates for two markers M_1 and M_2 (multiplied by 100), empirical coverage probabilities of 95% confidence intervals and Type I errors and powers of the test of $\mathcal{H}_0 : \text{AUC}_{M_1}(t) = \text{AUC}_{M_2}(t)$, depending on $\Delta\text{AUC}(t) = \text{AUC}_{M_1}(t) - \text{AUC}_{M_2}(t)$. IPCW estimators using the Kaplan-Meier estimator (IPCW KM) or a Cox proportional hazards model with covariates M_1 and M_2 (IPCW Cox) for weighting, and the Nearest Neighbor Estimator (NNE). Variances are computed from the estimated influence function (IF) for IPCW KM, or by bootstrapping 400 times (B).

	Method	ΔAUC(t)	AUC(t) True		AUC(t) Bias		Coverage Probability M_1		Coverage Probability M_2		Type I error or Power	
			M_1	M_2	M_1	M_2	IF	B	IF	B	IF	B
HR$_{C_1}$=2, HR$_{C_2}$=1												
$AUC^*(t)$	IPCW KM	0.0	85.6	85.6	2.6	0.6	80.9	81.6	91.9	92.1	7.3	7.2
	IPCW Cox				0.1	-0.1	-	90.1	-	93.5	-	5.2
	NNE				-1.6	-0.7	-	94.1	-	94.8	-	4.4
$AUC(t)$	IPCW KM	0.0	79.8	79.8	2.1	0.9	87.5	87.0	93.3	93.3	5.5	5.2
	IPCW Cox				0.1	-0.1	-	94.1	-	95.7	-	4.5
	NNE				-0.8	-0.2	-	95.2	-	95.5	-	5.0
HR$_{C_1}$=1.35, HR$_{C_2}$=0.6												
$AUC^*(t)$	IPCW KM	0.0	85.6	85.6	1.2	-2.8	89.0	89.8	93.8	94.2	10.3	9.5
	IPCW Cox				0.2	-0.1	-	90.6	-	93.4	-	5.1
	NNE				-0.6	-0.1	-	94.3	-	93.9	-	4.7
$AUC(t)$	IPCW KM	0.0	79.8	79.8	0.6	-1.8	94.4	94.5	94.9	95.2	6.9	6.6
	IPCW Cox				0.2	0.0	-	95.1	-	94.5	-	3.1
	NNE				-0.6	0.1	-	95.0	-	93.6	-	4.0

Table 4. Simulation results with sample size $n = 400$ (1000 replications), corresponding to the first scenrio of Table 1. Average of bias and estimated Asymptotic Standard Error (A.S.E) and coverage probability of 95% CI estimates. (Bias A.S.E = empirical standard error minus average of asymptotic standard error estimates)

	$\mathbb{P}(M_1 \leq c)$	True	Bias	A.S.E	Bias A.S.E	Coverage
$Se(c,t)$	30%	90.9	0.1	2.6	0.0	93.4
	40%	84.5	0.1	3.3	0.1	94.0
	50%	76.2	0.2	3.8	0.0	94.1
	60%	65.9	0.2	4.2	0.0	94.7
	70%	53.5	0.1	4.4	0.1	94.5
$Sp^*(c,t)$	30%	54.3	-0.1	6.4	0.0	94.3
	40%	67.3	-0.3	6.1	0.1	94.0
	50%	78.1	-0.2	5.4	0.1	92.8
	60%	86.6	0.0	4.4	0.1	92.5
	70%	93.0	-0.1	3.3	0.1	90.1
$Sp^*(c,t)$	30%	44.3	0.0	4.2	0.0	94.8
	40%	56.8	-0.2	4.1	-0.1	95.1
	50%	68.0	-0.2	3.8	0.0	95.3
	60%	77.8	-0.1	3.4	0.0	94.2
	70%	86.1	-0.2	2.7	0.0	94.1

IV.2 Compléments

IV.2.1 Une amélioration du calcul des régions de confiance

Les intervalles de confiance (et tests) que nous avons proposés sont « Wald-type », au sens où ils sont centrés autour de l'estimation, et où les bornes sont calculées à partir de $\pm z_{1-\alpha/2}\widehat{\sigma}_t$. Bien qu'ils soient asymptotiquement corrects et aient de bonnes performances dans nos simulations, leurs taux de couverture pourraient probablement être améliorés, notamment pour des tailles d'échantillon plus modestes.

Zou et Yue (2013) rappellent en effet que les AUCs étant des probabilités, il est dommage de fournir des intervalles de confiance qui ne sont pas par définition inclus dans $[0,1]$ (ce qui n'est pas le cas en présence de petits échantillons et/ou d'estimations proches de 1). Comme pour l'estimation de régions de confiance de simples proportions, on peut construire des alternatives (Newcombe, 2012, Chapitre 3). Une idée consiste à considérer une transformation de l'estimateur pour laquelle l'approximation normale est meilleure pour de petits échantillons. Pour l'estimateur d'une probabilité, un choix naturel est la transformation logit. On utilise alors la « méthode delta », qui implique que, si on a (comme dans l'article de la Section IV.1)

$$\sqrt{n}\left(\widehat{AUC}(t) - AUC(t)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_t^2\right),$$

comme la dérivée de la fonction logit : $x \mapsto \log\left(x/\{1-x\}\right)$, est $x \mapsto 1/\big(x(1-x)\big)$, on peut construire l'intervalle de confiance du logit de l'AUC $[l,u]$ par

$$[l,u] = \left\{\log\frac{\widehat{AUC}(t)}{1-\widehat{AUC}(t)} \pm z_{1-\alpha/2}\frac{\sigma_t/\sqrt{n}}{\widehat{AUC}(t)\left(1-\widehat{AUC}(t)\right)}\right\}$$

(dès que $\widehat{AUC}(t)$ est estimé différent de 0 ou 1). Pour améliorer la couverture des intervalles de confiance asymptotiques proposés par DeLong *et al.* (1988), Zou et Yue (2013) proposent de définir l'intervalle de confiance de l'AUC comme l'image de l'intervalle de confiance du logit de l'AUC par la fonction expit (la fonction réciproque de logit), défini par $x \mapsto \exp(x)/\big\{1+\exp(x)\big\}$. Ici, cela correspond alors à $\big[\text{expit}(l), \text{expit}(u)\big]$. Ainsi, ces intervalles sont toujours dans

$[0, 1]$ et ne sont pas symétriques autour de l'estimation. Bien qu'asymptotiquement équivalents aux intervalles « Wald-type », cela leur permet souvent d'avoir de meilleurs taux de couvertures.

Par ailleurs, Zou et Yue (2013) notent qu'il existe aussi d'autres opportunités intéressantes pour raffiner la construction d'intervalles de confiance de probabilités, notamment le « Wilson score interval » (Newcombe, 2012, Section 3.4.3), dont ils proposent une adaptation suivant le raisonnement de Newcombe (2001) pour le cas d'une simple proportion.

En parallèle, Di Termini *et al.* (2012) ont montré par simulation que la couverture de bandes de confiance pour l'incidence cumulée, calculées par une approche similaire à celle de l'article présenté en Section IV.1, pouvait de même être améliorées en utilisant des transformations. Suivant la même idée, on pourrait probablement améliorer la couverture des bandes de confiance de l'AUC en calculant la bande de confiance du logit de l'AUC pour ensuite calculer la bande de confiance de l'AUC par une transformation expit.

IV.2.2 Interprétation à l'aide des variables impropres de Fine et Gray et définition d'un C-index

En présence de risques concurrents, pour faciliter l'interprétation, il peut être intéressant d'utiliser de ce que Fine et Gray (1999) appellent des variables aléatoires « impropres ». Lorsque l'on s'intéresse principalement à l'événement $\eta = 1$, on peut en effet introduire la variable impropre :

$$T^* = \mathbb{1}_{(\eta=1)} \times T + \mathbb{1}_{(\eta\neq 1)} \times \infty,$$

avec la convention $0 \times \infty = 0$. Sachant M, la variable aléatoire impropre T^* a alors une fonction de répartition égale à la fonction d'incidence cumulée $F_1(t|M) = \mathbb{P}(T \leqslant t, \eta = 1|M)$ pour tout $t < \infty$, et une masse égale à $\mathbb{P}(T^* = \infty|M) = 1 - F_1(\infty|M)$ en $t = \infty$. C'est en fait le modèle du risque instantané de T^* qui est à la base du populaire modèle prédictif de Fine et Gray (1999).

Lorsque l'on souhaite faire de la prédiction en présence de risques concurrents, T^* joue

souvent un rôle analogue à celui du temps de survie T sans risque concurrent. En notant que

$$\begin{aligned} AUC(t) &= \mathbb{P}\Big(M_i > M_j \Big| T_i \leqslant t, \eta_i = 1, \{T_j > t\} \cup \ \{T_j \leqslant t, \eta_j \neq 1\}\Big) \\ &= \mathbb{P}\big(M_i > M_j | T_i^* \leqslant t, T_j^* > t\big), \end{aligned}$$

la définition d'un contrôle comme un sujet i vérifiant $T_i \leqslant t, \eta_i \neq 1$ ou $T_i > t$ apparaît une fois de plus assez naturelle.

De même, en rappelant que le C-index tronqué en τ_0 est, sans risque concurrent, défini par $\mathcal{C}(\tau_0) = \mathbb{P}(M_i > M_j | T_i < T_j, T_i < \tau_0)$, alors on peut aussi naturellement définir un C-index en présence de risques concurrents par

$$\begin{aligned} \mathcal{C}(\tau_0) &= \mathbb{P}(M_i > M_j | T_i^* < T_j^*, T_i^* < \tau_0) \\ &= \mathbb{P}\Big(M_i > M_j \Big| T_i < \tau_0, \eta_i = 1, \{T_i < T_j\} \cup \{T_i \geqslant T_j, \eta_j \neq 1\}\Big). \end{aligned}$$

Comme pour l'AUC, on peut estimer $\mathcal{C}(\tau_0)$ par une méthode IPCW. Sous l'hypothèse que la censure ne dépend pas du marqueur M, on peut aussi utiliser l'estimateur de Kaplan-Meier pour définir les poids, montrer que l'estimateur a de bonnes propriétés asymptotiques, et notamment une représentation i.i.d de la même forme que celle du lemme 1 de l'article de la Section IV.1. De même, on peut estimer ces termes i.i.d simplement à partir d'estimateurs empiriques et en dériver des régions de confiance ou des tests (Wolbers *et al.*, 2013).

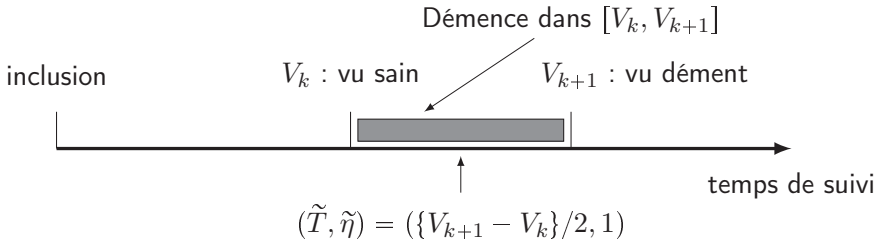
La définition de ce C-index avait été informellement suggérée par Wolbers *et al.* (2009). Plus récemment, nous avons complété ce travail en proposant des estimateurs IPCW consistants et asymptotiquement gaussiens (Wolbers *et al.*, 2013). Une étude de simulation a montré de bons résultats pour les procédures d'inférence proposées, et une application à l'étude des capacités pronostiques d'un score de risque d'événement cardio-vasculaire a illustré l'intérêt pratique de la méthode (Wolbers *et al.*, 2013).

IV.2.3 Prise en compte de la censure par intervalle par un modèle multi-états

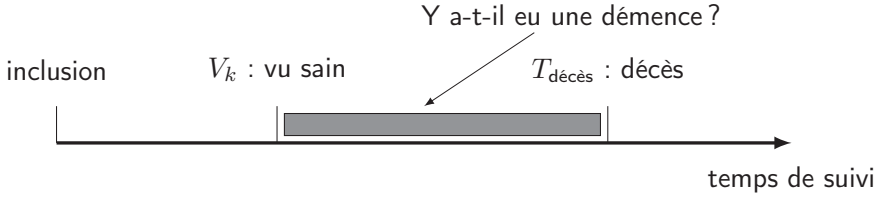
Une particularité des données de cohorte de vieillissement est que l'événement démence est censuré par intervalle. En effet, on ne connaît pas la date exacte du début de la démence d'un sujet. On peut seulement savoir qu'elle apparaît entre deux visites auxquelles le sujet est vu par un psychologue : la démence apparaît entre la date de la dernière visite à laquelle le sujet est vu sain, et la première à laquelle il est vu dément. Par ailleurs, lorsqu'un sujet décède sans avoir été diagnostiqué dément, on ne sait pas s'il a développé une démence entre la date de dernière visite à laquelle il a été vu sain et son décès. Par opposition, les dates de décès sont connues de manière exacte.

Pour étudier ces données censurées par intervalle dans l'application de l'article de la Section IV.1, nous avons utilisé une règle d'imputation « brutale » pour créer les données $\{(\widetilde{T}_i, \widetilde{\eta}_i), i = 1, \ldots, n\}$. Elle consistait en la méthode suivante, illustrée par la Figure IV.1. Lorsqu'une démence était diagnostiquée à une visite, le temps de démence était imputé égal au milieu de l'intervalle entre la dernière visite précédant le diagnostic (V_k) et la visite de diagnostic (V_{k+1}) (Figure IV.1a). Lorsqu'un sujet décédait sans avoir été diagnostiqué dément, il était considéré comme décédé sans démence s'il décédait dans les deux ans suivant la dernière visite (V_k), et comme censuré à la dernière visite autrement (Figure IV.1b). Enfin, un sujet non diagnostiqué dément et non décédé durant le suivi était considéré censuré à sa dernière visite.

Dans un autre travail (Jacqmin-Gadda *et al.*, 2013c), nous avons proposé deux méthodes de calcul de l'AUC pour ce type de données censurées par intervalle, qui permettent d'éviter cette méthode d'imputation « brutale ». Pour cela, nous avons utilisé un modèle multi-états « illness-death » markovien non-homogène, représenté à la Figure IV.2. L'idée consiste à modéliser un processus stochastique, que nous noterons $(E(t), t \in \mathbb{R})$, qui modélise l' état dans lequel se trouve un sujet au temps t : sain, noté $E(t) = 0$, dément, noté $E(t) = 1$ ou décédé, noté $E(t) = 2$. Avec ces notations, on a la correspondance suivante avec les notations de l'article précédent $T = \inf\{t \geqslant 0 : E(t) \neq 0\}$, $\eta = E(T)$. Pour les modèles multi-états markoviens, les



(a) Démence observée en V_{k+1}



$$(\tilde{T}, \tilde{\eta}) = \begin{cases} (T_{\text{décès}}, 2) & \text{si } T_{\text{décès}} - V_k < 2 \text{ ans} \\ (V_k, 0) & \text{sinon} \end{cases}$$

(b) Décès sans démence observée

FIGURE IV.1 – Imputation des données de survie dans l'article Blanche *et al.* (2013a)

quantités clés sont les intensités de transition. Elles sont définies (sachant M) par

$$\lambda_{hj}(t|M) = \lim_{dt \downarrow 0} \frac{1}{dt} \mathbb{P}\big(E(t+dt) = j | E(t) = h, M\big),$$

pour $h = 0, 1$ et $j = 1, 2$. L'intérêt de ce modèle est qu'il permet d'écrire la vraisemblance des données censurées par intervalle sans avoir recours à aucune approximation. En effet, la vraisemblance des données s'écrit naturellement en fonction des intensités de transition et des observations.

Par exemple, dans le cas de la Figure IV.1b, pour lequel on ne sait pas si un sujet est devenu dément avant son décès, sa vraisemblance s'écrit comme la somme de deux probabilités : celle d'être devenu dément avant de décéder plus celle d'être décédé sans démence. C'est en fait la somme des probabilités des deux trajectoires possibles dans le modèle multi-états markovien non-homogène, sachant que l'on observe un état sain et un décès à deux instants ponctuels

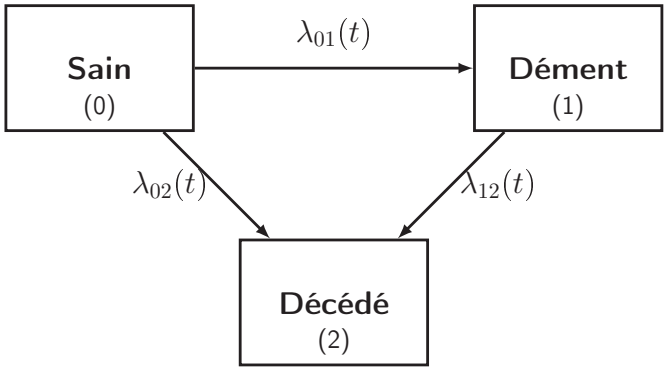


FIGURE IV.2 – Le modèle *illness-death*.

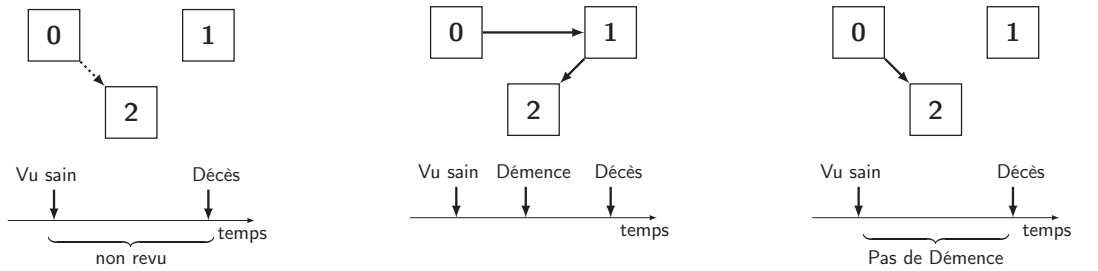
et rien entre les deux, comme illustré à la Figure IV.3. Dans ce cas, la contribution à la vraisemblance du sujet s'écrira (en omettant le conditionnement sur M par souci de brièveté) :

$$\underbrace{\exp\left(-\int_0^{T_{\text{décès}}}\Big\{\lambda_{01}(u)+\lambda_{02}(u)\Big\}du\right)}_{\text{sain entre } 0 \text{ et } T_{\text{décès}}-}\overbrace{\lambda_{02}(T_{\text{décès}})}^{\text{décès en } T_{\text{décès}} \text{ sachant sain en } T_{\text{décès}}-}$$

$$+\int_{V_k}^{T_{\text{décès}}}\underbrace{\exp\left(-\int_0^{s}\Big\{\lambda_{01}(u)+\lambda_{02}(u)\Big\}du\right)}_{\text{sain entre } 0 \text{ et } s-}\overbrace{\lambda_{01}(s)}^{\text{démence en } s \text{ sachant sain en } s-}\underbrace{\exp\left(-\int_s^{T_{\text{décès}}}\lambda_{12}(u)du\right)}_{\text{dément entre } s \text{ et } T_{\text{décès}}-}\overbrace{\lambda_{12}(T_{\text{décès}})}^{\text{décès en } T_{\text{décès}} \text{ sachant dément en } T_{\text{décès}}-}ds.$$

Le premier terme correspond à la trajectoire de la Figure IV.3c, le second à celle de la Figure IV.3b.

Pour davantage de détails sur l'écriture de la vraisemblance, on pourra consulter Joly *et al.* (2002). Par ailleurs, pour une discussion plus théorique justifiant l'écriture de cette vraisemblance et les hypothèses supposées à partir de la théorie des processus de comptage, on pourra consulter Commenges et Gégout-Petit (2007).



(a) Trajectoire observée

(b) Possibilité 1 : démence non observée puis décès

(c) Possibilité 2 : décès sans démence

FIGURE IV.3 – Trajectoires observées et possibles des sujets décédant après une dernière visite à laquelle ils ont été vus sains.

En paramétrant les intensités de transition par un modèle à risques proportionnels, i.e.

$$\lambda_{hj}(t|M) = \lambda^0_{hj}(t)\exp(\beta_{hj}M), \tag{IV.1}$$

pour $h = 0, 1$, $j = 1, 2$, on peut estimer les paramètres du modèle par la méthode du maximum de vraisemblance. Les risques de base $\lambda^0_{hj}(t)$, $h = 0, 1$, $j = 1, 2$, peuvent être modélisés de diverses façons, via des lois de Weibull ou des splines par exemple. Pour l'utilisation de splines, Joly *et al.* (2002) proposent en fait une méthode de maximum de vraisemblance pénalisée, en pénalisant sur les dérivées secondes des risques de base. L'idée sous-jacente est d'éviter que la grande souplesse des splines, qui mime une estimation complètement non paramétrique, induise l'estimation de risques de base trop irréguliers, qui surajouterait les données.

En notant que

$$\begin{aligned} Se(c,t) &= \mathbb{P}(M > c | T \leqslant t, \eta = 1) \\ &= \left\{ \int_c^\infty F_1(t|m) f_M(m) dm \right\} \Big/ \left\{ \int_{-\infty}^\infty F_1(t|m) f_M(m) dm \right\}, \end{aligned}$$

avec $f_M(\cdot)$ la densité de distribution du marqueur M, et

$$F_1(t|m) = \mathbb{P}(T \leqslant t, \eta = 1|M = m) \tag{IV.2}$$

$$= \int_0^t \lambda_{01}(u|m) \exp\left(-\int_0^u \Big\{\lambda_{01}(s|m) + \lambda_{02}(s|m)\Big\} ds\right) du, \tag{IV.3}$$

on peut définir un estimateur semi-paramétrique de $Se(c,t)$ à partir du $n-$échantillon des données par

$$\widehat{Se}(c,t) = \left\{\frac{1}{n}\sum_{i=1}^n \widehat{F}_1(t|M_i)\mathbb{1}_{(M_i>c)}\right\} \Big/ \left\{\frac{1}{n}\sum_{i=1}^n \widehat{F}_1(t|M_i)\right\}, \tag{IV.4}$$

où $\widehat{F}_1(t|M_i)$ est défini par (IV.1) et (IV.3) en remplaçant les paramètres par leur estimation. Cet estimateur est semiparamétrique au sens où il est paramétrique sur la distribution de (T,η) sachant M, mais non paramétrique sur la distribution de M, puisque l'espérance sur M est estimée par une moyenne empirique. Suivant le même raisonnement, on peut aussi estimer les spécificités et les AUCs.

Un inconvénient de cette méthode est néanmoins qu'elle suppose quelques hypothèses liées à la modélisation, comme la log-linéarité de l'effet de M sur chaque intensité de transition, la proportionnalité des risques et l'hypothèse markovienne. Pourtant, de telles hypothèses ne sont pas toujours souhaitables ni raisonnables en pratique. De plus, vérifier ces hypothèses par des méthodes graphiques, basées sur les résidus martingales par exemple (Therneau, 2000, Chapitre 4), comme on peut le faire en présence de données censurées à droite, n'est pas, à notre connaissance, faisable avec ce type de données censurées par intervalle.

En même temps, pour de nombreux sujets on connaît leur état sain, dément ou décédé sans démence en t, donc on sait s'ils sont des cas ou des contrôles en t. Il est donc dommage de ne pas utiliser cette information. Rappelons que si elle était disponible pour tous les sujets, on pourrait utiliser les estimateurs non paramétriques usuels, basés sur de simples proportions.

Nous avons donc également proposé des estimateurs alternatifs, basés sur une méthode dite d'imputation (Jacqmin-Gadda *et al.*, 2013c). L'idée consiste à définir des estimateurs par de simples proportions, pour lesquelles on compte 1 ou 0 pour les sujets dont on sait s'ils sont des cas ou des contrôles en t, et pour lesquelles on compte un score entre 0 et 1 pour les

autres. Ce score correspond à la probabilité d'être un cas ou un contrôle en t, sachant toute l'information sur le sujet. L'information disponible sur le sujet inclut sa valeur du biomarqueur, mais aussi la date de la dernière visite à laquelle il a été vu sans démence, le statut vital connu (décédé ou vivant) et éventuellement la date de la première visite à laquelle il a été vu dément. Les probabilités d'être un cas ou un contrôle sont estimées à partir des estimations du modèle *illness-death* markovien obtenues par maximisation de la vraisemblance ou de la vraisemblance pénalisée. Les estimateurs par imputations font donc les mêmes hypothèses que les estimateurs semiparamétriques précédemment décrits. Néanmoins, ils seront généralement moins biaisés en cas de mauvaise spécification du modèle, dès lors que la proportion de sujets pour lesquels on ne connaît pas le statut cas ou contrôle en t est faible.

Pour comparer les différents estimateurs en présence de données censurées par intervalle, nous avons présenté une étude de simulation (Jacqmin-Gadda *et al.*, 2013c). L'une des conclusions était que l'estimateur par imputation était le moins biaisé des deux estimateurs utilisant le modèle multi-états en cas de mauvaise spécification. Les résultats ont aussi montré que les biais engendrés par l'utilisation de la méthode d'imputation « brutale » et des estimateurs IPCW étaient généralement faibles, et étaient du même ordre ou plus faibles que ceux des estimateurs basés sur le modèle multi-états, en cas de mauvaise spécification.

En conclusion, en présence de données censurées par intervalle du type de celle de Paquid et Trois-Cités, les estimateurs semi-paramétriques comme celui de l'équation (IV.4) sont à éviter et on recommande plutôt d'estimer la courbe ROC par les estimateurs IPCW et la méthode d'imputation « brutale ». Cependant, on conseille de valider les résultats avec l'estimateur par imputation basé sur le modèle illness-death, dans l'esprit d'une analyse de sensibilité.

Enfin, notons qu'aujourd'hui il n'existe pas, à notre connaissance, de méthode du maximum de vraisemblance non-paramétrique pour des données censurées par intervalle comme celle de Paquid. Certains travaux ont cependant récemment fait quelques avancées dans ce sens (Frydman et Szarek, 2009; Frydman *et al.*, 2013). S'ils étaient étendus à la prise en compte de covariables, ils pourraient cependant représenter des opportunités intéressantes pour définir

des estimateurs non paramétriques consistants.

IV.2.4 Application à un score pronostique composite de la démence

Dans un travail épidémiologique connexe (Jacqmin-Gadda *et al.*, 2013b), l'objectif était de proposer un score prédictif de la démence, principalement à 10 ans, en fonction de différents scores cognitifs, du sexe, de l'âge et du niveau d'études. Pour tenir compte du risque compétitif de décès sans démence et de la censure par intervalle, le score pronostique a été estimé avec le modèle multi-états « illness-death » (Figure IV.2), par la méthode du maximum de vraisemblance (Joly *et al.*, 2002).

La combinaison linéaire correspondant au prédicteur linéaire d'un modèle à risques proportionnels pour l'intensité de transition de l'état sain à l'état dément a été retenue comme score pronostique.

Bien que d'autres approches étaient possibles, ce choix a été motivé par les bonnes performances de ce score et surtout sa relative simplicité, qui en fait un score à la fois transparent et probablement assez portable d'une population à l'autre (car ne dépendant pas des risques de bases des diverses intensités de transitions par exemple, i.e. des incidences de la démence, du décès sans démence ou du décès des déments).

Le score prédictif a été estimé sur un échantillon final de 2795 sujets de la cohorte Paquid. Il combinait les variables suivantes : l'âge du sujet, son niveau d'éducation, ses plaintes mnésiques, les trois tests psychométriques du *Mini-Mental State Examination* (MMSE), du *Digit symbol substitution test* et du *Test de fluence verbal d'Isaac* (présenté en Section I.2.3) et le sous score du MMSE correspondant à l'évaluation de la mémoire épisodique.

Les estimations des AUCs ont été réalisées avec deux méthodes : la méthode d'imputation « brutale » des données utilisant les estimateurs IPCW (notée « IPCW ») et la méthode d'imputation utilisant le modèle multi-états (notée « imputation »). Les estimations des AUCs à 10 ans étaient, respectivement pour les méthodes « IPCW » et « imputation », $\widehat{AUC}^*(t = 10) = 81.4\%$ et 82.8% (écarts types estimés (s.e)= 1.5% et 1.4%) et $\widehat{AUC}(t = 10) = 75.0\%$ et 77.0% (s.e=1.6% et 1.6%). Comme ces estimations d'AUCs étaient réalisées sur les mêmes

données que celles ayant été utilisées pour estimer le score prédictif, une estimation par « 10-fold cross-validation » a aussi été réalisée pour estimer un éventuel biais d'optimisme. Elle a mené aux mêmes résultats à plus ou moins 0.6% et 0.9%, respectivement. Cette faible différence entre les AUCs estimés avec ou sans cross-validation peut s'expliquer par le faible nombre de paramètres du modèle multi-états retenu, comparé à la taille de l'échantillon.

IV.2.5 Implémentation dans le package 'timeROC'

Le package R **'timeROC'** permet de calculer les estimateurs présentés dans l'article de la Section IV.1. Il est disponible sur le site du CRAN (« Comprehensive R Archive Network »), à l'adresse `http://cran.r-project.org`.

La fonction principale est la fonction `timeROC`. Elle calcule les estimateurs IPCW de la sensibilité, de la spécificité et de l'AUC en présence de données censurées, avec ou sans risque compétitif. L'estimation des termes i.i.d des décompositions de l'AUC est également fournie.

Plusieurs fonctions permettent de réutiliser les sorties de la fonction `timeROC` et de présenter l'essentiel des résultats :

- ▷ `print` permet de visualiser un résumé succinct des estimations.
- ▷ `confint` permet de calculer des intervalles de confiance et des bandes de confiance simultanées d'AUCs.
- ▷ `compare` permet de tester des différences d'AUCs.
- ▷ `plot` permet de tracer des courbes ROC.
- ▷ `plotAUCcurve` permet de tracer des courbes de l'évolution de l'AUC, de ses intervalles de confiance et de ses bandes de confiance simultanées au cours du temps.
- ▷ `plotAUCcurveDiff` permet de tracer des courbes de l'évolution d'une différence d'AUCs, de ses intervalles de confiance et de ses bandes de confiance simultanées au cours du temps.

La fonction indépendante `SeSpPPVNPV` est aussi incluse. Elle permet d'estimer des sensibilités, des spécificités, des valeurs positives prédictives (PPV) et des valeurs positives négatives (NPV) ainsi que leurs écart-types pour une valeur seuil donnée.

Enfin, notons aussi que le package R **'smoothHazard'** (Touraine *et al.*, 2013) permet d'estimer le modèle multi-états markovien *illness-death* décrit en Section IV.2.3. Les sorties du package permettent ensuite de calculer les estimateurs alternatifs proposés en Section IV.2.3, qui prennent en compte la censure par intervalle à l'aide d'un modèle multi-états.

IV.2.6 Conclusion du chapitre

1. Dans ce chapitre, on a proposé des méthodes d'estimations IPCW (« Inverse Probability of Censoring Weighting ») pour la courbe ROC dépendant du temps, et l'aire sous cette courbe, en présence de risques compétitifs et de données censurées à droite. Quelques résultats asymptotiques ont été présentés pour la construction de régions de confiance et de tests de comparaison d'AUCs, et des simulations ont illustré leurs bonnes performances.

2. Pour faciliter l'accessibilité de la méthode à d'autres utilisateurs, nous avons créé le package **'timeROC'**.

3. Bien que les méthodes développées ne supposent qu'une censure à droite, et pas de censure par intervalle, nos simulations ont montré qu'une approche d'approximation des données censurées par intervalle, par une imputation « brutale », ne semblait pas introduire de biais majeurs. La méthode apparaît ainsi robuste à une application aux données des cohortes Paquid et Trois-Cités.

4. Nous avons également proposé des méthodes alternatives, basées sur un modèle multi-états *illness-death*, pour spécifiquement traiter la censure par intervalle. En particulier une méthode d'imputation paramétrique que nous recommandons d'utiliser dans l'esprit d'une analyse de sensibilité.

5. La motivation épidémiologique de cette thèse était le développement et l'évaluation d'un score pronostique pour la démence. Nous avons proposé un tel score. À notre connaissance, c'est le premier score pronostique proposé et évalué en tenant compte du risque compétitif de décès et des problèmes de censure par intervalle. Par ailleurs, un accent a

été mis sur la clarification de la notion de capacité pronostique de ce score, en contrastant deux définitions de l'AUC.

6. Enfin, bien que le travail de ce chapitre porte sur l'étude de la démence, les méthodes proposées pourront s'appliquer au delà de ce contexte, car le développement et la validation de score pronostique connaissent aussi un intérêt pour de nombreuses autres applications. Notons d'ailleurs que pour les pronostics de cancer et de risques cardio-vasculaires, entre autres, les données sont le plus souvent uniquement censurées à droite. Les méthodes proposées dans la Section IV.1 et implémentées dans **'timeROC'** peuvent donc directement être appliquées à ces pathologies.

V. Comparaisons de prédictions dynamiques basées sur des mesures répétées d'un marqueur

V.1 Manuscrit soumis à publication

Résumé : Dans ce troisième travail, on s'intéresse à évaluer et comparer des prédictions dynamiques issues de modèles conjoints pour un marqueur longitudinal et des données de survie. On dit que les prédictions sont dynamiques car elles sont actualisées au fur et à mesure que l'information sur le sujet croît avec le temps landmark à partir duquel on fait la prédiction. Pour cela, on propose l'utilisation du Brier score (BS), de l'aire sous la courbe ROC (AUC) et de leurs estimateurs IPCW adaptés à la présence de risques concurrents et de données censurées. Quelques propriétés asymptotiques sont étudiées, dont on dérive des régions de confiance et des tests pour comparer des courbes de BS et d'AUC. Les tests permettent notamment de comparer simultanément les prédictions dynamiques de deux modèles pronostiques à plusieurs temps landmark.

Les procédures d'inférence sont évaluées à partir de simulations avant d'être appliquées à nos données de motivation : les cohortes Paquid et Trois-Cités. Deux modèles conjoints sont estimés sur les données de Paquid, et deux outils de prédictions dynamiques de démence en sont dérivés pour les sujets de Trois-Cités, chacun utilisant un test cognitif différent. Les méthodes proposées permettent de mettre en évidence des différences significatives entre les capacités pronostiques des deux outils pronostiques dynamiques.

Ci-après sont présentés le manuscrit principal et son *web-appendix*.

XXX 000, 000–000 DOI: 000
000 0000

Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks

Paul Blanche [1,2,*], Cécile Proust-Lima[1,2], Lucie Loubère[1,2], Claudine Berr[3]

and

Jean-François Dartigues[1,2], Hélène Jacqmin-Gadda[1,2]

[1] Univ. Bordeaux Segalen, ISPED, Inserm Research Center U897, F33076, Bordeaux, France

[2] INSERM, ISPED, Centre INSERM U897-Epidémiologie-Biostatistique, F-33000 Bordeaux, France

[3] INSERM, Centre INSERM U1061, Université Montpellier 1, Montpellier, France

*email: Paul.Blanche@isped.u-bordeaux2.fr

Summary:

Motivated by the growing interest on personalized medicine, joint modeling of longitudinal marker and time-to-event has recently started to be used to derive dynamic individual risk predictions. Individual predictions are called dynamic because there are updated when information on the subject's health profile grows with time. We focus in this work on statistical methods for quantifying and comparing dynamic predictive accuracy of this kind of prognostic models, accounting for right censoring and possibly competing events. Dynamic area under the ROC curve (AUC) and Brier Score (BS) are used to quantify predictive accuracy. Nonparametric inverse probability of censoring weighting technique is used to estimate dynamic curves of AUC and BS as functions of the time at which predictions are made. Asymptotic results are established and pointwise confidence intervals as well as simultaneous confidence bands are derived. Tests are also proposed to compare the dynamic prediction accuracy curves of two prognostic models. The finite sample behavior of the inference procedures are assessed via simulations. We apply the proposed methodology to compare different prediction models using repeated measures of different psychometric tests to predict dementia in the elderly, accounting for the competing risk of death. Models are estimated on the French Paquid cohort and predictive accuracies are evaluated and compared on the French Three-City cohort.

Key words: Competing risks; Dynamic prediction; Joint model; Longitudinal data; Prediction accuracy

1. Introduction

Risk prediction models play important roles in numerous medical fields such as oncology and cardiology. For patient counseling or targeting early detection of disease, they constitute a basis of personalized medicine. In neurology for instance, current researches focus on preventive treatments against Alzheimer's disease that could be administered in the pre-clinical phase to subjects at high risk of dementia (Aisen et al., 2011). Accurate prediction models for dementia onset are thus required. Since the decline in cognitive functions could begin long before a dementia diagnosis (Amieva et al., 2008), repeated measures of cognitive tests over time could be helpful for the prediction of Alzheimer's disease.

Whatever the clinical setting, ideally predictions should be as personalized as possible. For instance in Alzheimer's disease context, they should make the best use of the available information about individual cognitive decline available at the landmark time s at which predictions are made. We thus use the terminology of individual *dynamic predictions* to define such predictions that are expected to be updated when information on the subject's clinical profile grows with time. In Section 2, we remind how the joint modeling of a longitudinal marker and of a time-to-event can be used for building such dynamic predictions (Rizopoulos, 2011; Proust-Lima et al., 2012; Rizopoulos, 2012).

To be useful in clinical practice and health care, dynamic predictions need to have a good prediction accuracy. Appropriate measures for quantifying prediction accuracy and proposals for corresponding estimators are thus essential. Both the dynamic nature of the prediction and the potential presence of competing events should impact the definitions of predictive accuracy since they are real phenomena. For example, in the context of dementia prognosis in the elderly, predictive accuracy definitions should depend on both, the landmark time at which predictions are made, and of which depends the amount of information available, and on the competing risk of dementia-free death. By contrast, censoring induced by lost

to follow-up is a nuisance that should not impact the definition of predictive accuracy but that has to be handled by the estimators. Thus, in Section 3 we propose adaptations of definitions of Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) and expected Brier Score (BS) to the setting of dynamic prediction with competing risks. The proposed definitions combine previously proposed definitions of AUC and BS for the dynamic prediction setting (Parast et al., 2012; Schoop et al., 2008) and for the competing risk setting (Schoop et al., 2011; Zheng et al., 2012; Blanche et al., 2013). We then discuss their meaning and propose non-parametric Inverse Probability of Censoring Weighting (IPCW) estimators.

Induced by the recent advances on medical and statistical knowledge and the novel computation opportunities, more and more prediction models are becoming available. Before using one of them in clinical practice, it would be useful to rigorously compare their predictive abilities. Motivated by the prediction of Alzheimer's disease using aging cohort data, the main objective of this work is therefore to provide inference procedures to rigorously quantify and compare several dynamic prediction strategies. Thus, in Section 3 we provide large sample results. They rigorously justify and enable computation of pointwise confidence intervals, simultaneous confidence bands as well as tests for comparing two dynamic prediction accuracy curves of two dynamic prediction strategies. Within the context of dynamic predictions based on joint modeling, simultaneous confidence bands are useful to compare dynamic predictions simultaneously over a set of landmark times.The finite sample behavior of the asymptotic inference procedures are investigated through a simulation study in Section 4.

Finally, Section 5 illustrates the interest of the proposed methodology. Two joint models for dementia prediction using repeated measures of two different cognitive tests are estimated on the French cohort Paquid (training set). The 5-years predictive accuracy of the two models is then compared on a validation data set coming from the French Three-City cohort.

2. Building dynamic prediction tools by joint modeling

2.1 *Notations*

Let T denote a time-to-event and η the cause of the event. For sake of simplicity, we assume only two competing events and denote $\eta = 1$ the main event (e.g: dementia) and $\eta = 2$ the competing event (e.g: dementia-free death). Due to censoring C induced by lost to follow-up, we observe $\widetilde{T} = \min(T, C)$ and $\widetilde{\eta} = \Delta\eta$ where $\Delta = \mathbb{1}_{(T \leqslant C)}$ and $\mathbb{1}_{(\cdot)}$ denotes the indicator function. Let $Y(t)$ be a marker measurement at time t. We assume that we observe the independent and identically distributed (i.i.d) sample of n subjects $\left\{(\widetilde{T}_i, \Delta_i, \widetilde{\eta}_i, \mathbf{Y}_i, \mathbf{X}_i), i = 1, \ldots, n\right\}$, where $\mathbf{X}_i$ denotes a vector of baseline covariates (e.g: gender and education level) and $\mathbf{Y}_i$ denotes the vector of n_i observed repeated marker measurements $Y_{ij} \equiv Y_i(t_{ij})$ for subject i at time t_{ij}, $j = 1, \ldots, n_i$ (e.g: repeated measurements of a cognitive test).

2.2 *Joint modeling of longitudinal marker and time to event with competing risks*

Joint models aim at jointly modeling the marker trajectory $Y(\cdot)$ and the time and cause of the event (T, η). Most often, a fully parametric approach is used. The joint probability distribution of $\big(T, \eta, Y(\cdot)\big)$ is thus parametrized by a vector of parameters $\boldsymbol{\xi}$. Modeling strategies are diverse in the literature, but in most of the cases, the full likelihood of the data $\left\{(\widetilde{T}_i, \widetilde{\eta}_i, \mathbf{Y}_i), i = 1, \ldots, n\right\}$ given baseline covariates $\mathbf{X}_i, i = 1, \ldots, n$, denoted by $\mathcal{L}_{(\widetilde{T}, \widetilde{\eta}, \mathbf{Y})}$, can be written as

$$\mathcal{L}_{(\widetilde{T}, \widetilde{\eta}, \mathbf{Y})} = \prod_{i=1}^{n} \int_{\boldsymbol{\gamma}_i} \mathcal{L}_{\mathbf{Y}} \left\{\mathbf{Y}_i | \boldsymbol{\gamma}_i\right\} \mathcal{L}_{(\widetilde{T}, \widetilde{\eta})} \left\{(\widetilde{T}_i, \widetilde{\eta}_i) | \boldsymbol{\gamma}_i\right\} f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_i) d\boldsymbol{\gamma}_i.$$

The term $\mathcal{L}_{\mathbf{Y}} \left\{\mathbf{Y}_i | \boldsymbol{\gamma}_i\right\}$ denotes the contribution to the likelihood of the repeated marker measurements $\mathbf{Y}_i$ of subject i given a shared latent variable $\boldsymbol{\gamma}_i$ (discrete or continuous) of distribution $f_{\boldsymbol{\gamma}}(\cdot)$. The contribution to the likelihood of the survival data $(\widetilde{T}_i, \widetilde{\eta}_i)$ of subject i given $\boldsymbol{\gamma}_i$ is defined as

$$\mathcal{L}_{(\widetilde{T}, \widetilde{\eta})} \left\{(\widetilde{T}_i, \widetilde{\eta}_i) | \boldsymbol{\gamma}_i\right\} = \lambda_1(\widetilde{T}_i | \boldsymbol{\gamma}_i)^{\mathbb{1}_{(\widetilde{\eta}=1)}} \lambda_2(\widetilde{T}_i | \boldsymbol{\gamma}_i)^{\mathbb{1}_{(\widetilde{\eta}=2)}} S(\widetilde{T}_i | \boldsymbol{\gamma}_i),$$

138

where $S(\widetilde{T}_i|\boldsymbol{\gamma}_i) = \exp\left(-\int_0^{\widetilde{T}_i}\{\lambda_1(u|\boldsymbol{\gamma}_i) + \lambda_2(u|\boldsymbol{\gamma}_i)\}\,du\right)$. In practice, parametric proportional hazards models are usually chosen for modeling the cause specific hazards given $\boldsymbol{\gamma}_i$ denoted by $\lambda_k(\cdot|\boldsymbol{\gamma}_i), k = 1, 2$, while a linear mixed model is the most often chosen for modeling the marker trajectory $Y(\cdot)$, leading $\mathcal{L}_{\mathbf{Y}}\{\mathbf{Y}_i|\boldsymbol{\gamma}_i\}$ to be the likelihood of a multivariate gaussian variable.

By exploiting the link between the two contributions to the likelihoods $\mathcal{L}_{\mathbf{Y}}\{\mathbf{Y}_i|\boldsymbol{\gamma}_i\}$ and $\mathcal{L}_{(\widetilde{T},\widetilde{\eta})}\left\{(\widetilde{T}_i, \widetilde{\eta}_i)|\boldsymbol{\gamma}_i\right\}$ modeled by $\boldsymbol{\gamma}_i$, the maximization of the full likelihood $\mathcal{L}_{(\widetilde{T},\widetilde{\eta},\mathbf{Y})}$ enables to consistently estimate the vector of model parameters $\boldsymbol{\xi}$. Of particular interest, this approach enables the consistent estimation of numerous models for evaluating the association between a marker trajectory $Y(\cdot)$, that is an internal time-dependent covariate (Kalbfleisch and Prentice, 2002, Section 6.3), and an event-time and an event-type (T, η) (Rizopoulos, 2012; Tsiatis and Davidian, 2004).

Besides, let us note that two main approaches are often used for modeling the shared latent variable $\boldsymbol{\gamma}$. Either a continuous or a discrete distribution can be assumed for $\boldsymbol{\gamma}$, leading to the so-called "shared random effect" or "latent class" joint models respectively. We refer to Proust-Lima et al. (2012) for a recent overview contrasting the two approaches. In this work, we will make use of the joint latent class approach for our application, taking advantage of its computational assets and of our previous experiences, in particular for prediction purpose (Proust-Lima and Taylor, 2009; Proust-Lima et al., 2012). However, the following methodology can be applied whatever the kind of joint model fitted.

2.3 *Predictions from joint models*

There is currently an increasing interest on making prediction using the joint model framework (Henderson et al., 2002; Proust-Lima and Taylor, 2009; Rizopoulos, 2011, 2012; Proust-Lima et al., 2012; Taylor et al., 2013). Based on the model specification and the vector of estimated parameters $\widehat{\boldsymbol{\xi}}$, various subject-specific probabilities can be computed. Of particular

139

interest are the subject-specific probabilities of experiencing the main event within a time interval $(s, s+t]$ given the whole information available on the subject accumulated till landmark time s. Here, t denotes a fixed window of prediction whereas the varying landmark time s denotes the time at which predictions are made conditionally to the subject-specific history. For $0 \leqslant s < T_i$, we thus define the dynamic predictions of event 1 (main event) for subject i by

$$\pi_i(s,t) = \mathbb{P}_{\widehat{\boldsymbol{\xi}}}(s < T_i \leqslant s+t, \eta_i = 1 | T_i > s, \mathcal{Y}_i(s), \mathbf{X}_i),$$

where $\mathbb{P}_{\widehat{\boldsymbol{\xi}}}$ denotes the probability distribution parametrized by $\widehat{\boldsymbol{\xi}}$, $\mathbf{X}_i$ denotes the vector of baseline covariates and $\mathcal{Y}_i(s)$ denotes the entire information on the marker trajectory available for subject i by time s, i.e $\mathcal{Y}_i(s) = \{Y_{ij} : 0 \leqslant t_{ij} \leqslant s, j = 1, \ldots, n_i\}$. These predictions are dynamic in the sense that they change with increasing landmark time s and increasing available information $\{\mathcal{Y}_i(s), T_i > s\}$. From a practical point a view, $\pi_i(s,t), i = 1, \ldots, n$, can be computed from different approaches implying the use of Bayesian formulae, Monte-Carlo simulations or numerical integrations (Proust-Lima et al., 2012; Rizopoulos, 2012, Sec. 7.1). For the joint latent class approach used in our application, additional details are provided in Section 5 and in the web-appendix B.

3. Quantifying and comparing dynamic predictive accuracy

In the following, we assume to observe an i.i.d sample of n subjects $\{(\widetilde{T}_i, \Delta_i, \widetilde{\eta}_i, \pi_i(\cdot,\cdot)), i = 1, \ldots, n\}$, where $\pi_i(\cdot,\cdot)$ denotes a subject-i-specific prediction process computable for all landmark times s and prediction horizon t. Note that assuming an i.i.d sample implies that the prediction model (or any kind of prediction algorithm) used to compute $\pi_i(s,t)$ from subject specific characteristic has been fitted on an independent learning dataset. By convention and without loss of generality, we set $\pi_i(s,t) = 0$ for all subjects i that are no

longer at risk at landmark time s, and we assume that we are interested in the prediction of event $\eta = 1$ (main event).

3.1 *Definitions and meaning*

We propose an adaptation of the definitions of two well established prediction accuracy measures, the Area Under the ROC Curve (AUC) and the expected Brier score (BS), to simultaneously account for both (i) the dynamic nature of the predictions and (ii) the competing risks setting.

3.1.1 *Dynamic AUC with competing risks.* When developing a prediction model, it is desirable to quantify its predictive accuracy in term of discrimination. In practice, one would like a prediction tool that gives higher predicted risks of event for subjects who are more likely to experience the event, than for subjects who are less likely to experience the event. This is precisely what aims at measuring the AUC.

Formally, combining the definition of ROC curve for competing risks (Zheng et al., 2012; Blanche et al., 2013) and the one for dynamic prediction (Parast et al., 2012), we propose the following definition of the dynamic AUC, at landmark time s for a prediction horizon t:

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

where $D_i(s,t) = \mathbb{1}_{(s<T_i\leqslant s+t, \eta_i=1)}$. With this notation (where "D" is for "diseased"), for any subject i at risk at time s, $D_i(s,t) = 1$ when subject i experiences the main event within the time interval $(s, s+t]$, and $D_i(s,t) = 0$ when either subject i experiences a competing event within the time interval $(s, s+t]$ or subject i is event free at time $s+t$. Within the terminology of the ROC methodology, at fixed landmark time s and prediction horizon t, subject i at risk at time s is defined as a case when $D_i(s,t) = 1$ and a control when $D_i(s,t) = 0$ (Zheng et al., 2012; Blanche et al., 2013).

Note that an alternative definition of controls, excluding subjects experiencing the com-

141

peting event, could also be proposed following Saha and Heagerty (2010). However, we here prefer to define controls as all subjects that are not cases. The reason is that, using similar argument as in McIntosh and Pepe (2002), with our definition of controls it can be shown that the "true underlying" risk of event has the best (i.e the highest) of all possible AUC, which is a desirable property for prediction accuracy assessment.

3.1.2 *Dynamic Brier score with competing risks.* Combining the definition of the Expected Brier score for competing risks (Schoop et al., 2011) and the one for dynamic prediction (Schoop et al., 2008; Parast et al., 2012), we propose the following definition for the dynamic expected Brier score

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

which is a mean squared error. As recalled in Graf et al. (1999), the Brier score can be expressed as

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

where $\mathcal{H}(s) = \{\mathbf{X}, \mathcal{Y}(s), T > s\}$ denotes the history at time s used for computing the prediction $\pi(s,t)$. The second term in (1), named "inseparability", does not depend on the distribution of the predictions. The first term, named "imprecision" or "calibration", measures how close are the predictions to $\mathbb{E}\big[D(s,t)\big|\mathcal{H}(s)\big]$, i.e the "true underlying" risk of event in $(s, s+t]$ given $\mathcal{H}(s)$. Note that if the predictions were perfect, i.e if $\pi(s,t) = \mathbb{E}\big[D(s,t)\big|\mathcal{H}(s)\big]$, then the best (i.e the lowest) of all possible BS for predictions using $\mathcal{H}(s)$ would be reached and the "calibration" term would be equal to zero.

3.1.3 *Contrasts between AUC and BS.* The two predictive accuracy measures are interesting and complete each other. AUC is more convenient for communication purpose, as

142

it has a simple interpretation as a concordance index, does not depend on the main event cumulative incidence $\mathbb{P}(s < T \leqslant s+t, \eta = 1 | T > s)$, and so has an easily understandable scaling. By contrast, BS has the advantage of being a more complete predictive accuracy measure as it quantifies both calibration and discrimination. Indeed, it can be shown that the "inseparability" term in (1) depends on the inherent discrimination ability of the information $\mathcal{H}(s)$. However, as BS depends on the main event cumulative incidence, note that the scaling of this predictive accuracy measure changes with s and has to be interpreted carefully.

3.2 *Estimation*

In presence of censored data induced by lost-to-follow-up, for all subjects i censored within $(s, s+t]$ the indicator $D_i(s,t)$ cannot be computed and is thus unknown. To overcome such "missing data" issue, the Inverse Probability of Censoring Weighting (IPCW) technique has been successfully applied in lots of closely related settings in the last decade (Graf et al. (1999); Hung and Chiang (2010); Blanche et al. (2013); Parast et al. (2012) among others). Thus, we propose the IPCW estimators :

$$\widehat{AUC}(s,t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}_{(\pi_i(s,t)>\pi_j(s,t))} \widetilde{D}_i(s,t)\left(1-\widetilde{D}_j(s,t)\right)\widehat{W}_i(s,t)\widehat{W}_j(s,t)}{\sum_{i=1}^{n}\sum_{j=1}^{n} \widetilde{D}_i(s,t)\left(1-\widetilde{D}_j(s,t)\right)\widehat{W}_i(s,t)\widehat{W}_j(s,t)}$$

and

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

where the indicator $\widetilde{D}_i(s,t) = \mathbb{1}_{(s<\widetilde{T}_i\leqslant s+t, \widetilde{\eta}_i=1)}$ equals to 1 when subject i is known to have experienced the main event within $(s, s+t]$, and equals to 0 otherwise. To account for censoring, the weights are defined as

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

143

where $\widehat{G}(u)$ is the Kaplan-Meier estimator of survival function of the censoring at time u, i.e $\mathbb{P}(C > u)$, and $\forall u > s, \widehat{G}(u|s) = \widehat{G}(u)/\widehat{G}(s)$ estimates the conditional probability of not being censored at time u conditionally on being uncensored at time s.

An important remark is that these estimators are model free in the sense that there is no assumption about the correctness of the specification of the joint model used for computing $\pi_i(s,t), i = 1, \ldots, n$. This is of particular importance when we aim at comparing several predictions built from different joint models. Indeed, for a comparison purpose it would appear paradoxical to us to use "model-based" predictive accuracy estimator (as used in Proust-Lima and Taylor (2009) and Henderson et al. (2002)) since in any case several joint models cannot likely be simultaneously well-specified.

Moreover, most often it is not realistic to assume that a model is well-specified whereas it makes sense to claim that it is useful enough for prediction purpose. The estimation of its predictive accuracy in this case is thus of interest and must be model-free to be consistent.

3.3 *Large sample results*

For sake of brevity, let $\theta(s,t)$ denote either $AUC(s,t)$ or $BS(s,t)$ and $\widehat{\theta}(s,t)$ the corresponding IPCW estimator. Let $\tau_0 < \sup\left\{u : \mathbb{P}(\widetilde{T} > u) > 0\right\}$, $\tau_1(s) > \inf\left\{u : P(T \leqslant s+u, \eta = 1|T > s) > 0\right\}$ and $\tau_2(s) < \sup\left\{u : \mathbb{P}(\widetilde{T} > s+u) > 0\right\}$. Thus, $[0, \tau_0]$ represents a time window in which the probability of observing a subject at risk is non null, and $[\tau_1(s), \tau_2(s)]$ represents a time window in which there is both a non null probability of observing the main event in $(s, s+t]$ and a non null probability of observing someone at risk at time $s+t$.

The key result that both justifies and enables the computation of inference procedures for $\theta(s,t)$ is the following.

Lemma 1: *Assume that the censoring time C is independent of $(T, \eta, \pi(\cdot,\cdot))$, then $\forall s \in$*

144

$[0, \tau_0]$, $\forall t \in [\tau_1(s), \tau_2(s)]$:

$$\sqrt{n}\left(\widehat{\theta}(s,t) - \theta(s,t)\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IF_\theta(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i(s,t), s, t) + o_p(1)$$

where $\mathbb{E}\left[IF_\theta(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i(s,t), s, t)\right] = 0$ *and* $IF_\theta(\cdot)$ *is the influence function of the estimator which is detailed in the web-appendix A, at formulae (8-9) and (20-21).*

Proof. The proof follows similar arguments as the proof of Lemma 1 of Blanche et al. (2013) and of Theorem 1 of Hung and Chiang (2010) and is provided in the web-appendix A.

From the decomposition of the estimator $\widehat{\theta}(s,t)$ in a sum of asymptotically i.i.d terms in lemma 1, the central limit theorem induces the asymptotic normality of the estimator:

$$\sqrt{n}\left(\widehat{\theta}(s,t) - \theta(s,t)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2_{s,t}\right). \tag{2}$$

Using a simple plug-in estimator $\widehat{\mathrm{IF}}_\theta(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i(s,t), s, t)$ detailed in the web-appendix A.3, the variance $\sigma^2_{s,t}$ can be consistently estimated by the empirical estimator

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

It can be seen that $\sigma^2_{s,t}$ is inversely proportional to the probability of being at risk at landmark time s, i.e $S_{\tilde{T}}(s) = \mathbb{P}(\widetilde{T} > s)$. Therefore, note that the $\sqrt{n}$ convergence rate of $\widehat{\theta}(s,t)$ is thus consistent with the lemma 1 of Blanche et al. (2013) applied to any sub-sample of n_s subjects at risk at time s, since $\sqrt{n_s} = \sqrt{n} \times \sqrt{S_{\tilde{T}}(s)} + o_p(1)$.

3.4 *Confidence regions and tests*

Let $\pi^{(1)}(\cdot,\cdot)$ and $\pi^{(2)}(\cdot,\cdot)$ be two rival prediction processes computable for all landmark times s and prediction horizon t and respectively $\theta^{(1)}(\cdot,\cdot)$, and $\theta^{(2)}(\cdot,\cdot)$ their predictive accuracy processes. For sake of brevity, we only display confidence regions and tests for the difference $\triangle\theta(\cdot,t) = \theta^{(1)}(\cdot,t) - \theta^{(2)}(\cdot,t)$ at a given prediction horizon t.

145

3.4.1 *Inference procedures at a specific landmark time s.* Let $\triangle\widehat{\theta}(\cdot,t) = \widehat{\theta}^{(1)}(\cdot,t) - \widehat{\theta}^{(2)}(\cdot,t)$ define the IPCW estimator of $\triangle\theta(\cdot,t)$. As a consequence of lemma 1, for any landmark time s, we directly obtain the asymptotic $(1-\alpha)$-level pointwise confidence interval

$$\left\{\triangle\widehat{\theta}(s,t) \pm z_{1-\alpha/2}\frac{\widehat{\sigma}_{\triangle,s,t}}{\sqrt{n}}\right\} \tag{3}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the univariate standard normal distribution and

$$\widehat{\sigma}^2_{\triangle,s,t} = \frac{1}{n}\sum_{i=1}^{n}\triangle\widehat{\mathrm{IF}}_\theta(\widetilde{T}_i,\widetilde{\eta}_i,\pi_i^{(1)}(s,t),\pi_i^{(2)}(s,t),s,t)^2,$$

with $\triangle\widehat{\mathrm{IF}}_\theta(\widetilde{T}_i,\widetilde{\eta}_i,\pi_i^{(1)}(s,t),\pi_i^{(2)}(s,t),s,t) = \widehat{\mathrm{IF}}_\theta(\widetilde{T}_i,\widetilde{\eta}_i,\pi_i^{(1)}(s,t),s,t) - \widehat{\mathrm{IF}}_\theta(\widetilde{T}_i,\widetilde{\eta}_i,\pi_i^{(2)}(s,t),s,t)$. Instead of observing whether or not zero is contained within the confidence interval, for all landmark times s and prediction horizons t satisfying condition of lemma 1, a test for comparing the two prediction accuracy measures $\theta^{(1)}(s,t)$ and $\theta^{(2)}(s,t)$ can be equivalently derived. Indeed, under $\mathcal{H}_0^{(s)} : \triangle\theta(s,t) = 0$, as $n \to \infty$ then

$$\frac{\triangle\widehat{\theta}(s,t)}{\widehat{\sigma}_{\triangle,s,t}/\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

3.4.2 *Simultaneous inference procedures over a set of landmark time s.* Pointwise confidence intervals and tests are of interest when aiming at quantifying and comparing predictive accuracy at a specific landmark time s. However, within the joint modeling framework that enables computations of dynamic predictions for a set (or an interval) of landmark times s, denoted $\mathcal{S}$, it is also desirable to quantify the overall predictive accuracy over $\mathcal{S}$ for a fixed prediction horizon t. A simultaneous confidence band is then useful to estimate the variability of the estimation of the curve $\left\{\left(s,\triangle\theta(s,t)\right), s\in\mathcal{S}\right\}$. A $(1-\alpha)$-simultaneous confidence band for this curve is defined as a region containing this curve with probability level $1-\alpha$. By definition, a simultaneous confidence band is larger that the band of pointwise confidence intervals. We propose to compute an asymptotic $(1-\alpha)$-simultaneous confidence band by :

$$\left\{\triangle\widehat{\theta}(s,t) \pm \widehat{q}_{1-\alpha}^{(\mathcal{S},t)}\frac{\widehat{\sigma}_{\triangle,s,t}}{\sqrt{n}}\right\}, \quad s\in\mathcal{S}, \tag{4}$$

146

where $\widehat{q}_{1-\alpha}^{(\mathcal{S},t)}$ is a $(1-\alpha)$-quantile that depends on both the landmark time set $\mathcal{S}$ and the prediction horizon t. By use of lemma 1, the following simulation-based technique is applied to estimate these quantiles by properly accounting for the autocorrelation of the process $\triangle\widehat{\theta}(\cdot,t)$:

(1) For $b = 1,\ldots,B$, with B large enough, say $B = 4000$ for instance:

(a) Generate a random sample $(\omega_1^b,\ldots,\omega_n^b)$ from n independent standard normal distributions.

(b) Using the estimator $\widehat{\mathrm{IF}}_\theta(\cdot)$ detailed in web-appendix A.3, compute :

$$\Upsilon^b = \sup_{s\in\mathcal{S}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\omega_i^b\frac{\triangle\widehat{\mathrm{IF}}_\theta(\widetilde{T}_i,\widetilde{\eta}_i,\pi_i^{(1)}(s,t),\pi_i^{(2)}(s,t),s,t)}{\widehat{\sigma}_{\triangle,s,t}}\right|$$

(2) Compute $\widehat{q}_{1-\alpha}^{(\mathcal{S},t)}$ as the $100(1-\alpha)$th percentile of $\left\{\Upsilon^1,\ldots,\Upsilon^B\right\}$.

Note that when $\mathcal{S}$ reduces to any singleton $\{s\}$ then $\widehat{q}_{1-\alpha}^{(\mathcal{S},t)}$ is close enough from $z_{1-\alpha/2}$ as soon as B is large enough. By contrast, larger is the set $\mathcal{S}$ and weaker is the autocorrelation of the process $\widehat{\theta}(\cdot,t)$ and larger will be $\widehat{q}_{1-\alpha}^{(\mathcal{S},t)}$ compared to $z_{1-\alpha/2}$. Based on the conditional multiplier central limit theorem (Van der Vaart and Wellner, 1996), this kind of simulation approach has already been successfully applied in related settings, by Lin et al. (1994) and Martinussen and Scheike (2006) among others.

The confidence band of the curve $\left\{\left(s,\triangle\theta(s,t)\right),s\in\mathcal{S}\right\}$ can either be used :

- to test whether or not the two dynamic prediction accuracy curves are different, i.e $\mathcal{H}_0$: $\forall s\in\mathcal{S}\ \triangle\,\theta(s,t)=0$, by observing whether or not the zero line is contained within the band;
- or to assert that one dynamic prediction accuracy is significantly uniformly higher than another in the set of landmark times, i.e $\forall s\in\mathcal{S}\ \triangle\,\theta(s,t)>0$, by observing whether or not the confidence band overlaps the zero line.

3.4.3 *Remarks.* Similarly, confidence regions can be derived for $\theta(s,t)$ instead of $\triangle\theta(s,t)$. The previous methodology applies by replacing $\triangle\widehat{\theta}(s,t)$ by $\widehat{\theta}(s,t)$, $\triangle\widehat{\mathrm{IF}}_\theta(\cdot)$ by $\widehat{\mathrm{IF}}_\theta(\cdot)$ and $\widehat{\sigma}_{\triangle,s,t}$ by $\widehat{\sigma}_{s,t}$ in the formulae (3) and (4) and in the algorithm to compute $\widehat{q}_{1-\alpha}^{(\mathcal{S},t)}$.
Note that there is no equivalence between the confidence regions for predictive accuracy overlapping and the significance of the differences, in particular because of paired data. Confidence regions for $\theta(s,t)$ and $\triangle\theta(s,t)$ either at a single landmark time s or simultaneously over $s \in \mathcal{S}$ are thus complementary to each other.

4. Simulation study

The objective of the simulation study was to check the finite sample behavior of the inference procedures, in particular those for simultaneous inference procedures over a set of landmark times. Scenarii choices were partly guided by the cognitive aging cohort data described in Section 5.

4.1 *Simulation scenarii*

We considered a setting with two competing risks and two longitudinal markers $Y^{(l)}(\cdot)$, $l = 1, 2$, both associated with the two competing events. For simplicity and to gain insights, we generated the data from simple joint latent class models that consider a discrete shared latent variable γ. We generated only two classes for γ from a Bernoulli distribution with parameter $p = 0.5$. Class-specific constant hazard functions were chosen for cause-specific hazards of the two events and the following linear mixed models were considered for the two maker trajectories $Y^{(l)}(\cdot)$, $l = 1, 2$:

$$Y_i^{(l)}(t_{ij})|_{\gamma_i=g} = (\beta_{0g}^{(l)} + b_{i0}^{(l)}) + (\beta_{1g}^{(l)} + b_{i1}^{(l)}) \times t_{ij} + \varepsilon_i^{(l)}(t_{ij}),$$

for class $g = 1, 2$, with $(b_{i0}^{(l)}, b_{i1}^{(l)}) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{b}}^{(l)})$ independent of γ_i, $\mathbf{\Sigma}_b^{(l)} = \mathrm{diag}\left({\sigma_{b_0}^{(l)}}^2, {\sigma_{b_1}^{(l)}}^2\right)$ and an independent noise $\varepsilon_i^{(l)}(t_{ij}) \sim \mathcal{N}(0, {\sigma_\varepsilon^{(l)}}^2)$.

We generated a 10-year follow-up for the longitudinal data with repeated measurements

148

every year. Independent censoring was generated with an exponential distribution. The landmark times were chosen equal to the measurement times, i.e $\mathcal{S} = \{0, 1, \ldots, 10\}$ and the prediction horizon was set to $t = 5$ years. Parameters for survival data generation were chosen such that the proportions of observed main event, competing event and censored observations within the time interval $(s, s+t]$ and observed event free at time $s+t$ were respectively around 20%, 10%, 20% and 50% for all landmark times s.

For each time s, two rival dynamic predictions $\pi_i^{(l)}(s,t)$, $l = 1, 2$, were computed from two joint latent class models, using formulae of web-appendicies B and C, each using one of the two sets of repeated marker measurements $\mathcal{Y}_i^{(l)}(s)$, $l = 1, 2$. Eight scenarii illustrated in Figure 1 were investigated with sample sizes $n = 2000$ and 3000. Changes from one scenario to another only consisted from changes in the marker trajectories generation. Two scenarii were generated under the null hypothesis asserting that the dynamic predictive accuracy curves were equal, by generating $Y_i^{(1)}(\cdot)$ and $Y_i^{(2)}(\cdot)$ from the same distribution. The two scenarii differed by the distance between the two class-specific slopes $\beta_{11}^{(l)}$ and $\beta_{12}^{(l)}$, $l = 1, 2$ (H0.1 and H0.2). Six scenarii were generated under the alternative hypothesis by breaking the symmetry between the distribution of the two marker trajectories $Y_i^{(1)}(\cdot)$ and $Y_i^{(2)}(\cdot)$. Three of them were generated by increasing the difference $\beta_{11}^{(1)} - \beta_{12}^{(1)}$ between the class-specific slopes for marker 1 (H1.1.1, H1.1.2, H1.1.3). Two scenarii were generated by making the standard deviations $\sigma_{b_0}^{(l)}, \sigma_{b_1}^{(l)}$ and $\sigma_{\varepsilon}^{(l)}$ 1.5 or 2 times larger for $l = 1$ compared to $l = 2$ (H1.2.1, H1.2.2). Contrary to the previous scenarii, one additional scenario was generated such that the two class-specific trajectories of one of the two markers crossed (H1.3). Parameter values for data generation are given in web-appendix C.

[Figure 1 about here.]

4.2 *Simulation results*

For each scenario, we ran 1000 simulations and we estimated curves of $AUC(s,t)$ and $BS(s,t)$ versus $s \in \mathcal{S}$ for the two dynamic predictions $\pi^{(l)}(s,t)$, $l = 1, 2$, and the curves of differences $\triangle AUC(s,t)$ and $\triangle BS(s,t)$ versus $s \in \mathcal{S}$. For all curves, pointwise confidence intervals and simultaneous confidence bands at α-level equal to 5% were estimated as proposed in Section 3.2 (with $B = 4000$). Pointwise comparison tests with $\mathcal{H}_0^{(s)} : \triangle\theta(s,t) = 0$ at each landmark time $s \in \mathcal{S}$ as well as simultaneous tests with $\mathcal{H}_0 : \forall s \in \mathcal{S} \triangle \theta(s,t) = 0$ were computed.

For sake of brevity, we mainly present here the empirical coverage probabilities of simultaneous confidence bands for the two dynamic predictions and the type one error and power of the test of $\mathcal{H}_0 : \forall s \in \mathcal{S} \triangle \theta(s,t) = 0$ (Table 1). Figure 2 illustrates the power of the tests of $\mathcal{H}_0^{(s)} : \triangle\theta(s,t) = 0$ for each landmark time $s \in \mathcal{S}$ and Figure 3 displays the mean curves of $AUC(s,t)$ versus s for $\pi^{(2)}(s,t)$ and the coverage probabilities of the pointwise confidence intervals. Additional details about simulation results for pointwise inference are moved in the web-appendix D.

[Table 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

As expected, for both BS and AUC empirical coverage of confidence regions and type one error are close to the chosen 95% and 5%-levels (Table 1). The simulation study thus confirms that the asymptotic inference procedures behave well under our data generated scenarii. The two sample sizes $n = 2000$ and $n = 3000$ illustrate the gain in power when the sample size increases.

The BS comparison tests seem more powerful than the AUC one in these simulations. However, this difference in power may largely depend on the simulation setting and is not

confirmed in our application displayed in Section 5. It could come from the fact that, for sake of simplicity, we have generated well calibrated dynamic predictions.

Note that Figue 3 displays an increase in $AUC(s,t)$ for early increasing s that precedes a decrease. Even if the models are accumulating more and more information when s increases, and are thus more and more accurate, this decrease in discrimination is realistic and rational. Indeed, the decrease essentially reflects the fact that the selection mechanism for the at risk population makes it naturally more and more homogeneous.

Finally, Figure 2 displays an increase in power with increasing s that precedes a decrease under the alternatives. This decrease may also be due to the selection mechanism, to the decrease in the difference between the two dynamic prediction models with increasing s, and to the decreasing number of subjects at risk.

5. Application to prediction of dementia onset in the elderly

5.1 *Objective*

The objective of this analysis is the quantification and the comparison of 5-year dynamic predictions of dementia built from two different joint latent class models. The first one is based on repeated measurements of the Mini Mental Score Examination (MMSE) (Folstein et al., 1975), the second one is based on repeated measurements of the Isaacs Set Test shortened at 15 seconds (IST) (Isaacs and Kennie, 1973). The MMSE is a sum-score evaluating various dimensions of cognition such as memory, calculation and language, and is a widely used test for dementia screening or evaluating cognitive impairment in the elderly. The IST shortened at 15 seconds mainly evaluates speed of verbal production and has been shown to be particularly sensitive to small changes in cognition in the elderly (Proust-Lima et al., 2007). Roughly, IST consists in asking a subject to give the longest list of words belonging to 4 specific semantic categories in 15 seconds (truncated at 10 words by categories).

151

The choice of a 5-year prediction horizon was guided by the clinical perspective of targeting subjects at high risk of dementia, who could benefit from prevention programs or preventive treatments.

5.2 *Data from Paquid cohort (training cohort) and 3C cohort (validation cohort)*

Paquid and Three-City (3C) are two French population-based studies that aimed at studying cognitive aging and dementia onset. They included 3777 and 9294 subjects aged 65 years and older and living at home at enrollment in 1988 and 1999, respectively. In both studies, subjects were seen approximately every two or three years. Each visit included evaluations of their cognitive abilities through a battery of cognitive tests and a standardized diagnosis of dementia (Dartigues et al., 1992; The 3C Study Group, 2003).

Subjects demented, blind, deaf or confined to bed at the initial visit were excluded. At time of analyses, 10-year follow-up data were not available for subjects from Dijon city. Only subjects from Bordeaux and Montpellier cities were thus included in our 3C sample. Final samples from Paquid and 3C data include $n = 2970$ and $n = 3880$ subjects with follow-up durations of 20 and 10 years, respectively.

Time-to-dementia was computed as the time elapsed between study entry and the mid-point between the time at the visit of diagnosis and the time at the last visit without dementia. Subjects who died without a dementia diagnosis were considered as free of dementia at their death time if the last visit was less than two years before the death and were considered as censored at the last visit if duration between the last negative dementia diagnosis and death was longer.

The two samples are described in the web-Table I. Age at inclusion (about 74 years) and gender frequencies (about 40% of male) are similar in the two cohorts. Primary education level is higher in the 3C cohort (32.5% of high education) than in the Paquid cohort (8.6% of high education) and baseline cognitive scores are slightly higher in 3C than in Paquid.

These differences mainly come from the fact that 3C is a more recent cohort than Paquid and only includes subjects living in large cities.

5.3 *Fitting a joint latent class model using the Paquid (training) sample*

After preliminary analyzes, we chose the same model specification for the two joint latent class models based either on IST or on MMSE. Paquid data was used to fit the two corresponding models. Based on BIC criteria, we retained models with the three classes for both IST and MMSE. Conditionally on each latent class $\gamma_i = g$, $g = 1, 2, 3$, we modeled the subject i cognitive test trajectory by

$$\begin{aligned} Y_i(t_{ij})|_{\gamma_i=g} =& \beta_0 + \beta_{0,age}\texttt{AGE}_i + \beta_{0,educ}\texttt{EDUC}_i + \beta_{0,learn}\mathbb{1}_{(t_{ij}=0)} + b_{i0|\gamma_i=g} \\ &+ \left(\beta_{1g} + \beta_{1,age}\texttt{AGE}_i + b_{i1|\gamma_i=g}\right) \times t_{ij} + \left(\beta_{2g} + \beta_{2,age}\texttt{AGE}_i + b_{i2|\gamma_i=g}\right) \times t_{ij}^2 + \varepsilon_i(t_{ij}), \end{aligned}$$

with $(b_{i0|\gamma_i=g}, b_{i1|\gamma_i=g}, b_{i2|\gamma_i=g}) \sim \mathcal{N}(\mathbf{0}, \sigma_g^2\mathbf{B})$, where $\mathbf{B}$ is unstructured, and an independent noise $\varepsilon_i(t_{ij}) \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. Binary variable `EDUC` and continuous variable `AGE` represent the primary education level and the age at baseline. Note that the term $\beta_{0,learn}\mathbb{1}_{(t_{ij}=0)}$ was added to account for the learning effect after the first interview (Jacqmin-Gadda et al., 1997) and that the variables `EDUC` and `AGE` have a common effect over classes.

For cause-specific hazards of dementia (denoted by event-type 1) and dementia-free death (denoted by event-type 2), we respectively modeled

$$\begin{aligned} &\lambda_{i,1}(t|\gamma_i = g) = \lambda_{01,g}(t) \exp\left(\alpha_{11,g}\texttt{AGE}_i + \alpha_{21,g}\texttt{EDUC}_i\right) \\ \text{and} \quad &\lambda_{i,2}(t|\gamma_i = g) = \lambda_{02,g}(t) \exp\left(\alpha_{12,g}\texttt{AGE}_i + \alpha_{22,g}\texttt{EDUC}_i + \alpha_{32,g}\texttt{SEX}_i\right). \end{aligned}$$

Class-specific baseline hazard of both event types $\lambda_{0k,g}(t)$, $k = 1, 2$, $g = 1, 2, 3$, were parametrized by Weibull hazard functions. We directly used IST repeated measurements as input in the first joint model, i.e as the $Y_i(t_{ij})$. However, in order to account for ceiling effects and curvilinearity of MMSE, we used a recently proposed monotonic transformation

153

of the MMSE (Philipps et al., 2013) as input of the second joint model. Parameters of the two models $\boldsymbol{\xi}^{(l)}$, $l = 1, 2$, were estimated by maximizing the corresponding log-likelihoods.

5.4 *Building predictions for the subjects of the 3C (validation) sample*

Using the joint latent class models fitted on Paquid data, we computed subject-specific predictions at each landmark times s = 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5 and 4 years with a prediction window of $t = 5$ years for all subjects of the 3C sample. By 3C study design, dynamic predictions use one cognitive test measurement at $s = 0$ and up to three repeated measurements when s increases till $s = 4$. For $l = 1, 2$, using estimated parameters $\widehat{\boldsymbol{\xi}}^{(l)}$, baseline covariates $\mathbf{X} = (\texttt{AGE}, \texttt{SEX}, \texttt{EDUC})$, and MMSE ($l = 1$) or IST ($l = 2$) repeated measurements collected by time s, denoted by $\mathcal{Y}_i^{(l)}(s)$, dynamic subject-specific predictions for subjects of the 3C sample were computed as:

$$\begin{aligned}\pi_i^{(l)}(s,t) &= \mathbb{P}_{\widehat{\boldsymbol{\xi}}^{(l)}}(s < T_i \leqslant s+t, \eta_i = 1 | T > s, \mathcal{Y}_i^{(l)}(s), \mathbf{X}_i) \\ &= \frac{\sum_{g=1}^{3} \mathbb{P}_{\widehat{\boldsymbol{\xi}}^{(l)}}(\gamma_i = g) \mathcal{L}_{g,\widehat{\boldsymbol{\xi}}^{(l)}}(\mathcal{Y}_i^{(l)}(s)|\mathbf{X}_i) \left\{ F_{1g,\widehat{\boldsymbol{\xi}}^{(l)}}(s+t|\mathbf{X}_i) - F_{1g,\widehat{\boldsymbol{\xi}}^{(l)}}(s|\mathbf{X}_i) \right\}}{\sum_{g=1}^{3} \mathbb{P}_{\widehat{\boldsymbol{\xi}}^{(l)}}(\gamma_i = g) \mathcal{L}_{g,\widehat{\boldsymbol{\xi}}^{(l)}}(\mathcal{Y}_i^{(l)}(s)|\mathbf{X}_i) S_{g,\widehat{\boldsymbol{\xi}}^{(l)}}(s|\mathbf{X}_i)}\end{aligned}$$

where $\mathcal{L}_{g,\widehat{\boldsymbol{\xi}}^{(l)}}(\mathcal{Y}_i^{(l)}(s)|\mathbf{X}_i)$, $F_{1g,\widehat{\boldsymbol{\xi}}^{(l)}}(s|\mathbf{X}_i)$ and $S_{g,\widehat{\boldsymbol{\xi}}^{(l)}}(s|\mathbf{X}_i)$ are respectively the density of $\mathcal{Y}_i^{(l)}(s)$, the cumulative incidence of dementia at time s and the all-event-free survival function at time s conditionally on $\mathbf{X}_i$ and on the latent class $\gamma_i = g$ (detailed formulae are given in web-appendix B).

5.5 *Quantifying and comparing predictive accuracy on the (validation) 3C cohort*

Figure 4 shows that estimated AUCs for both curves corresponding to the two prediction models are high (range from 0.79 to 0.90). This suggests that the proposed dynamic prediction tools based on joint latent class models using repeated cognitive test measurements have a good predictive accuracy in term of discrimination.

[Figure 4 about here.]

154

Although the proportions of observed subjects getting dementia within each time window $(s, s+t]$ is low (range from 4.7% to 7.6%), the numbers are large enough (range from 182 to 227) due to the large sample size $n = 3880$. This makes confidence intervals and confidence bands relatively narrow. Pointwise confidence intervals of the differences of AUCs do not overlap the zero line. Consequently, for each landmark time $s = 0, 0.5, \ldots, 4$ the null hypotheses $\mathcal{H}_0^{(s)} : \triangle AUC(s,t) = 0$ are rejected with significance level of $\alpha = 5\%$. As the simultaneous confidence bands do not contain the zero line then the simultaneous null hypothesis $\mathcal{H}_0 : \forall s \in \mathcal{S} \triangle AUC(s,t) = 0$ is also significantly rejected. More importantly, as the simultaneous confidence band of the differences of AUCs does not even overlap the zero line, we can also assert with confidence level of of 95% that predictions from IST have a uniformly better prediction accuracy in term of AUC over all landmark times $s \in \mathcal{S} = \{0, 0.5, \ldots, 4\}$ than those from MMSE, i.e $\mathbb{P}\left(\forall s \in \mathcal{S}\ AUC^{(IST)}(s,t) > AUC^{(MMSE)}(s,t)\right) \geqslant 95\%$. Although not significant, BS for predictions from IST are also estimated lower than those from MMSE (the lower the better).

Finally, as similarly discussed for simulation results in Section 4.2, the decreasing trends for $AUC(s,t)$ with increasing s is probably the consequence of a selection process that makes the at risk population more and more homogeneous when s increases.

6. Discussion

We proposed two dynamic predictive accuracy curves to quantify and compare different dynamic risk prediction tools derived from the joint modeling framework. The dynamic AUC curve is easy to interpret and quantifies discrimination abilities. The dynamic BS curve allows to additionally compare calibration. Estimators dealing with competing risk and censored data were proposed. Asymptotic results were established and we derived pointwise and simultaneous confidence regions as well as comparison tests that performed well in the simulation study. Besides, by applying the proposed methodology to cohort data, we were

155

able to show that: (i) dynamic predictions of dementia using repeated measurements of cognitive tests have a high predictive accuracy in term of discrimination and that (ii) the discrimination is uniformly better for the IST cognitive test than for the MMSE.

Although we focused on the competing risks setting, note that the proposed predictive accuracy definitions, estimators and inferences procedures simplify and are still relevant when there is only one type of event. In addition, in the more usual survival setting, the proportions of observed events in each prediction window is often higher and so better are the finite sample behaviors of the asymptotic procedures (simulation results not shown).

While pointwise simulation results about AUC confirmed those of Blanche et al. (2013) applied to the at risk population at each landmark time, to our knowledge, the corresponding pointwise inference procedures for BS have never been proposed. In addition to the investigation of confidence bands for dynamic AUC and BS, this work thus also contributed to introduce innovative pointwise inference procedures for BS in the competing risk setting.

In practice, a specific risk prediction threshold, say 20% for instance, is sometimes of clinical interest for decision making. For such cases, it would be complementary to derive inference procedures for dynamic sensitivity, specificity and positive and negative predictive values following similar lines than in Section 3.2. Explained variation criteria could also be derived, by comparing the BS of any prediction tool with the BS of a "null" prediction tool ignoring any subject-specific information (Graf et al., 1999). For a comparison purpose, comparing a difference in explained variation would however be equivalent to comparing a difference in BS.

Being model free, our inference procedures do not assume any correct model specification and could also be applied beyond the joint modeling framework, to the rival landmarking approach for example (van Houwelingen and Putter, 2012). Comparison of predictions from joint modeling and landmarking is appealing and could also be investigated using our

156

methodology. However, as the two approaches do not deal equivalently with random marker measurement time and missing data as it happens in our cohort data, caution is needed and more insights would be required for a fair comparison.

7. Acknowledgments

Computer time for this study was provided by the computing facilities MCIA (Mésocentre de Calcul Intensif Aquitain) of the Université de Bordeaux and of the Université de Pau et des Pays de l'Adour. This work was partly funded by a grant from France Alzheimer awarded to Hélène Jacqmin-Gadda in 2009. The Paquid study is funded by IPSEN and Novartis laboratories. The 3C Study supports are listed on the Study website (www.three-city-study.com).

Supplementary Materials

Web-appendix is available with this paper at XXX.

References

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,** 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,** 492–498.

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, in press* .

Dartigues, J., Gagnon, M., Barberger-Gateau, P., Letenneur, L., Commenges, D., Sauvel,

157

C., Michel, P., and Salamon, R. (1992). The Paquid epidemiological program on brain ageing. *Neuroepidemiology* **11,** 14–18.

Folstein, M., Folstein, S., and P, M. (1975). "Mini-Mental State." A practical method for grading the cognitive decline state of patients for the clinician. *Journal of Psychiatric Research* **12,** 189–198.

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

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

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

Isaacs, B. and Kennie, A. (1973). The set test as an aid to the detection of dementia in old people. *The British Journal of Psychiatry* **123,** 467–470.

Jacqmin-Gadda, H., Fabrigoule, C., Commenges, D., and Dartigues, J.-F. (1997). A 5-year longitudinal study of the mini-mental state examination in normal aging. *American Journal of Epidemiology* **145,** 498–506.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The statistical analysis of failure time data.* Wiley-Interscience.

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

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

McIntosh, M. and Pepe, M. (2002). Combining several screening tests: optimality of the risk score. *Biometrics* **58,** 657–664.

Parast, L., Cheng, S.-C., and Cai, T. (2012). Landmark Prediction of Long-Term Survival

158
Incorporing Short-Term Event Time Information. *Journal of the American Statistical Association* **107,** 1492–1501.

Philipps, V., Amieva, H., Andrieu, S., Dufouil C.and Berr, C., Dartigues, J.-F., Jacqmin-Gadda, H., and C., P.-L. (2013). Normalized MMSE for assessing cognitive change in population-based aging studies. *submitted* .

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,** 344–350.

Proust-Lima, C., Mbéry, S., Taylor, J., and Jacqmin-Gadda, H. (2012). Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research* .

Proust-Lima, C. and Taylor, J. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: a joint modeling approach. *Biostatistics* **10,** 535–549.

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

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-event Data: With Applications in R*, volume 6. Chapman & Hall.

Saha, P. and Heagerty, P. (2010). Time-dependent predictive accuracy in the presence of competing risks. *Biometrics* **66,** 999–1011.

Schoop, R., Beyersmann, J., Schumacher, M., and Binder, H. (2011). Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal* **53,** 88–112.

Schoop, R., Graf, E., and Schumacher, M. (2008). Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics*

64, 603–610.

Taylor, J., Park, Y., Ankerst, D., 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* .

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.

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

Van der Vaart, A. and Wellner, J. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer.

van Houwelingen, H. and Putter, H. (2012). *Dynamic Prediction in Clinical Survival Analysis*. CRC Press.

Zheng, Y., Cai, T., Jin, Y., and Feng, Z. (2012). Evaluating Prognostic Accuracy of Biomarkers under Competing Risk. *Biometrics* **68,** 388–396.

160

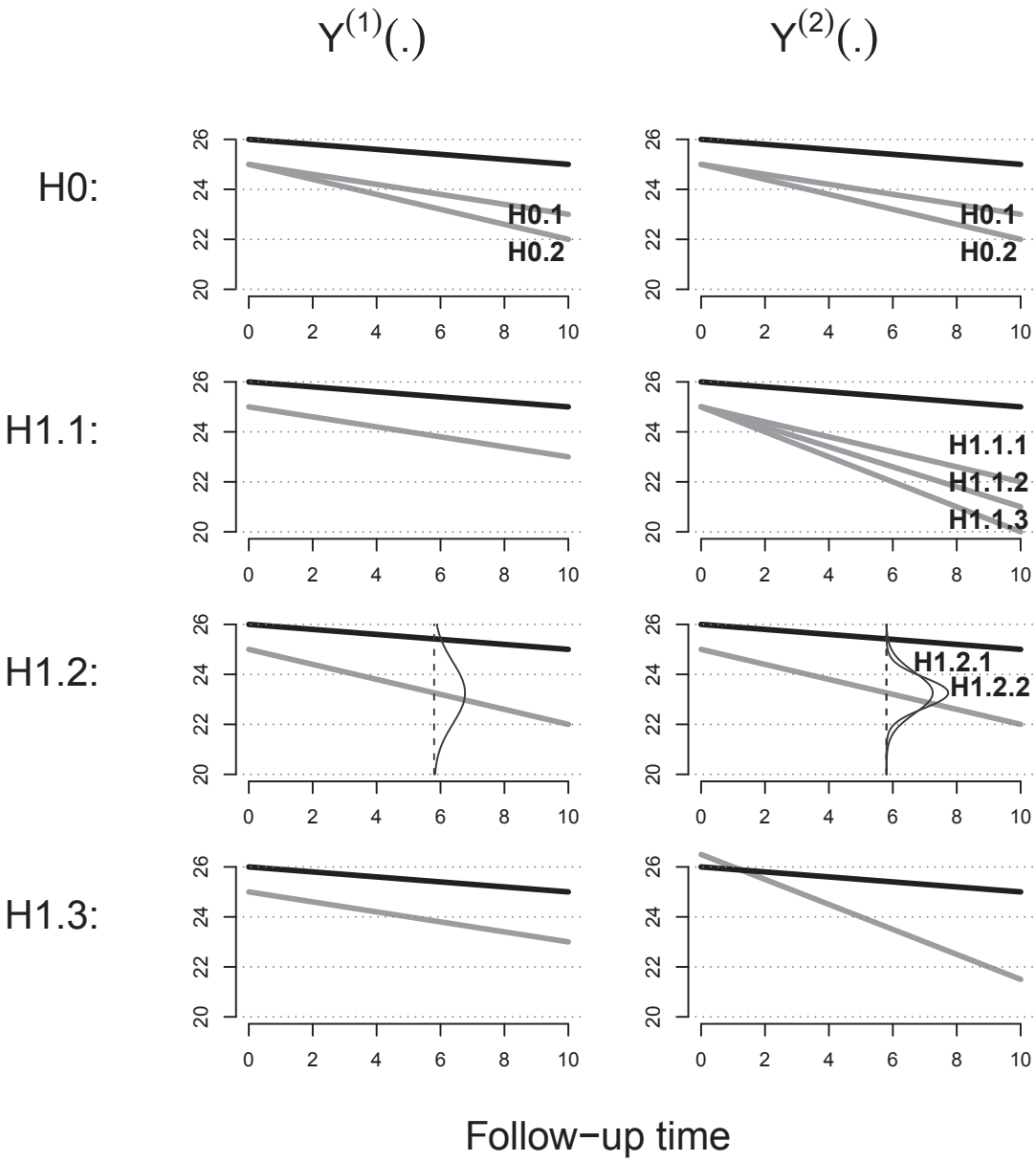


Figure 1. Comparison of the marker trajectories according to each simulation scenarii. Marker trajectories $Y_i^{(1)}(\cdot)$ and $Y_i^{(2)}(\cdot)$ conditionally to latent class $\gamma = 1$ (black) and $\gamma = 2$ (grey) are displayed in solid line. In scenario H1.2, gaussian densities of $Y^{(l)}(t = 6)|_{\gamma=2}$, $l = 1, 2$, display the impact of dividing the standard errors by 1.5 or 2.

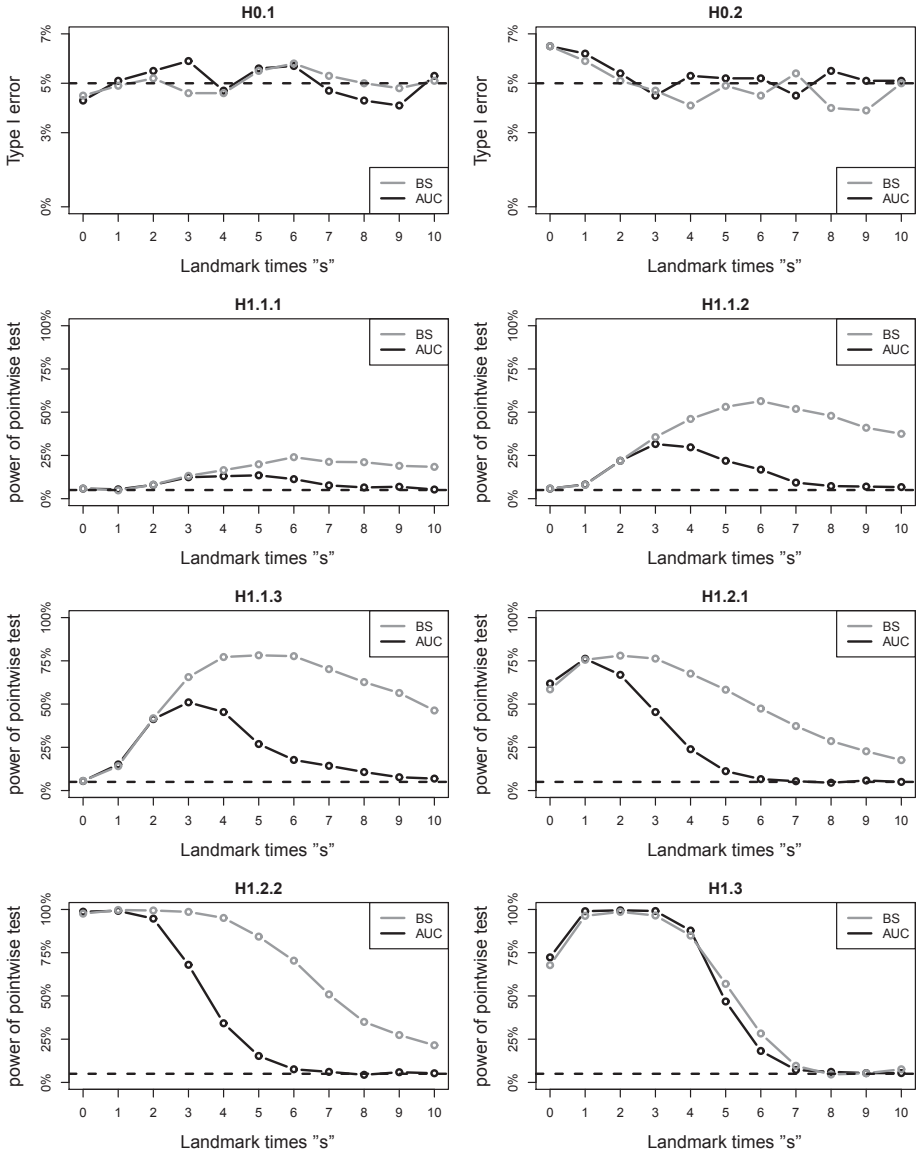


Figure 2. Type I error and power of pointwise tests for the difference in $AUC(s, t = 5)$ or $BS(s, t = 5)$ between the dynamic predictions $\pi^{(l)}(s, t = 5)$, $l = 1, 2$, when landmark time s changes, according to each simulation scenario. $n = 2000$ subjects, 1000 runs.

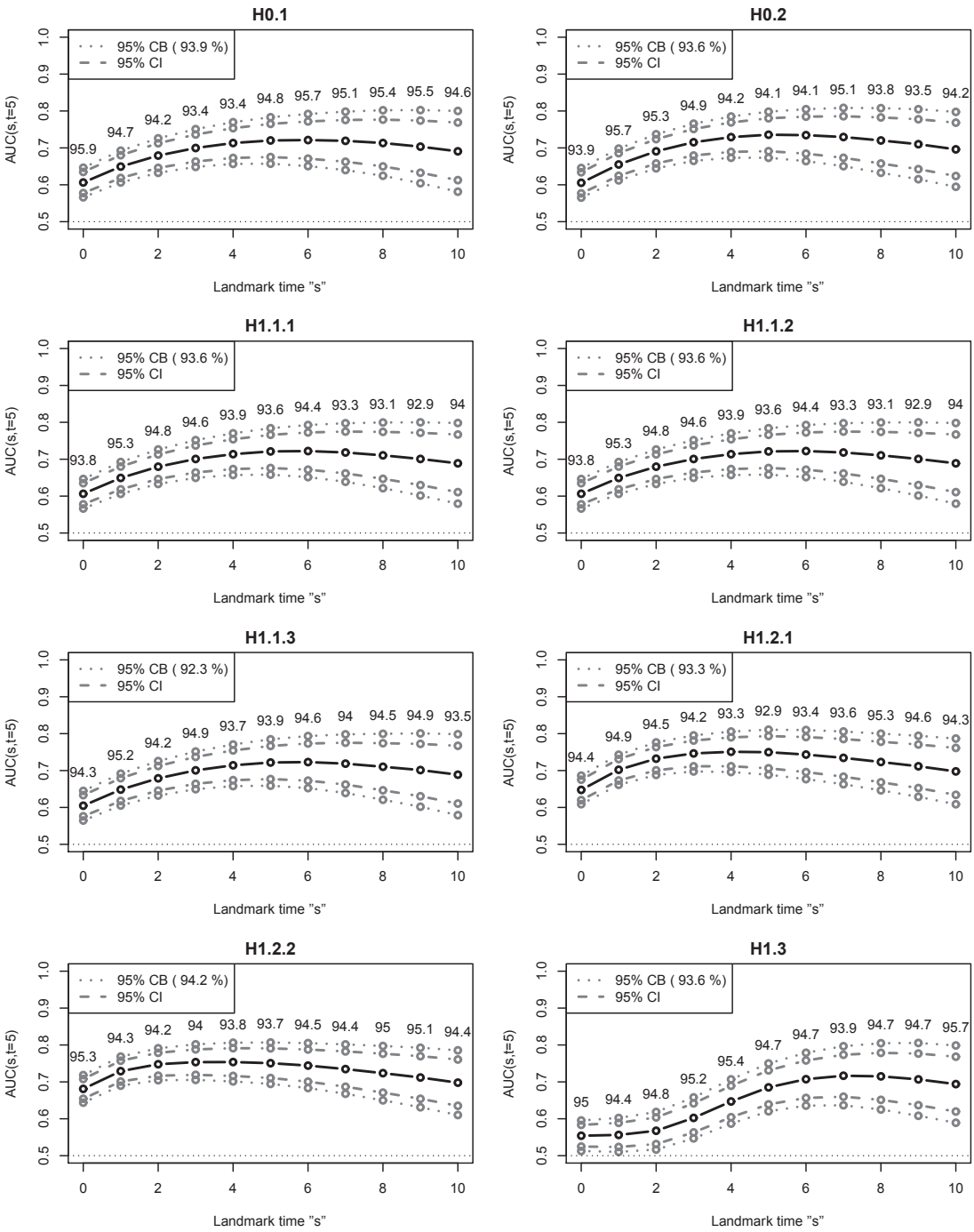


Figure 3. Average estimates of the curve $AUC(s, t = 5)$ versus s for $\pi^{(2)}(s, t)$, $n = 2000$. Average 95% pointwise confidence intervals (CI) and 95% simultaneous confidence band (CB) are plotted from the average standard errors and quantiles over the 1000 runs. The empirical coverage probabilities of the 95% CB is given in the legends and of the 95% CI for each time s are displayed in the plot.

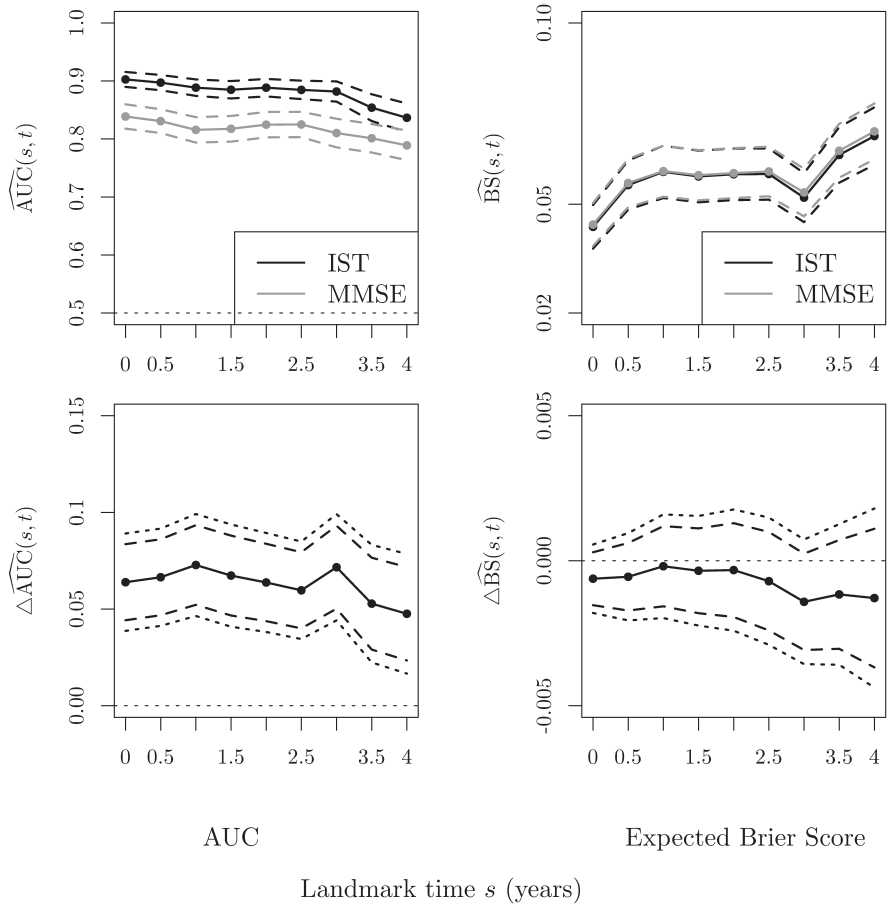


Figure 4. Comparison of predictive accuracy of the two predicted risks of dementia within time window $(s, s+t)$ when $s = \{0, 0.5, 1, 1.5, \ldots, 4\}$ and $t = 5$ years. 95% pointwise confidence intervals are displayed in dashed lines, 95% simultaneous confidence bands in dotted lines. 3C data, $n = 3880$ subjects.

	Coverage rates for confidence bands				Type I error or power	
	$AUC_{\pi^{(1)}}$	$BS_{\pi^{(1)}}$	$AUC_{\pi^{(2)}}$	$BS_{\pi^{(2)}}$	$\triangle AUC$	$\triangle BS$
$n = 2000$						
H0.1	94.2	93.5	93.9	93.9	5.7	5.7
H0.2	92.8	93.1	93.6	94.0	5.0	5.2
H1.1.1	94.7	95.5	93.6	95.9	10.5	21.6
H1.1.2	95.3	95.2	93.6	95.9	21.9	56.9
H1.1.3	93.9	93.4	92.3	93.4	39.5	82.7
H1.2.3	93.7	94.0	93.3	93.4	68.4	85.1
H1.2.1	93.7	94.0	94.2	92.9	98.5	99.9
H1.3.3	94.1	93.2	93.6	94.7	99.3	97.7
$n = 3000$						
H0.1	94.5	93.9	94.0	94.2	4.2	3.7
H0.2	93.0	95.4	93.7	94.3	4.7	4.7
H1.1.1	94.3	93.9	94.4	95.4	14.5	34.4
H1.1.2	94.1	94.3	94.4	95.4	34.0	80.7
H1.1.3	94.5	94.8	93.6	94.4	59.3	95.8
H1.2.3	94.7	94.9	93.7	93.9	85.6	95.6
H1.2.1	94.7	94.9	93.4	93.9	99.8	100.0
H1.3.3	92.9	94.4	93.2	94.2	100.0	99.9

Table 1

Simulation results. Empirical coverage probabilities of simultaneous confidence bands and type I error and power of tests with $\mathcal{H}_0 : \forall s \in \mathcal{S} \triangle \theta(s,t) = 0$, *for comparing dynamic predictions* $\pi^{(1)}(\cdot,t)$ *and* $\pi^{(1)}(\cdot,t)$ *(*$t = 5$, $s \in \{0, 1, \ldots, 10\}$, *1000 runs).*

Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks: WEB APPENDIX

Paul Blanche[1,2], Cécile Proust-Lima[1,2], Loubère Lucie[1,2], Claudine Berr[3], Dartigues Jean-François[1,2], and Hélène Jacqmin-Gadda[1,2]

[1]Univ. Bordeaux Segalen, ISPED, Inserm Research Center U897, F33076, Bordeaux, France
[2]INSERM, ISPED, Centre INSERM U897-Epidémiologie-Biostatistique, F-33000 Bordeaux, France
[2]INSERM, Centre INSERM U1061, Université Montpellier 1, Montpellier, France

A Proof of lemma 1, and formulae of $\mathbf{IF}_\theta(\cdot)$ and $\widehat{\mathbf{IF}}_\theta(\cdot)$

Let $S_{\widetilde{T}}(u) = \mathbb{P}(\widetilde{T} > u)$, $\tau_0 < \sup\left\{u : S_{\widetilde{T}}(u) > 0\right\}$, $\tau_1(s) > \inf\left\{u : \mathbb{P}(T \leq s+u, \eta = 1 | T > s) > 0\right\}$ and $\tau_2(s) < \sup\left\{u : S_{\widetilde{T}}(s+u) > 0\right\}$. Hereafter, we assume that $s \in \mathcal{S} \subset [0, \tau_0]$ and $t \in \mathcal{T}_s \subset [\tau_1(s), \tau_2(s)]$.

The martingale representation of the Kaplan-Meier estimator of the censoring survival function entails that, $\forall s \in \mathcal{S}$,

$$\sup_{t \in \mathcal{T}_s} \left| \sqrt{n}\left(\widehat{G}(t) - G(t)\right) + \frac{G(t)}{\sqrt{n}} \sum_{i=1}^{n} \int_0^t \frac{dM_{C_i}(u)}{S_{\widetilde{T}}(u)} \right| = o_p(1) \tag{1}$$

with the usual martingale $M_{C_i}(t) = \mathbb{1}_{(\widetilde{\eta}_i=0,\widetilde{T}_i\leq t)} - \int_0^t \mathbb{1}_{(\widetilde{T}_i\geq t)} d\Lambda_C(u)$, where $\Lambda_C(\cdot)$ denote the cumulative hazard function of the censoring variable C (see for instance [1]).

Let us recall the following notations: $\forall u > s$ $G(u|s) = G(u)/G(s)$ and $\widehat{G}(u|s) = \widehat{G}(u)/\widehat{G}(s)$. By combining (1) with the first order Taylor expansion of the function $(x, y) \mapsto x/y$ at $\left(\widehat{G}(s+t), \widehat{G}(s)\right)$, it therefore follows that

$$\sup_{s\in\mathcal{S},t\in\mathcal{T}_s} \left| \sqrt{n}\left(\widehat{G}(s+t|s) - G(s+t|s)\right) + \frac{G(s+t|s)}{\sqrt{n}} \sum_{i=1}^{n} \int_s^{s+t} \frac{dM_{C_i}(u)}{S_{\widetilde{T}}(u)} \right| = o_p(1). \quad (2)$$

Besides, before going any further, let us note that the two following proofs follow similar arguments as the proof of Lemma 1 of [4] and of Theorem 1 of [5].

A.1 Proof of lemma 1 for BS

The estimator $\widehat{BS}(s,t)$ (Section 3.2 of the main manuscript) can be written as

$$\widehat{BS}(s,t) = \frac{1}{n}\sum_{i=1}^{n} \left\{ \widehat{h}^{BS}_{i,st,1} + \widehat{h}^{BS}_{i,st,2} \right\}, \quad (3)$$

with,

$$\widehat{h}^{BS}_{i,st,1} = \frac{\mathbb{1}_{(\widetilde{T}_i>s+t)}}{\widehat{G}(s+t|s)} \pi_i(s,t)^2 \quad (4)$$

$$\text{and} \quad \widehat{h}^{BS}_{i,st,2} = \frac{\mathbb{1}_{(s<\widetilde{T}_i\leq s+t)}}{\widehat{G}(\widetilde{T}_i|s)} \left(\mathbb{1}_{(\widetilde{\eta}_i\notin\{0,1\})} \pi_i(s,t)^2 + \mathbb{1}_{(\widetilde{\eta}_i=1)} \left[1 - \pi_i(s,t)\right]^2 \right). \quad (5)$$

Besides, let $h^{BS}_{i,st,1}$ and $h^{BS}_{i,st,2}$ be defined by replacing $\widehat{G}$ by G in formulae (4) and (5). Then, note that

$$\widehat{h}^{BS}_{i,st,1} = h^{BS}_{i,st,1} \left(1 - \frac{\widehat{G}(s+t|s) - G(s+t|s)}{\widehat{G}(s+t|s)} \right) \quad (6)$$

$$\text{and} \quad \widehat{h}^{BS}_{i,st,2} = h^{BS}_{i,st,2} \left(1 - \frac{\widehat{G}(\widetilde{T}_i|s) - G(\widetilde{T}_i|s)}{\widehat{G}(\widetilde{T}_i|s)} \right). \quad (7)$$

Combining (3),(6),(7) and (2), it follows that :

$$\sup_{s\in\mathcal{S},t\in\mathcal{T}_s} \left| \sqrt{n}\left(\widehat{\mathsf{BS}}(s,t) - \mathsf{BS}(s,t)\right) - \frac{\sqrt{n}}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \psi^{BS}_{ij}(s,t) \right| = o_p(1)$$

where

$$\psi_{ij}^{BS}(s,t) = h_{i,st,1}^{BS}\left(1+\int_s^{s+t}\frac{dM_{C_j}(u)}{S_{\widetilde{T}}(u)}\right) + h_{i,st,2}^{BS}\left(1+\int_s^{\widetilde{T}_i}\frac{dM_{C_j}(u)}{S_{\widetilde{T}}(u)}\right) - \mathsf{BS}(s,t). \quad (8)$$

Then, let us also define the Hájek projection [9, Sec. 11.3]:

$$\begin{aligned}
&\mathsf{IF}_{BS}(\widetilde{T}_i,\widetilde{\eta}_i,\pi_i(s,t),s,t)\\
&= \mathbb{E}\Big[\psi_{ij}^{BS}(s,t)+\psi_{ji}^{BS}(s,t)\Big|\big(\widetilde{T}_i,\widetilde{\eta}_i,\pi_i(s,t)\big)\Big]\\
&= h_{i,st,1}^{BS}+h_{i,st,2}^{BS}-\mathsf{BS}(s,t)\\
&\qquad + \mathbb{E}\left[h_{jst,1}^{BS}\right]\int_s^{s+t}\frac{dM_{C_i}(u)}{S_{\widetilde{T}}(u)} + \mathbb{E}\left[h_{jst,2}^{BS}\int_s^{\widetilde{T}_j}\frac{dM_{C_i}(u)}{S_{\widetilde{T}}(u)}\middle|\big(\widetilde{T}_i,\widetilde{\eta}_i,\pi_i(s+t)\big)\right]. \quad (9)
\end{aligned}$$

The following i.i.d decomposition therefore holds from U-statistic theory (see for instance: [8, Sec. 5.1.6], [7, Sec. 3.2.1] or [9, Chap. 12]) :

$$\sup_{s\in\mathcal{S},t\in\mathcal{T}_s}\left|\sqrt{n}\left(\widehat{\mathsf{BS}}(s,t)-\mathsf{BS}(s,t)\right)-\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathsf{IF}_{BS}(\widetilde{T}_i,\widetilde{\eta}_i,\pi_i(s,t),s,t)\right| = o_p(1)$$

A.2 Proof of lemma 1 for AUC

Let us introduce the following notations,

$$h_{ij,st,1}^{AUC} = \frac{\mathbb{1}_{(s<\widetilde{T}_i\le s+t,\widetilde{\eta}_i=1)}\mathbb{1}_{(\widetilde{T}_j>s+t)}}{G(\widetilde{T}_i|s)G(s+t|s)}\mathbb{1}_{(\pi_i(s,t)>\pi_j(s,t))}, \quad (10)$$

$$h_{ij,st,2}^{AUC} = \frac{\mathbb{1}_{(s<\widetilde{T}_i\le s+t,\widetilde{\eta}_i=1)}\mathbb{1}_{(s<\widetilde{T}_j\le s+t,\widetilde{\eta}_j\notin\{0,1\})}}{G(\widetilde{T}_i|s)G(\widetilde{T}_j|s)}\mathbb{1}_{(\pi_i(s,t)>\pi_j(s,t))}, \quad (11)$$

$$\text{and}\quad f_{i,st} = \frac{\mathbb{1}_{(s<\widetilde{T}_i\le s+t,\widetilde{\eta}_i=1)}}{G(\widetilde{T}_i|s)}. \quad (12)$$

In addition, let $\widehat{h}_{ij,st,1}^{AUC}$, $\widehat{h}_{ij,st,2}^{AUC}$ and $\widehat{f}_{i,st}$ be defined be replacing G by $\widehat{G}$ in formulae (10), (11) and (12). Let also define $h_{st}^{AUC} = \mathbb{E}\left[h_{ij,st,1}^{AUC}+h_{ij,st,2}^{AUC}\right]$ and $F_1(s+t|s) = \mathbb{P}(s<T\le s+t,\eta=1|T>s)$, and the corresponding consistent estimators:

$$\widehat{h}_{st}^{AUC} = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \widehat{h}_{ij,st,1}^{AUC}+\widehat{h}_{ij,st,2}^{AUC}, \quad (13)$$

$$\text{and}\quad \widehat{F}_1(s+t|s) = \frac{1}{n}\sum_{i=1}^n \widehat{f}_{i,st}. \quad (14)$$

Then, note that:

$$\widehat{h}_{ij,st,1}^{AUC} = h_{ij,st,1}^{AUC}\left(1-\frac{\widehat{G}(s+t|s)-G(s+t|s)}{\widehat{G}(s+t|s)}\right)\left(1-\frac{\widehat{G}(\widetilde{T}_i|s)-G(\widetilde{T}_i|s)}{\widehat{G}(\widetilde{T}_i|s)}\right) \quad (15)$$

$$\widehat{h}_{ij,st,2}^{AUC} = h_{ij,st,2}^{AUC}\left(1-\frac{\widehat{G}(\widetilde{T}_i|s)-G(\widetilde{T}_i|s)}{\widehat{G}(\widetilde{T}_i|s)}\right)\left(1-\frac{\widehat{G}(\widetilde{T}_j|s)-G(\widetilde{T}_j|s)}{\widehat{G}(\widetilde{T}_j|s)}\right) \quad (16)$$

$$\text{and} \quad \widehat{h}_{i,st} = f_{i,st}\left(1-\frac{\widehat{G}(\widetilde{T}_i|s)-G(\widetilde{T}_i|s)}{\widehat{G}(\widetilde{T}_i|s)}\right). \quad (17)$$

Combining (13), (15),(16) and (2), it follows that :

$$\sup_{s\in\mathcal{S},t\in\mathcal{T}_s}\left|\sqrt{n}\left(\widehat{h}_{st}^{AUC}-h_{st}^{AUC}\right)-\frac{\sqrt{n}}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\left\{h_{ij,st,1}^{AUC}\left(1+\int_0^{\widetilde{T}_i}\frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}+\int_0^{t}\frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right)\right.\right.$$
$$\left.\left.+h_{ij,st,2}^{AUC}\left(1+\int_0^{\widetilde{T}_i}\frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}+\int_0^{\widetilde{T}_j}\frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right)-h_{st}^{AUC}\right\}\right| = o_p(1). \quad (18)$$

Combining (14), (17) and (2), it also follows that :

$$\sup_{s\in\mathcal{S},t\in\mathcal{T}_s}\left|\sqrt{n}\left(\widehat{F}_1(s+t|s)-F_1(s+t|s)\right)-\frac{\sqrt{n}}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\left\{f_{i,st}\left(1+\int_0^{\widetilde{T}_i}\frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right)\right.\right.$$
$$\left.\left.-F_1(s+t|s)\right\}\right| = o_p(1). \quad (19)$$

As reminded in [4, Sec 3.1], based on results from [6] and [2] it can be seen that:

$$\sum_{j=1}^{n}\left\{\frac{\mathbb{1}_{(\widetilde{T}_j>s+t)}}{\widehat{G}(s+t|s)}+\frac{\mathbb{1}_{(s<\widetilde{T}_j\le s+t,\widetilde{\eta}_j\notin\{0,1\})}}{\widehat{G}(\widetilde{T}_j|s)}\right\}=\sum_{i=1}^{n}\left\{1-\frac{\mathbb{1}_{(s<\widetilde{T}_i\le s+t,\widetilde{\eta}_i=1)}}{\widehat{G}(\widetilde{T}_i|s)}\right\}.$$

As a consequence, the denominator of $\widehat{AUC}(s,t)$ reduces to $\widehat{F}_1(s+t|s)\left\{1-\widehat{F}_1(s+t|s)\right\}$ and this leads to $\widehat{AUC}(s,t)=\widehat{h}_{st}^{AUC}\Big/\left[\widehat{F}_1(s+t|s)\left\{1-\widehat{F}_1(s+t|s)\right\}\right]$. Thus, by combining (18), (19) and a first order Taylor expansion of the function $(x,y)\mapsto x\big/\{y(1-y)\}$ at $(x,y)=(\widehat{h}_{st},\widehat{F}_1(s+t|s))$, we further derive

$$\sup_{s\in\mathcal{S},t\in\mathcal{T}_s}\left|\sqrt{n}\left(\widehat{AUC}(s,t)-AUC(s,t)\right)-\frac{\sqrt{n}}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\Psi_{ijk}^{AUC}(s,t)\right| = o_p(1)$$

where

$$\Psi_{ijk}^{AUC}(s,t) = \Bigg\{ h_{ij,st,1}^{AUC}\left(1 + \int_0^{\widetilde{T}_i} \frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)} + \int_0^t \frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right)$$
$$+ h_{ij,st,2}^{AUC}\left(1 + \int_0^{\widetilde{T}_i} \frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)} + \int_0^{\widetilde{T}_j} \frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right)$$
$$- h_{st}^{AUC} - \frac{h_{st}^{AUC}(1-2F_1(s+t|s))}{F_1(s+t|s)\big(1-F_1(s+t|s)\big)}\left[f_{i,st}\left(1 + \int_0^{\widetilde{T}_i} \frac{dM_{C_k}(u)}{S_{\widetilde{T}}(u)}\right) - F_1(s+t|s)\right]$$
$$\Bigg\} \Big/ \Big[F_1(s+t|s)\big\{1 - F_1(s+t|s)\big\}\Big]. \tag{20}$$

Let us also define the Hájek projection [9, Sec. 11.3]:

$$\begin{aligned}&\mathsf{IF}_{AUC}(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i(s,t), s, t)\\ &= \mathbb{E}\Big[\psi_{ijk}^{AUC}(s,t) + \psi_{jik}^{AUC}(s,t) + \psi_{jki}^{AUC}(s,t)\Big|\big(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i(s,t)\big)\Big].\end{aligned} \tag{21}$$

The following i.i.d decomposition therefore holds from U-statistic theory (see for instance [8, Sec. 5.1.6], [7, Sec. 3.2.1] or [9, Chap. 12]):

$$\sup_{s\in\mathcal{S}, t\in\mathcal{T}_s}\left|\sqrt{n}\left(\widehat{\mathsf{AUC}}(s,t) - \mathsf{AUC}(s,t)\right) - \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathsf{IF}_{AUC}(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i(s,t), s, t)\right| = o_p(1)$$

A.3 Estimators of the influence functions for BS and AUC

Let $\widehat{S}_{\widetilde{T}}(u)$ be the empirical survival function of $S_{\widetilde{T}}(u)$ and let $\widehat{M}_C$ be defined by plugging the usual Nelson-Aalen $\widehat{\Lambda}_C(\cdot)$ estimator of the cumulative incidence function of the censoring $\Lambda_C(\cdot)$ into M_C, i.e $\widehat{M}_C(t) = \mathbb{1}_{(\widetilde{\eta}=0,\widetilde{T}\leq t)} - \int_0^t \mathbb{1}_{(\widetilde{T}_i\geq t)} d\widehat{\Lambda}_C(u)$.

We further define the plug-in estimator of $\Psi_{ij}^{BS}(s,t)$, denoted by $\widehat{\Psi}_{ij}^{BS}(s,t)$, by respectively replacing $h_{i,st,1}^{BS}$, $h_{i,st,2}^{BS}$, $M_C(\cdot)$, $S_{\widetilde{T}}(\cdot)$ and $\mathsf{BS}(s,t)$ by $\widehat{h}_{i,st,1}^{BS}$, $\widehat{h}_{i,st,2}^{BS}$, $\widehat{M}_C(\cdot)$, $\widehat{S}_{\widetilde{T}}(\cdot)$ and $\widehat{\mathsf{BS}}(s,t)$ in equation (8).

Similarly, we also define the estimator of $\Psi_{ijk}^{AUC}(s,t)$, denoted by $\widehat{\Psi}_{ijk}^{AUC}(s,t)$, by respectively replacing $h_{ij,st,1}^{AUC}$, $h_{ij,st,2}^{AUC}$, h_{st}^{AUC}, $f_{i,st}$, $M_C(\cdot)$, $S_{\widetilde{T}}(\cdot)$ and $F_1(s+t|s)$ by $\widehat{h}_{ij,st,1}^{AUC}$, $\widehat{h}_{ij,st,2}^{AUC}$, $\widehat{h}_{st}^{AUC}$, $\widehat{f}_{i,st}$, $\widehat{M}_C(\cdot)$, $\widehat{S}_{\widetilde{T}}(\cdot)$ and $\widehat{F}_1(s+t|s)$ in equations (20).

Consequently, by plugging in $\widehat{\psi}_{ij}^{BS}(s,t)$ and $\widehat{\psi}_{ij}^{AUC}(s,t)$ and estimating conditional expectation by empirical means, we respectively define the estimator of the influence functions $\mathsf{IF}_{BS}(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i(s,t), s, t)$ and $\mathsf{IF}_{AUC}(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i(s,t), s, t)$ by:

$$\widehat{\mathsf{IF}}_{BS}(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i(s,t), s, t) = \frac{1}{n}\sum_{j=1}^{n}\left\{\widehat{\psi}_{ij}^{BS}(s,t) + \widehat{\psi}_{ji}^{BS}(s,t)\right\}$$

and

$$\begin{aligned}&\widehat{\mathsf{IF}}_{AUC}(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i(s,t), s, t)\\ &\qquad = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\left\{\widehat{\psi}_{ijk}^{AUC}(s,t) + \widehat{\psi}_{jik}^{AUC}(s,t) + \widehat{\psi}_{jki}^{AUC}(s,t)\right\}.\end{aligned}$$

B Computing dynamic predictions from a joint latent class model

For a joint latent class model we have the following decomposition,

$$\begin{aligned}\pi_i(s,t) &= \mathbb{P}(s < T_i \leq s+t, \eta_i = 1 | T > s, \mathcal{Y}_i(s), \mathbf{X}_i)\\ &= \frac{\sum_g \mathbb{P}(\gamma_i = g|\mathbf{X}_i)\mathcal{L}_g(\mathcal{Y}_i(s)|\mathbf{X}_i)\Big\{F_{1g}(s+t|\mathbf{X}_i) - F_{1g}(s|\mathbf{X}_i)\Big\}}{\sum_g \mathbb{P}(\gamma_i = g|\mathbf{X}_i)\mathcal{L}_g(\mathcal{Y}_i(s)|\mathbf{X}_i)S_g(s|\mathbf{X}_i)}\end{aligned}$$

with:

- $\mathcal{L}_g(\mathcal{Y}_i(s)|\mathbf{X}_i)$, the density of the multivariate gaussian density of the vector $\mathcal{Y}_i(s) = \{Y_{ij} : 0 \leq t_{ij} \leq s, j = 1, \ldots, n_i\}$ conditionally on $\mathbf{X}_i$ and on the latent class $\gamma_i = g$.
- $S_g(s|\mathbf{X}_i)$: the overall-event-free survival function at time s conditionally on $\mathbf{X}_i$ and on the latent class $\gamma_i = g$, that is

$$S_g(s|\mathbf{X}_i) = \exp\left(-\int_0^s \Big\{\lambda_{i,1,g}(u|\mathbf{X}_i) + \lambda_{i,2,g}(u|\mathbf{X}_i)\Big\}du\right). \tag{22}$$

- $F_{1g}(s|\mathbf{X}_i)$: the cumulative incidence of the main event at time s conditionally on $\mathbf{X}_i$ and on the latent class $\gamma_i = g$, that is

$$F_{1g}(s|\mathbf{X}_i) = \int_0^s \lambda_{i,1,g}(u|\mathbf{X}_i)S_g(u|\mathbf{X}_i)du. \tag{23}$$

And denoting by $\lambda_{i,k,g}(u|\mathbf{X}_i)$, $k = 1, 2$, the class specific cause specific hazard of cause k at time u, conditionally on $\mathbf{X}_i$ and on the latent class $\gamma_i = g$, that is

$$\lambda_{i,k,g}(u|\mathbf{X}_i) = \lim_{dt \downarrow 0} \frac{1}{dt} \mathbb{P}(u \leq T_i < u + dt, \eta = k | T_i \geq u, \mathbf{X}_i, \gamma_i = g),$$

for which we used the notation $\lambda_{i,k}(u|\gamma_i = g)$ in the main manuscript.

C Parameter values for the simulation study

C.1 Generation of survival data (common for all scenarii)

Constant class and cause specific hazard was chosen in the simulation, i.e $\forall u$, $\lambda_{i,k,g}(u|\mathbf{X}_i) \equiv \lambda_{k,g}$. Therefore formulae (22) and (23) reduce to

$$S_g(s|\mathbf{X}_i) \equiv S_g(s) = \exp\Big(- (\lambda_{1,g} + \lambda_{2,g})s \Big)$$

and

$$F_{1g}(s|\mathbf{X}_i) \equiv F_{1g}(s) = \frac{\lambda_{1,g}}{\lambda_{1,g} + \lambda_{2,g}} \left(1 - e^{-(\lambda_{1,g}+\lambda_{2,g})s}\right).$$

Web Figure I displays the functions $S_g(\cdot)$ and $F_{kg}(\cdot)$, $k = 1, 2$, $g = 1, 2$. The values of constant class-cause-specific hazard $\lambda_{k,g}$ of event $k = 1, 2$ given the class $\gamma = g$ was chosen equal to:

g=1: $\lambda_{1,1} = 1/6$ for event 1, $\lambda_{1,2} = 1/45$ for event 2,

g=2: $\lambda_{2,1} = 1/20$ for event 1, $\lambda_{1,2} = 1/40$ for event 2,

such that the proportions of observed main event, i.e $\{i : s < \widetilde{T}_i \leq s + t, \widetilde{\eta}_i = 1\}$, competing event, i.e $\{i : s < \widetilde{T}_i \leq s + t, \widetilde{\eta}_i = 2\}$ and censored observations, i.e $\{i : s < \widetilde{T}_i \leq s + t, \widetilde{\eta}_i = 0\}$ within the time interval $(s, s + t]$ and observed event free at time $s + t$, i.e $\{i : \widetilde{T}_i > s + t\}$ were respectively around 20%, 10%, 20% and 50% of subjects at risk at time s, for all landmark times s.

We inverted the survival function $u \mapsto S_g(u)$ to generate T_i given class γ_i and generate η_i given class γ_i from a Bernoulli distribution with parameter $\lambda_{g,1}/(\lambda_{g,1} + \lambda_{g,1})$, as described in [3]. Independent censoring C was generated with an exponential distribution such that $\mathbb{E}(C) = 1/\lambda_C = 20$.

C.2 Generation of longitudinal data according to each scenario

We generated the data from the 8 following scenarii that are displayed in Figure 1 of the main manuscript. Parameter values for each scenario were:

H0 : Under the null hypothesis.

H0.1 : For $l = 1, 2$, we chose:

g=1: Intercept $\beta_{01}^{(l)} = 25$, slope $\beta_{11}^{(l)} = -0.2$.

g=2: Intercept $\beta_{02}^{(l)} = 26$, slope $\beta_{12}^{(l)} = -0.1$.

For both classes $g = 1, 2$, the values of variances of random intercept and slope were equal to $\Sigma_b = \text{diag}(0.25^2, 0.05^2)$, and variance of the noise was $\sigma_\varepsilon^2 = 1.2$.

H0.2 : Compared to H0.1, for class $g = 1$ we have now $\beta_{11}^{(l)} = -0.3$, $l = 1, 2$.

H1.1 : Alternative hypotheses with increasing difference $\beta_{11}^{(1)} - \beta_{12}^{(1)}$.

H1.1.1 : For $Y^{(1)}(\cdot)$, we have $\beta_{11}^{(1)} - \beta_{12}^{(1)} = -0.3 - (-0.1) = -0.2$, whereas for $Y^{(2)}(\cdot)$, as in H0.1 we still have $\beta_{11}^{(2)} - \beta_{12}^{(2)} = -0.2 - (-0.1) = -0.1$.

H1.1.2 : As in H1.1.1, except that $\beta_{11}^{(1)} - \beta_{12}^{(1)} = -0.4 - (-0.1) = -0.3$.

H1.1.3 : As in H1.1.1 and H1.1.2, except that $\beta_{11}^{(1)} - \beta_{12}^{(1)} = -0.5 - (-0.1) = -0.4$.

H1.2 : Alternative hypotheses with reduced variances for marker trajectory $Y^{(1)}(\cdot)$.

H1.2.1 : Compared to H0.2, standard deviations of both noise, σ_ε, and random effect, $\sigma_{b_0}, \sigma_{b_0}$, are divided by 1.5 for marker trajectory $Y^{(1)}(\cdot)$.

H1.2.2 : Compared to H0.2, standard deviations are divided by 2 for marker trajectory $Y^{(1)}(\cdot)$.

H1.3 : Additional alternative hypothesis scenario where data were generated such that the two class-specific trajectories of marker $Y^{(1)}(\cdot)$ crossed (at $t = 1.25$).
Marker $Y^{(2)}(\cdot)$ has the same trajectory than in H0.1, except that $\beta_{11}^{(2)} = -0.25$. Marker $Y^{(1)}(\cdot)$ has the following parameters

g=1: Intercept $\beta_{01}^{(1)} = 26.5$, slope $\beta_{11}^{(1)} = -0.5$.

g=2: Intercept $\beta_{02}^{(1)} = 26$, slope $\beta_{02}^{(1)} = -0.1$.

D Additional simulation results

As Figure 2 of the main manuscript, Web Figure II of this document displays the mean curves of $BS(s,t)$ versus s for $\pi^{(2)}(s,t)$, and the coverage probabilities of the pointwise confidence intervals. Web Figure III and Web Figure IV respectively display similar plots for the difference in AUC and in BS. In particular, they show that the asymptotic inference procedures behave well under our data generated scenarii.

[Web Figure II about here]

[Web Figure III about here]

[Web Figure IV about here]

References

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

[2] L. Antolini, E.M. Biganzoli, and P. Boracchi. Crude Cumulative Incidence in the form of a Horvitz-Thompson like and Kaplan-Meier like Estimator. *COBRA Preprint Series*, 2006.

[3] J. Beyersmann, A. Latouche, A. Buchholz, and M. Schumacher. Simulating competing risks data in survival analysis. *Statistics in medicine*, 28(6):956–971, 2009.

[4] P. Blanche, J-F Dartigues, and H. Jacqmin-Gadda. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine, in press*, 2013.

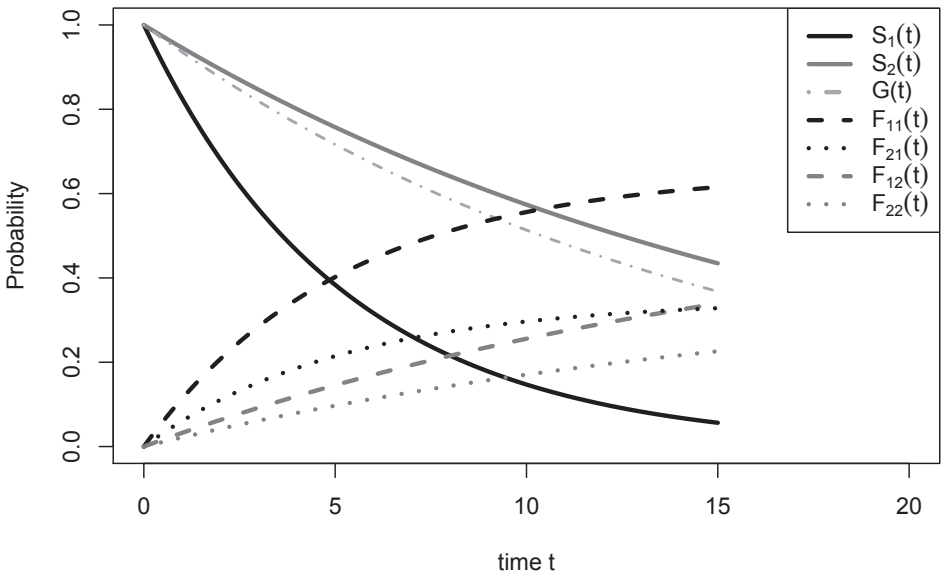
[5] H. Hung and C.T. Chiang. Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics*, 38(1):8–26, 2010.

[6] N.P. Jewell, X. Lei, A.C. Ghani, C.A. Donnelly, G.M. Leung, L.M. Ho, B.J. Cowling, and A.J. Hedley. Non-parametric estimation of the case fatality ratio with competing risks data: an application to Severe Acute Respiratory Syndrome (SARS). *Statistics in medicine*, 26(9):1982–1998, 2007.

[7] Jeanne Kowalski and Xin M Tu. *Modern applied U-statistics*, volume 714. Wiley. com, 2008.

[8] R.J. Serfling. Approximation theorems of mathematical statistics. 1980.

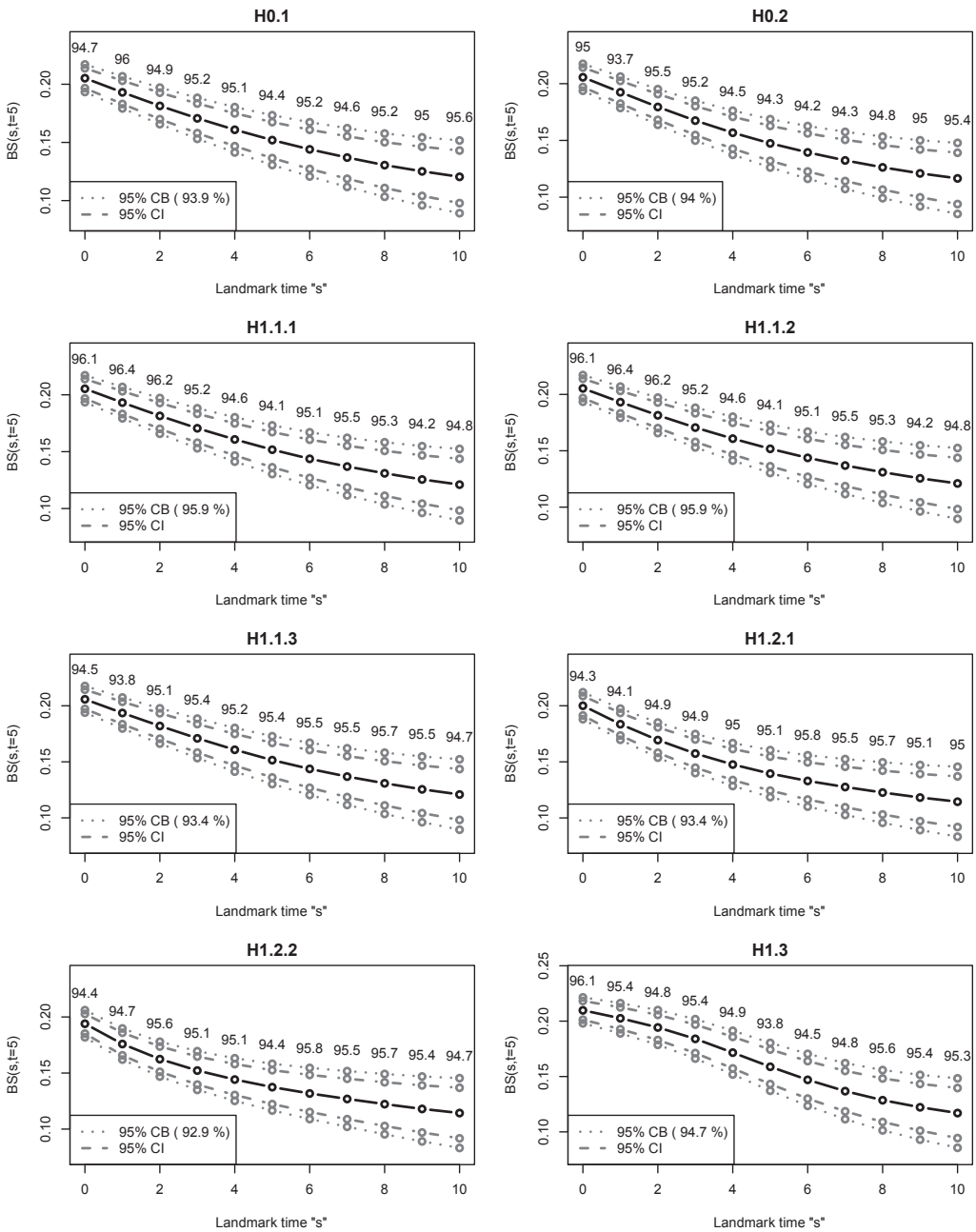
[9] A.W. Van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998.



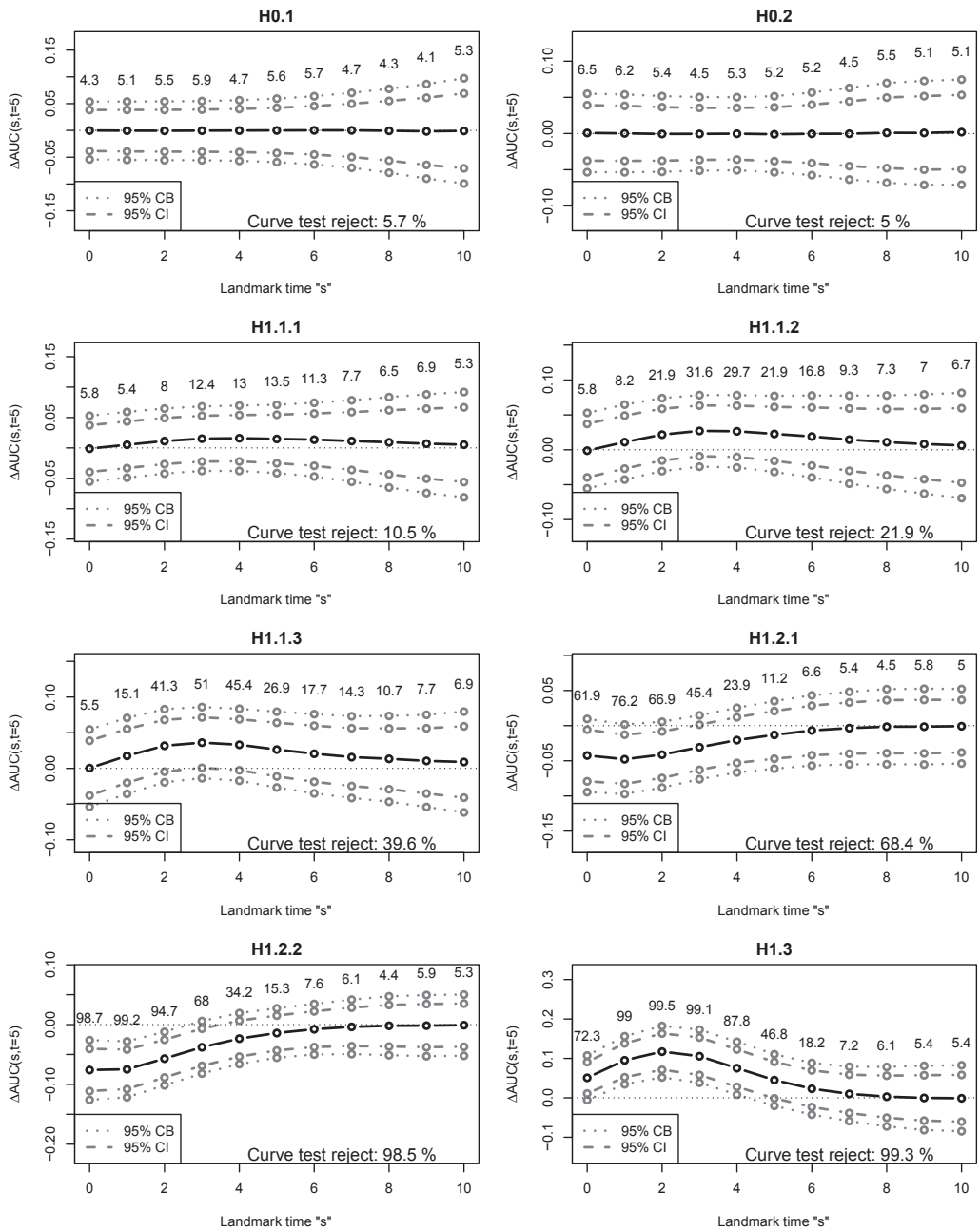
Web Figure I: Class specific cumulative incidence functions $F_{kg}(\cdot)$ generated for event-type $k = 1, 2$ and class $g = 1, 2$; class specific event-free survival functions $S_g(\cdot)$, $g = 1, 2$; censoring survival distribution $G(\cdot)$.

		PAQUID (n=2970)		3C (n=3880)	
Age		74.8	(6.5)	73.7	(5.3)
MMSE		26.1	(3.1)	27.3	(2.0)
IST		26.9	(6.3)	29.5	(5.6)
Gender					
	male	1258	(42.4%)	1557	(40.1%)
	female	1712	(57.6%)	2323	(59.9%)
Education level					
	low	965	(32.5%)	335	(8.6%)
	high	2005	(67.5%)	3545	(91.4%)

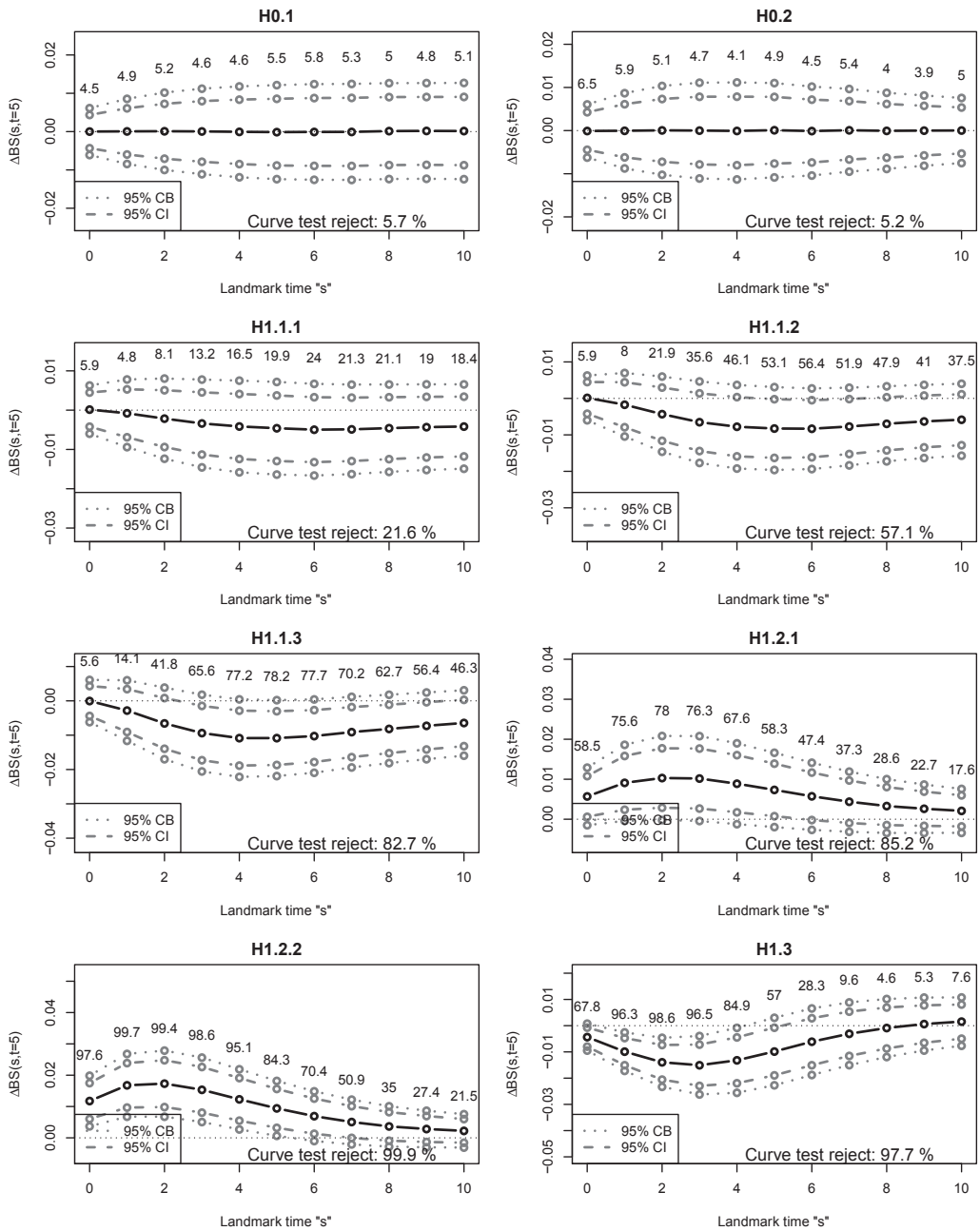
Web Table I: Descriptive statistics of the covariates at baseline for Paquid and 3C cohort data. Means and standard deviations, frequencies and proportions.



Web Figure II: Average estimates of the curve $BS(s, t = 5)$ versus s for $\pi^{(2)}(s, t)$, $n = 2000$. Average 95% pointwise confidence intervals (CI) and 95% simultaneous confidence band (CB) are plotted from the average standard errors and quantiles over the 1000 runs. The empirical coverage probabilities of the 95% CB are given in the legends and of the 95% CI for each time s are displayed in the plot.



Web Figure III: Average of estimates of the curve of the difference in $AUC(s, t = 5)$ for $\pi^{(1)}(s, t)$ and $\pi^{(2)}(s, t)$ versus s, $n = 2000$. Average 95% pointwise confidence intervals (CI) and 95% simultaneous confidence band (CB) are plotted from the average standard errors and quantiles over the 1000 runs. The empirical type one error or power of the simultaneous test of comparison of the two AUC are given ("curve test reject"). The empirical type one errors or powers of the pointwise tests for each time s are displayed in the plot.



Web Figure IV: Average of estimates of the curve of the difference in $BS(s, t = 5)$ for $\pi^{(1)}(s, t)$ and $\pi^{(2)}(s, t)$ versus s, $n = 2000$. Average 95% pointwise confidence intervals (CI) and 95% simultaneous confidence band (CB) are plotted from the average standard errors and quantiles over the 1000 runs. The empirical type one error or power of the simultaneous test of comparison of the two BS are given ("curve test reject"). The empirical type one errors or powers of the pointwise tests for each time s are displayed in the plot.

V.2 Compléments

V.2.1 Une amélioration du calcul des régions de confiance

Similairement à ce que l'on a vu en Section IV.2.1 pour le cas non dynamique, on pourrait, ici aussi, améliorer le calcul des intervalles de confiance et des bandes de confiance. En particulier, pour le cas des bandes de confiance des courbes $\left\{\left(AUC(s,t), s\right), s \in \mathcal{S}\right\}$ à t fixé, l'amélioration sera probablement encore meilleure que pour le cas de celles du type $\left\{\left(AUC(0,t), t\right), t_{min} \leqslant t \leqslant t_{max}\right\}$ proposées dans l'article de la Section IV.1. En effet, lorsque s croît, l'effectif n_s des sujets à risque d'événement en s décroît. Bien qu'en théorie cela ne pose pas de problème, puisque, comme on l'a mentionné dans l'article précédent, $\sqrt{n_s} = \sqrt{n} \times \sqrt{S_{\tilde{T}}(s)} + o_p(1)$, en pratique cela pourrait poser quelques soucis si les effectifs n_s devenaient trop petits. En effet, il en résulterait alors que la variance des $\widehat{AUC}(s,t)$ serait grande et que les approximations normales de $\widehat{AUC}(s,t)$ seraient mauvaises. Pour le calcul des bandes de confiance des courbes de $AUC(s,t)$ versus s, les méthodes d'amélioration du calcul des régions de confiance présentées en Section IV.2.1, et des tests associés, sont donc potentiellement très intéressantes.

Dans notre application aux données de la cohorte Trois-Cités, les échantillons sont cependant de tailles importantes (Figure V.1), et les approximations normales nous paraissent donc raisonnables.

V.2.2 Critère du type $R^2(s,t)$

Le Brier score $BS(s,t)$ dépend de l'incidence cumulée $F_1(s+t|s) = \mathbb{P}(s < T \leqslant s+t, \eta = 1|T > S)$. Comme on l'a déjà évoqué en Section II.3.2.4, l'impact des qualités d'une prédiction dynamique $\pi(s,t)$ sur la valeur du Brier score peut alors être difficile à différencier de celui de $F_1(s,t)$. Cela ne facilite pas l'interprétation de $BS(s,t)$ pour un unique temps landmark s, et encore moins l'interprétation de la courbe de $BS(s,t)$ versus s. En effet, le sens croissant (ou décroissant) d'une telle courbe ne peut pas être interprété comme le fait que les prédictions prédisent de mieux en mieux (ou de moins en moins bien) quand s augmente.

Une solution pour contourner ces difficultés consiste à définir un critère du type « R^2 »,

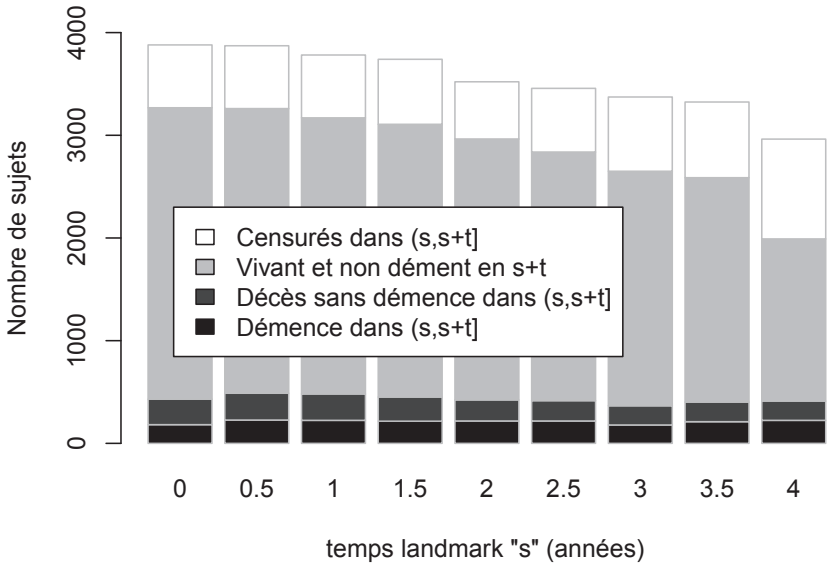


FIGURE V.1 – Nombres de sujets à risque $n_s = \sum_i \mathbb{1}_{(\tilde{T}_i > s)}$ et statistiques descriptives à chaque temps landmark $s \in \{0, 0.5, \ldots, 4\}$ ans pour l'horizon de prévision $t = 5$ ans (échantillon de Trois-Cités).

comme Parast *et al.* (2012), dont on a déjà discuté en Section II.3.2.4. On le définit par

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

avec $BS_0(s,t)$ le Brier score du modèle « nul », n'incluant aucune covariable, c'est-à-dire le prédicteur qui prédit un risque égal à $F_1(s+t|s)$ pour tous les sujets. En pratique, on l'estimera par

$$\widehat{R}^2(s,t) = 1 - \frac{\widehat{BS}(s,t)}{\widehat{BS}_0(s,t)},$$

avec $\widehat{BS}_0(s,t) = \widehat{F}_1(s+t|s)\Big(1 - \widehat{F}_1(s+t|s)\Big)$, où $\widehat{F}_1(s+t|s)$ est l'estimateur non paramétrique de l'incidence cumulée de Aalen et Johansen (1978). Définissant un critère de type « variance expliquée », comme on l'a vu en Section II.3.2.4, son interprétation est plus aisée, comme celle d'une quantité entre 0 et 1, pour laquelle une valeur est d'autant meilleure qu'elle est élevée. L'intérêt particulier d'un tel critère pour le cas dynamique est que la courbe de $\widehat{R}^2(s,t)$ versus s aura la même « échelle » pour tout temps landmark s et sera ainsi plus facile à interpréter.

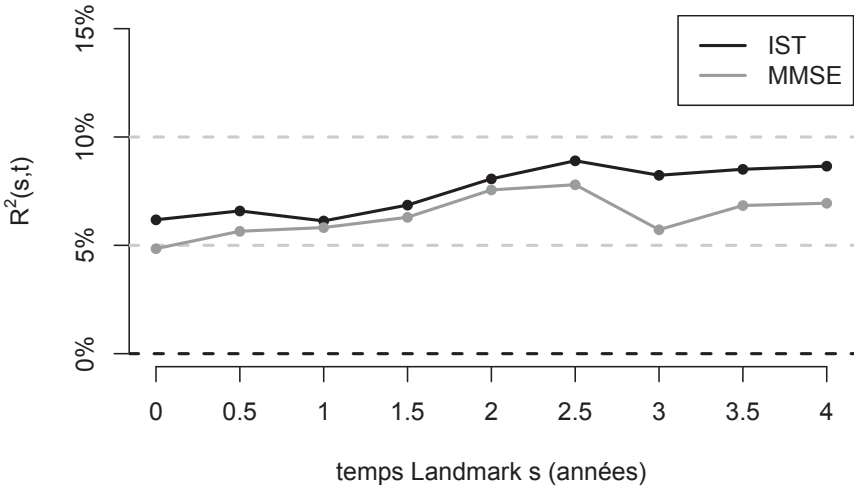


FIGURE V.2 – Critère $R^2(s,t)$ à chaque temps landmark $s \in \{0, 0.5, \ldots, 4\}$ ans pour l'horizon de prévision $t = 5$ ans (échantillon de Trois-Cités).

Pour l'application de notre article, des courbes de $\widehat{R}^2(s,t)$ versus s sont présentées à la Figure V.2 et montrent des $\widehat{R}^2(s,t)$ allant de 6 à 9% quand s varie. De manière intéressante, on constate que les courbes semblent croissantes, suggérant que les prédictions ont un Brier score de plus en plus élevé, non pas en « valeur brute », comme présentée à la Figure 4 de l'article précédant, mais relativement aux prédictions du modèle nul. Ainsi, d'un point de vue de l'interprétation, cela signifie que les prédictions sont de plus en plus précises comparées à des prédictions n'utilisant pas d'information spécifique à chaque sujet car l'information $\mathcal{H}(s)$ utilisée par les prédictions $\pi(s,t)$ croît avec s.

Cela conforte l'idée que c'est le mécanisme de sélection de l'échantillon à risque qui rend les courbes $AUC(s,t)$ versus s décroissantes à la Figure 4 de l'article précédent. En effet, l'échantillon est de plus en plus homogène quand s augmente, et donc difficile à discriminer.

Enfin, en complément du travail de l'article précédent, il aurait aussi été intéressant de proposer (et d'évaluer) des régions de confiance pour des courbes de $\widehat{R}^2(s,t)$ versus s. Elles pourraient aisément être dérivées des propriétés asymptotiques de l'estimateur IPCW du Brier score, et de la « méthode delta » ou de quelques raffinements comme ceux discutés en Sec-

tion IV.2.1 et V.2.1 pour l'AUC. Rappelons que pour comparer des outils de prédiction dynamique, les comparaisons des courbes de $R^2(s,t)$ versus s ou de $BS^2(s,t)$ versus s sont cependant équivalentes.

VI. Conclusion et perspectives

VI.1 Résumé des travaux

Dans cette thèse on s'est principalement intéressé à l'estimation de courbes ROC dépendant du temps et des aires sous ces courbes en présence de données censurées, ainsi qu'au Brier score. Pour définir des estimateurs consistants, et ce même en présence d'une censure possiblement dépendante de l'outil pronostique étudié, on a proposé d'utiliser une méthode « Inverse Probability of Censoring Weighting ». On a ensuite adapté les définitions et les estimateurs à la situation de risques concurrents, particulièrement intéressante pour tenir compte du risque concurrent de décès sans démence lors de l'évaluation d'un marqueur pronostique de démence. Pour compléter les procédures d'estimation, on a aussi proposé la construction de régions de confiance et de tests. Enfin, on s'est intéressé à l'évaluation d'un marqueur pronostique répété, permettant de définir des outils prédictifs dynamiques. Ce dernier cas nous parait particulièrement intéressant en pratique, puisque les travaux d'Amieva *et al.* (2005, 2008), qui ont en partie motivé cette thèse, suggèrent que les mesures répétées de tests cognitifs ont beaucoup de potentiels pour prédire une démence. Toutes les procédures d'inférence ont été évaluées par simulation et ont montré des résultats satisfaisants. Nous avons aussi proposé quelques extensions et alternatives, pour traiter le cas spécifique des données censurées par intervalles de Paquid, par exemple. Par ailleurs, les diverses applications aux données des cohortes Paquid et Trois-Cités ont permis d'évaluer le potentiel prédictif de plusieurs tests cognitifs. Enfin, pour faciliter la diffusion des méthodes proposées, le package R timeROC a été déposé sur le site du CRAN (« Comprehensive R Archive Network »).

Dans l'ensemble de ces travaux, les méthodes développées ont été appliquées à l'évaluation de marqueurs pronostiques observés ou de modèles pronostiques construits sur un échantillon différent de l'échantillon de validation. Nous ne nous sommes pas intéressés à la validation interne et la plupart de nos perspectives de travail portent sur ce sujet.

VI.2 Perspectives

VI.2.1 Une limite de la validation externe

La construction de prédictions dynamiques dérivées de modèles conjoints est intéressante car ces modèles permettent l'exploitation d'une grande quantité d'information. Au Chapitre V on a employé des modèles conjoints basés sur un unique test cognitif répété. Pour améliorer les performances des prédictions dynamiques, il serait intéressant d'utiliser un modèle conjoint pour marqueurs longitudinaux multivariés afin de prendre en compte l'évolution simultanée de plusieurs tests cognitifs. On pourrait alors faire des prédictions basées sur les mesures répétées de plusieurs tests cognitifs complémentaires, évaluant différentes dimensions de la cognition.

Les données de la cohorte Paquid pourraient être utilisées pour estimer de tels modèles. Cependant, de nombreux tests cognitifs passés par les sujets de Paquid n'ont pas été passés par les sujets de Trois-Cités. Les prédictions dynamiques issues de modèles conjoints pour marqueurs longitudinaux multivariés ne pourraient donc pas être évaluées sur les données de validation « externe » de la cohorte des Trois-Cités.

Il serait donc intéressant d'étudier des procédures de construction de régions de confiance et de tests adaptées au cas d'une validation « interne », c'est-à-dire lorsque les mêmes données sont utilisées à la fois pour estimer les modèles prédictifs et pour évaluer les prédictions. Pour cela, on pourrait utiliser des méthodes de « cross-validation », que l'on a déjà discuté succinctement en Section II.3.5.

VI.2.2 Cross-validation « usuelle » et procédures d'inférence

Pour corriger des biais d'optimisme dans l'estimation de capacités prédictives, on pourrait envisager une cross-validation « usuelle », telle qu'une « k-fold » cross-validation déjà évoquée en Section II.3.5. Cette approche serait probablement très chronophage avec des modèles conjoints, car elle nécessiterait l'estimation de k modèles conjoints. Cependant, elle pourrait être envisagée en combinant les récents progrès en terme de calcul parallèle et l'accessibilité croissante à des calculateurs puissants et multi-tâches. Le calcul de régions de confiance et de

tests à partir d'estimations par cross-validation n'est cependant pas évident et des réflexions complémentaires dans cette direction sont nécessaires.

Cependant, une approche pourrait consister à suivre les raisonnements de Uno *et al.* (2007) et Tian *et al.* (2007), récemment repris par Parast et Cai (2013). Brièvement, pour calculer des intervalles de confiance, ils montrent qu'on peut les centrer sur une estimation cross-validée et calculer les bornes à partir de la variance estimée en considérant les données comme externes. Cependant, il existe quelques différences importantes entre les modèles pronostiques qu'ils considèrent et les modèles conjoints. Notamment concernant le ratio entre le nombre de paramètres à estimer et le nombre de sujets, qui peut être nettement plus élevé dans le cas des modèles conjoints. Des réflexions additionnelles sur les possibilités et les limites de cette approche, et éventuellement une étude de simulations, seraient donc les bienvenues avant de l'utiliser avec des modèles conjoints.

Pour construire des tests de comparaison d'outils de prédiction à partir d'une approche par cross-validation, la méthode de van de Wiel *et al.* (2009) parait également intéressante. Brièvement, elle consiste à calculer une *p-value* à chaque découpage des données en échantillons d'apprentissage et de validation lors de chaque itération de la procédure de cross-validation. Pour cela, on effectue un test de comparaison sur les données de validation en utilisant les modèles estimés sur les données d'apprentissage. On compare ensuite la valeur médiane de ces *p-value* (ou la moyenne) au seuil de significativité du test souhaité. Bien qu'elle soit d'une simplicité étonnante, cette méthode est justifiée par quelques résultats mathématiques rigoureux par van de Wiel *et al.* (2009). D'un point de vue pratique, le principal avantage de cette méthode réside dans son universalité et son implémentation triviale. Elle peut en effet être utilisée pour comparer des AUCs, des Brier scores ou n'importe quelle autre mesure de capacités pronostiques, dès que l'on dispose d'un test valide pour des données de validation externes. On pourrait notamment réutiliser les tests de Brier scores et d'AUCs proposés aux Chapitres IV et V en présence de données censurées. Ils pourraient être complémentaires au test initialement proposé par van de Wiel *et al.* (2009), basé sur la « signed rank statistic » présentée en Section II.5.2 et non adaptée aux données censurées. Une version IPCW de la « signed rank statistic » pourrait

aussi être envisagée, et un nouveau test dérivé, en suivant les mêmes raisonnements que ceux utilisés pour les estimateurs IPCW de l'AUC. L'hypothèse nulle du test de l'approche de van de Wiel *et al.* (2009) ne correspond cependant pas exactement à ce que l'on souhaite tester en général. En effet, l'hypothèse nulle du test que l'on souhaite en général correspond à une question du type : « Parmi les modèles comparés estimés avec mes données, quel est le modèle qui prédit le mieux ? ». La cross-validation quant à elle aide généralement à répondre à une question du type : « En moyenne sur des données similaires à mes données, parmi les modèles estimables comparés, quel modèle prédit le mieux ? ». Enfin, l'approche de van de Wiel *et al.* (2009) aide à répondre à une question du type « Parmi les modèles comparés, quel modèle estimable prédit le mieux pour tous les jeux de données similaires à mes données ? ». Lorsque les échantillons sont grands, il peut souvent être raisonnable de considérer que les réponses aux deux premières questions apportent des informations proches. Pour des raisons de puissance de la procédure de van de Wiel *et al.* (2009), la réponse à la troisième question peut aussi souvent être considérée comme donnant une information proche de la réponse à la première question, qui nous intéresse particulièrement. Un travail est en cours avec Mark van de Wiel et Thomas Gerds, notamment pour mieux comprendre le sens de l'hypothèse nulle du test de l'approche de van de Wiel *et al.* (2009), et à quel point elle est une bonne approximation de celle que l'on souhaite vraiment en pratique.

VI.2.3 Cross-validation « approchée » et procédures d'inférence

Pour éviter la cross-validation usuelle qui est très chronophage avec des modèles conjoints, Commenges *et al.* (2012) ont récemment proposé des méthodes de pénalisation, approximant une cross-validation de type « leave-one-out », pour corriger les biais d'optimisme lors de l'évaluation de modèles pronostiques sans données externes. La méthode permet également de construire des régions de confiance et des tests de comparaison. Proposée pour approcher l'évaluation de modèles pronostiques à partir d'une fonction de risque quelconque (vérifiant des conditions assez faibles), l'approche pourrait être reprise pour le cas spécifique du Brier score. En présence de données censurées, l'estimateur IPCW du Brier score n'étant pas une moyenne de

terme i.i.d, l'adaptation des résultats de Commenges *et al.* (2012) pour définir une pénalisation appropriée est loin d'être évidente. Cependant, en estimant le Brier score en présence de données censurées au moyen des « pseudo-values » du *jacknife* (Andersen *et al.*, 2003), évoquées en Section II.4.2.3, des extensions de la méthode de pénalisation de Commenges *et al.* (2012) pourraient probablement être dérivées. En effet, en utilisant l'estimateur du Brier score proposé par Cortese *et al.* (2013), et en notant le résultat du théorème 3 de Singh et Liu (1990) sur les propriétés des « pseudo-values » du *jacknife* de l'estimateur de Kaplan-Meier, on s'aperçoit qu'on est en présence de conditions proches de celles supposées par Commenges *et al.* (2012) pour dériver les pénalisations. Davantage de réflexions autour de cette idée nous parait donc une perspective de travail intéressante.

VI.2.4 À propos de la prédiction de la démence

Aujourd'hui, l'imagerie médicale et l'utilisation de biomarqueurs tels que ceux mesurant l'accumulation de β-amyloid sont considérées comme des opportunités majeures pour contruire des outils prédictifs de la démence (The 3C Study Group, 2003; Alzheimer's Association, 2012). Il serait intéressant de comparer leurs potentiels prédictifs à celui des tests cognitifs ou d'évaluer leurs apports à des modèles prédictifs basés sur des tests cognitifs.

Pour la prédiction dynamique du risque de démence, un modèle conjoint pour données longitudinales multivariées et risques compétitifs est en cours de développement dans l'équipe. Il pourra être utilisé pour développer un outil de prédiction dynamique combinant l'évolution de plusieurs tests cognitifs et dont les capacités pronostiques pourraient être évaluées avec les outils développés au chapitre V.

Dans tous les cas, quelle que soit l'information utilisée pour construire des outils pronostiques de la démence, il sera toujours important d'évaluer les capacités pronostiques de tels outils pronostiques de manière rigoureuse, en traitant les problèmes méthodologiques liés aux données censurées et au risque concurrent majeur de décès sans démence chez les sujets âgés.

VI.3 Conclusion générale

Au cours de cette thèse, nous avons développé des méthodes d'évaluation et de comparaison de capacités pronostiques de marqueurs adaptées aux données censurées, aux risques compétitifs et aux marqueurs dépendants du temps. L'ensemble de ces méthodes a été implémenté en R, et le package R timeROC a été créé afin de permettre leur diffusion. Elles ont également fait l'objet de plusieurs applications à la prédiction de la démence, qui était la principale motivation des travaux de cette thèse. Au Chapitre V, l'application des méthodes proposées a d'ailleurs permis de montrer que les capacités pronostiques de tests cognitifs étaient très bonnes.

Dans le cadre de mon postdoctorat, j'espère poursuivre ces travaux afin de proposer des procédures d'inférence corrigées pour l'optimisme dans le cadre d'une validation interne.

Bibliographie

Aalen, O. (1975). *Statistical inference for a family of counting processes*. Thèse de doctorat, University of California, Berkeley.

Aalen, O., Borgan, Ø., Gjessing, H. K. et Gjessing, S. (2008). *Survival and event history analysis : a process point of view*. Springer.

Aalen, O. O., Andersen, P. K., Borgan, Ø., Gill, R. D. et Keiding, N. (2009). History of applications of martingales in survival analysis. *Electronic Journal for History of Probability and Statistics*, 5(1):1–28.

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

Airola, A., Pahikkala, T., Waegeman, W., De Baets, B. et Salakoski, T. (2011). An experimental comparison of cross-validation techniques for estimating the area under the roc curve. *Computational Statistics & Data Analysis*, 55(4):1828–1844.

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.

Akritas, M. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics*, 22:1299–1327.

Alzheimer's Association (2012). 2012 Alzheimer's disease facts and figures. *Alzheimer's and Dementia*, 8(2):131 – 168.

Amieva, H., Jacqmin-Gadda, H., Orgogozo, J., Le Carret, N., Helmer, C., Letenneur, L., Barberger-Gateau, P., Fabrigoule, C. et Dartigues, J. (2005). The 9 year cognitive decline before dementia of the Alzheimer type : a prospective population-based study. *Brain*, 128(5):1093.

Amieva, H., Le Goff, M., Millet, X., Orgogozo, J. M., Pérès, K., Barberger-Gateau, P., Jacqmin-Gadda, H. et 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. et Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New york : Springer Verlag, New York.

Andersen, P. K., Geskus, R. B., de Witte, T. et Putter, H. (2012). Competing risks in epidemiology : possibilities and pitfalls. *International journal of epidemiology*, 41(3):861–870.

Andersen, P. K. et Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11(2):91–115.

Andersen, P. K. et Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in medicine*, 31(11-12):1074–1088.

Andersen, P. K., Klein, J. P. et Rosthøj, S. (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1):15–27.

Andersen, P. K. et Perme, M. P. (2010). Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, 19(1):71–99.

Andersen, P. K. et Skovgaard, L. T. (2010). *Regression with linear predictors*. Springer.

Andrieu, S., Ousset, P.-J., Coley, N., Ouzid, M., Mathiex-Fortunet, H. et Vellas, B. (2008). Guidage study : A 5-year double blind, randomised trial of egb 761 for the prevention of alzheimers disease in elderly subjects with memory complaints. i. rationale, design and baseline data. *Current Alzheimer Research*, 5(4):406–415.

Arlot, S. et Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Austin, P. C. et Steyerberg, E. W. (2013). Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Statistics in Medicine*.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. *Unpublished technical report, University of California, Berkeley*.

Beyersmann, J., Termini, S. D. et Pauly, M. (2012). Weak Convergence of the Wild Bootstrap for the Aalen–Johansen Estimator of the Cumulative Incidence Function of a Competing Risk. *Scandinavian Journal of Statistics, in press*.

Blanche, P., Dartigues, J.-F. et Jacqmin-Gadda, H. (2013a). Estimating and Comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*.

Blanche, P., Latouche, A. et Viallon, V. (2013b). Time-dependent auc with right-censored data : a survey study. *In* Risk Assessment and Evaluation of Predictions, éditeur : *Lee, M-L and Gail, G. and Cai, T. and Pfeiffer, R. and Gandy, A.* Springer.

Box, G. (1979). Robustness in the strategy of scientific model building. *In* Robustness in statistics : proceedings of a workshop, éditeur : *Launer, RL and Wilkinson GN*. New York : Academic press.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

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

Cai, T., Gerds, T., Zheng, Y. et Chen, J. (2011). Robust prediction of t-year survival with data from multiple studies. *Biometrics*, 67(2):436–444.

Cai, T. et Pepe, M. (2002). Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *Journal of the American Statistical Association*, 97(460):1099–1107.

Cai, T., Pepe, M., Zheng, Y., Lumley, T. et Jenny, N. (2006). The sensitivity and specificity of markers for event times. *Biostatistics*, 7(2):182.

Chambless, L. et Diao, G. (2006). Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in Medicine*, 25(20):3474–3486.

Chambless, L. E., Cummiskey, C. P. et Cui, G. (2011). Several methods to assess improvement in risk prediction models : extension to survival analysis. *Statistics in medicine*, 30(1):22–38.

Chiang, C. et Hung, H. (2010). Non-parametric estimation for time-dependent AUC. *Journal of Statistical Planning and Inference*, 140(5):1162–1174.

Commenges, D. et Gégout-Petit, A. (2007). Likelihood for generally coarsened observations from multistate or counting process models. *Scandinavian journal of statistics*, 34(2):432–450.

Commenges, D., Proust-Lima, C., Samieri, C. et Liquet, B. (2012). A universal approximate cross-validation criterion and its asymptotic distribution. *arXiv preprint arXiv :1206.1753*.

Cortese, G. et Andersen, P. K. (2010). Competing risks and time-dependent covariates. *Biometrical Journal*, 52(1):138–158.

Cortese, G., Gerds, T. A. et Andersen, P. K. (2013). Comparing predictions among competing risks models with time-dependent covariates. *Statistics in Medicine*, 32(18):3089–3101.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B*, pages 187–220.

Dabrowska, D. M. (1992). Variable bandwidth conditional Kaplan-Meier estimate. *Scandinavian journal of statistics*, pages 351–361.

Dantan, E., Joly, P., Dartigues, J.-F. et Jacqmin-Gadda, H. (2011). Joint model with latent state for longitudinal and multistate data. *Biostatistics*, 12(4):723–736.

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

Datta, S., Bandyopadhyay, D. et 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.

DeLong, E. R., DeLong, D. M. et Clarke-Pearson, D. L. (1988). Comparing the Areas Under Two or More correlated Receiver Operating Characteristic Curves : A Nonparametric Approach. *Biometrics*, 44(3):837–845.

Di Termini, S., Hieke, S., Schumacher, M. et Beyersmann, J. (2012). Nonparametric inference for the cumulative incidence function of a competing risk, with an emphasis on confidence bands in the presence of left-truncation. *Biometrical Journal*, 54(4):568–578.

Dodd, L. E. et Pepe, M. S. (2003). Semiparametric regression for the area under the Receiver Operating Characteristic curve. *Journal of the American Statistical Association*, 98(462):409–417.

D'Agostino, R. et Nam, B.-H. (2004). Evaluation of the performance of survival analysis models : discrimination and calibration measures. *Handbook of statistics*, 23:1–26.

Efron, B. et Tibshirani, R. (1997). Improvements on cross-validation : the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.

Fine, J. P. et 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.

Fisher, L. D. et Lin, D. (1999). Time-dependent covariates in the cox proportional-hazards regression model. *Annual review of public health*, 20(1):145–157.

Folstein, M., Folstein, S. et P, M. (1975). "Mini-Mental State." A practical method for grading the cognitive decline state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198.

Foucher, Y. et Danger, R. (2012). Time Dependent ROC Curves for the Estimation of True Prognostic Capacity of Microarray Data. *Statistical Applications in Genetics and Molecular Biology*, 11(6).

Foucher, Y., Giral, M., Soulillou, J. et Daures, J. (2010). Time-dependent ROC analysis for a three-class prognostic with application to kidney transplantation. *Statistics in Medicine*, 29(30):3079–3087.

Frydman, H., Gerds, T., Grøn, R. et Keiding, N. (2013). Nonparametric estimation in an "illness-death" model when all transition times are interval censored. *Biometrical Journal*.

Frydman, H. et Szarek, M. (2009). Nonparametric estimation in a markov "illness–death" process from interval censored observations with missing intermediate transition status. *Biometrics*, 65(1): 143–151.

Gail, M. H. (2008). Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *Journal of the National Cancer Institute*, 100(14):1037–1041.

Gallez, C. (2005). Rapport sur la maladie d'alzheimer et les maladies apparentées. Rapport technique, Office Parlementaire d'Évaluation des Politiques de Santé, Bibliothèque des Rapports Publics.

Geman, S., Bienenstock, E. et Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58.

Gerds, T., Andersen, P. et Kattan, M. (2013a). Calibration plots for risk prediction models in the presence of competing risks. Rapport technique 9, University of Copenhagen, Department of Biostatistics.

Gerds, T. et 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.

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

Gerds, T. A., Kattan, M. W., Schumacher, M. et Yu, C. (2013b). Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in medicine*, 32(13):2173–2184.

Gerds, T. A., Scheike, T. H. et Andersen, P. K. (2012). Absolute risk regression for competing risks : interpretation, link functions, and prediction. *Statistics in Medicine*, 31(29):3921–3930.

Gerds, T. A. et Schumacher, M. (2007). Efron-type measures of prediction error for survival analysis. *Biometrics*, 63(4):1283–1287.

Gerds, T. A. et van de Wiel, M. A. (2011). Confidence scores for prediction models. *Biometrical Journal*, 53(2):259–274.

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. et 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. et Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18:2529–2545.

Graw, F., Gerds, T. A. et Schumacher, M. (2009). On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis*, 15(2):241–255.

Gray, R. J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of statistics*, pages 1141–1154.

Harrell, F., Lee, K. L. et Mark, D. B. (1996). Tutorial in biostatistics multivariable prognostic models : issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15:361–387.

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

Heagerty, P., Lumley, T. et Pepe, M. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344.

Heagerty, P. et Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105.

Hedeker, D. et Gibbons, R. D. (2006). *Longitudinal data analysis*, volume 451. Wiley. com.

Henderson, R., Diggle, P. et Dobson, A. (2002). Identification and efficacy of longitudinal markers for survival. *Biostatistics*, 3(1):33–50.

Hilden, J. et Gerds, T. A. (2013). A note on the evaluation of novel biomarkers : do not rely on integrated discrimination improvement and net reclassification index. *Statistics in medicine*.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325.

Horvitz, D. G. et Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Hosmer, D. W. et Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics-Theory and Methods*, 9(10):1043–1069.

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

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

Isaacs, B. et Kennie, A. (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, H., Kogalur, U. B., Blackstone, E. H. et Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, pages 841–860.

Jacqmin-Gadda, H., Alperovitch, A., Montlahuc, C., Commenges, D., Leffondre, K., Dufouil, C., Elbaz, A., Tzourio, C., Ménard, J., Dartigues, J.-F. *et al.* (2013a). 20-year prevalence projections for dementia and impact of preventive policy about risk factors. *European journal of epidemiology*, pages 1–10.

Jacqmin-Gadda, H., Blanche, P., Chary, E., L., L., Amieva, H. et Dartigues, J. (2013b). Prognostic score to predict risk of dementia over 10 years accounting for death competing risk. *soumis*.

Jacqmin-Gadda, H., Blanche, P., Chary, E., Touraine, C. et Dartigues, J. (2013c). ROC curve estimation for time-to-event with semi-competing risks and interval censoring. *soumis*.

Jacqmin-Gadda, H., Fabrigoule, C., Commenges, D. et Dartigues, J. (1997). A 5-year longitudinal study of the Mini-Mental State Examination in normal aging. *American Journal of Epidemiology*, 145(6):498.

Jacqmin-Gadda, H., Thiébaut, R., Dartigues, J.-F. *et al.* (2004). Joint modeling of quantitative longitudinal data and censored survival time. *Revue d'épidémiologie et de santé publique*, 52(6):502–10.

Joly, P., Commenges, D., Helmer, C. et Letenneur, L. (2002). A penalized likelihood approach for an illness–death model with interval-censored data : application to age-specific incidence of dementia. *Biostatistics*, 3(3):433.

Kalbfleisch, J. D. et Prentice, R. L. (2002). *The statistical analysis of failure time data*. Second edition : Wiley-Interscience.

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

Kerr, K. et Pepe, M. (2011). Joint modeling, covariate adjustment, and interaction : contrasting notions in risk prediction models and risk prediction performance. *Epidemiology*, 22(6):805–812.

Kowalski, J. et Tu, X. M. (2008). *Modern applied U-statistics*. Wiley.

Lawless, J. F. et Yuan, Y. (2010). Estimation of prediction error for survival models. *Statistics in medicine*, 29(2):262–274.

Liang, K.-Y. et Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Lin, D., Fleming, T. et 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. et Scheike, T. (2006). *Dynamic regression models for survival data*. Springer.

McIntosh, M. et Pepe, M. (2002). Combining several screening tests : optimality of the risk score. *Biometrics*, 58(3):657–664.

Mogensen, U. B. et Gerds, T. A. (2013). A random forest approach for competing risks based on pseudo-values. *Statistics in medicine*.

Newcombe, R. G. (2001). Logit confidence intervals and the inverse sinh transformation. *The American Statistician*, 55(3):200–202.

Newcombe, R. G. (2012). *Confidence Intervals for Proportions and Related Measures of Effect Size*. CRC Press.

Nicolaie, M., Houwelingen, J., Witte, T. et Putter, H. (2013a). Dynamic prediction by landmarking in competing risks. *Statistics in medicine*, 32(12):2031–2047.

Nicolaie, M., van Houwelingen, J., de Witte, T. et Putter, H. (2013b). Dynamic pseudo-observations : A robust approach to dynamic prediction in competing risks. *Biometrics*, in press.

Parast, L. et Cai, T. (2013). Landmark risk prediction of residual life for breast cancer survival. *Statistics in medicine*, 32(20):3459–3471.

Parast, L., Cheng, S.-C. et 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.

Paul, P., Pennell, M. L. et Lemeshow, S. (2013). Standardizing the power of the hosmer–lemeshow goodness of fit test in large data sets. *Statistics in Medicine*, 32(1):67–80.

Pencina, M. J., D'Agostino, R. B. et Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker : from area under the roc curve to reclassification and beyond. *Statistics in medicine*, 27(2):157–172.

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

Pepe, M., Fang, J., Feng, Z., Gerds, T. et Hilden, J. (2013). The Net Reclassification Index (NRI) : a Misleading Measure of Prediction Improvement with Miscalibrated or Overfit Models. *UW Biostatistics Working Paper Series*.

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

Pepe, M., Zheng, Y., Jin, Y., Huang, Y., Parikh, C. et Levy, W. (2008a). Evaluating the ROC performance of markers for future events. *Lifetime Data Analysis*, 14(1):86–113.

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

Pepe, M. S., Janes, H., Longton, G., Leisenring, W. et Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American journal of epidemiology*, 159(9):882–890.

Proust-Lima, C., Amieva, H., Dartigues, J.-F. et 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., Mbéry, S., Taylor, J. et Jacqmin-Gadda, H. (2012). Joint latent class models for longitudinal and time-to-event data : A review. *Statistical Methods in Medical Research*.

Proust-Lima, C. et Taylor, J. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA : a joint modeling approach. *Biostatistics*, 10(3):535.

Putter, H., Fiocco, M. et Geskus, R. (2007). Tutorial in biostatistics : competing risks and multi-state models. *Statistics in medicine*, 26(11):2389–2430.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829.

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

Robins, J. M. et Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. *In* AIDS Epidemiology-Methodological Issues, éditeur : *Jewell, N. and Dietz, K. and Farewell, V.*, pages 297–331.

Saha, P. et Heagerty, P. (2010). Time-dependent predictive accuracy in the presence of competing risks. *Biometrics*, 66(4):999–1011.

Saporta, G. (2006). *Probabilités, analyses des données et statistiques*. Editions Technip.

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

Scheike, T., Zhang, M. et Gerds, T. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, 95:205–220.

Schemper, M. et Henderson, R. (2000). Predictive accuracy and explained variation in cox regression. *Biometrics*, 56(1):249–255.

Schoop, R., Graf, E. et Schumacher, M. (2008). Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics*, 64(2):603–610.

Schoop, R., Schumacher, M. et Graf, E. (2011). Measures of prediction error for survival data with longitudinal covariates. *Biometrical Journal*, 53(2):275–293.

Schumacher, M., Binder, H. et Gerds, T. (2007). Assessment of survival prediction models based on microarray data. *Bioinformatics*, 23(14):1768–1774.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1(1):43–62.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York : John Wiley & Sons.

Singh, K. et Liu, R. Y. (1990). On the validity of the jack-knife procedure. *Scandinavian journal of statistics*, pages 11–21.

Snitz, B. E., O'Meara, E. S., Carlson, M. C., Arnold, A. M., Ives, D. G., Rapp, S. R., Saxton, J., Lopez, O. L., Dunn, L. O., Sink, K. M. *et al.* (2009). Ginkgo biloba for preventing cognitive decline in older adults. *JAMA : the journal of the American Medical Association*, 302(24):2663–2670.

Song, X. et Zhou, X. (2008). A semiparametric approach for the covariate-specific ROC curve with survival outcome. *Statistica Sinica*, 18:947–965.

Stephan, B. C., Kurth, T., Matthews, F. E., Brayne, C. et Dufouil, C. (2010). Dementia risk prediction in the population : are screening models accurate ? *Nature Reviews Neurology*, 6(6):318–326.

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

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J. et Kattan, M. W. (2010). Assessing the performance of prediction models : a framework for some traditional and novel measures. *Epidemiology*, 21(1):128.

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis*, 45(1):89–103.

Stute, W. (1995). The central limit theorem under random censorship. *The Annals of Statistics*, pages 422–439.

Thas, O., Neve, J. D., Clement, L. et Ottoy, J.-P. (2012). Probabilistic index models. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 74(4):623–671.

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.

Therneau, T. M. (2000). *Modeling survival data : extending the Cox model*. Springer.

Tian, L., Cai, T., Goetghebeur, E. et Wei, L. (2007). Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika*, 94(2):297–311.

Touraine, C., Gerds, T. et Joly, P. (2013). The SmoothHazard package for R : Fitting regression models to interval-censored observations of illness-death models. Rapport technique, University of Copenhagen, Department of Biostatistics.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22.

Tsiatis, A. (2006). *Semiparametric theory and missing data*. New York : Springer Verlag.

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

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. et Wei, L. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117.

Uno, H., Cai, T., Tian, L. et Wei, L. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537.

Uno, H., Tian, L., Cai, T., Kohane, I. S. et Wei, L. (2013). A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data. *Statistics in Medicine*, 32(14):2430–2442.

van de Wiel, M. A., Berkhof, J. et van Wieringen, W. N. (2009). Testing the prediction error difference between 2 predictors. *Biostatistics*, 10(3):550–560.

van der Laan, M. et Robins, J. (2003). *Unified methods for censored longitudinal data and causality*. New York : Springer Verlag.

van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge University Press.

van Houwelingen, H. et 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.

Vellas, B., Aisen, P. S., Sampaio, C., Carrillo, M., Scheltens, P., Scherrer, B., Frisoni, G. B., Weiner, M., Schneider, L., Gauthier, S. *et al.* (2011). Prevention trials in alzheimer's disease : an eu-us task force report. *Progress in neurobiology*, 95(4):594–600.

Viallon, V. et Latouche, A. (2011). Discrimination measures for survival outcomes : Connection between the AUC and the predictiveness curve. *Biometrical Journal*, 53:217–236.

Wechsler, D. (1981). Wechsler adult intelligence scale (rev. ed.). *New York : Psychological Corporation*.

Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H. et Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18): 1837–1847.

Wolbers, M., Blanche, P., Koller, M. T., Witteman, J. C. et Gerds, T. A. (2013). Concordance for prognostic models with competing risks. *under revision*.

Wolbers, M., Koller, M. T., Witteman, J. C. et Steyerberg, E. W. (2009). Prognostic models with competing risks : methods and application to coronary risk prediction. *Epidemiology*, 20(4):555–561.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295.

Zheng, Y., Cai, T., Jin, Y. et Feng, Z. (2012a). Evaluating Prognostic Accuracy of Biomarkers under Competing Risk. *Biometrics*, 68(2):388–396.

Zheng, Y., Parast, L., Cai, T. et Brown, M. (2012b). Evaluating incremental values from new predictors with net reclassification improvement in survival analysis. *Lifetime data analysis*, pages 1–21.

Zhou, Q. M., Zheng, Y. et Cai, T. (2013). Subgroup specific incremental value of new markers for risk prediction. *Lifetime data analysis*, pages 1–28.

Zou, G. et Yue, L. (2013). Using confidence intervals to compare several correlated areas under the receiver operating characteristic curves. *Statistics in medicine*.

Table des figures

Liste des tableaux

VII. Annexe

Time-dependent AUC with right-censored data: a survey study

Paul Blanche, Aurélien Latouche, Vivian Viallon

Abstract The ROC curve and the corresponding AUC are popular tools for the evaluation of diagnostic tests. They have been recently extended to assess prognostic markers and predictive models. However, due to the many particularities of time-to-event outcomes, various definitions and estimators have been proposed in the literature. This review article aims at presenting the ones that accommodate to right-censoring, which is common when evaluating such prognostic markers.

1 Introduction

In the medical literature, a variety of general criteria have been used to assess diagnostic tests [24, 14]. Among them, the ROC curve and the area under it – the AUC – are popular tools, originally aimed at evaluating the discriminant power of continuous diagnostic tests. In this simple situation, the outcome status D is a binary variable (typically, $D = 1$ for cases and $D = 0$ for controls) and the ROC curve for a continuous diagnostic test X plots the true positive rate, or sensitivity, $\mathrm{TPR}(c) = \mathbb{P}(X > c|D = 1)$ against the false positive rate, or one minus the specificity, $\mathrm{FPR}(c) = \mathbb{P}(X > c|D = 0)$, when making threshold c vary. The AUC, which is the area under this curve, is a commonly used summary measure of the information contained in the sequences $(\mathrm{TPR})_{c\in\mathbb{R}}$ and $(\mathrm{FPR})_{c\in\mathbb{R}}$. As such, it inherits some of the properties of true and false positive rates. In particular, it is not affected by the disease prevalence (unlike positive and negative predictive values) and it can be evaluated from random samples of cases and controls. It also has a nice interpretation since it corresponds to the probability that the marker value of a randomly selected case exceeds that of a randomly selected control.

Paul Blanche
Univ. Bordeaux, ISPED and INSERM U897, Bordeaux, FRANCE e-mail: `Paul.Blanche@isped.u-bordeaux2.fr`

Aurélien Latouche
Conservatoire national des arts et métiers, Paris, FRANCE e-mail: `aurelien.latouche@cnam.fr`

Vivian Viallon
UMRESTTE (Univ. Lyon 1 and IFSTTAR), Bron, FRANCE e-mail: `viallon@math.univ-lyon1.fr`

The extension of the AUC (and other evaluation criteria) to the setting of prognostic markers has raised several issues. In particular, when evaluating such markers, the outcome status typically changes over time: in a cohort study for instance, patients are diseased-free when entering the study and may develop the disease during the study. This leads to three major differences with the evaluation of diagnostic tests. First, this time-dependent outcome status (which may be defined in several ways, as will be seen in Section 3) naturally implies that sensitivity, specificity, ROC curves, their AUC values and, more generally, any extension of the criteria used in the diagnostic setting, are functions of time as well. Second, the time-to-event, i.e. the time between the entry in the study and the disease onset, is usually *censored* and not fully observed, requiring dedicated inference.Third, the time lag between the entry in the study and the disease onset also leads to two further refinements: (i) the marker can be repeatedly measured over time and (ii) competing events (in addition to censoring) may be observed between the entry and the putative disease onset.
These particularities has led to the development of numerous methods aimed at estimating the time-dependent AUC for prognostic markers. In this paper, we review those that accommodate to right-censoring. Some notations are introduced in the following Section 2. Then, in Section 3 we will present several definitions and estimators of the time-dependent AUC in the "standard" setting of a baseline marker and univariate survival data. Section 4 will cover the case of longitudinal markers which corresponds to the marker being repeatedly measured over time, while we will discuss the competing events setting in Section 5. Finally, concluding remarks will be given in Section 6.

2 Notations

Let T_i and C_i denote survival and censoring times for subject i, $i = 1,\ldots,n$. We further let $Z_i = \min(T_i, C_i)$ and $\delta_i = \mathbb{1}(T_i \leq C_i)$ denote the observed time and the status indicator respectively. We will denote by $D_i(t)$ the time-dependent outcome status for subject i at time t, $t \geq 0$. Several definitions for $D_i(t)$ will be given hereafter, but we will always have $D_i(t) = 1$ if subject i is considered as a case at time t and $D_i(t) = 0$ if subject i is considered as a control at time t.
We will denote by X the marker under study, which can be a single biological marker or several biological markers combined into a predictive model (in this case, it is assumed throughout this article that the predictive model has been constructed on an independent data set; otherwise sub-sampling techniques are needed [21]). Without loss of generality, we will suppose that larger values of X are associated with greater risks (otherwise, X can be recoded to achieve this). We will denote by g and G^{-1} the probability density function and the quantile function of marker X. In Section 3, we assume that marker X is measured once at $t = 0$, and we will denote by X_i the marker value for subject i. In Section 4, which treats the longitudinal setting, the marker is measured repeatedly over time, and we will denote by $X_i(s)$ the marker value at time s for subject i.

3 Time-dependent AUCs in the standard setting

Definitions of time-dependent ROC curves, $\mathrm{ROC}(t)$, follow from definitions of usual ROC curves and thus rely on first defining time-dependent true and false positive rates. For any threshold c, these two functions of time are defined as $\mathrm{TPR}(c,t) = \mathbb{P}(X > c|D(t) = 1)$ and $\mathrm{FPR}(c,t) = \mathbb{P}(X > c|D(t) = 0)$. $\mathrm{ROC}(t)$ then simply plots $\mathrm{TPR}(c,t)$ against $\mathrm{FPR}(c,t)$ making threshold c vary. The time-dependent AUC at time t is then defined as the area under this curve,

$$\mathrm{AUC}(t) = \int_{-\infty}^{\infty} \mathrm{TPR}(c,t) \left| \frac{\partial \mathrm{FPR}(c,t)}{\partial c} \right| dc. \tag{1}$$

As a matter of fact, these definitions deeply rely on that of the outcome status at time t, $D(t)$. Heagerty and Zheng [18] described several definitions of *cases* and *controls* in this survival outcome setting. According to Heagerty and Zheng's terminology and still denoting by T_i survival time for subject i, cases are said to be *incident* if $T_i = t$ is used to define cases at time t, and *cumulative* if $T_i \leq t$ is used instead. Similarly, depending on whether $T_i > \tau$ for a large time $\tau > t$ or $T_i > t$ is used for defining controls at time t, they are said to be *static* or *dynamic* controls. Depending on the definition retained for cases and controls at time t, four definitions of the time-dependent AUC value may be put forward. In the following paragraphs, we will present formulas and estimators for the most commonly used ones and will discuss their respective interests.

3.1 The cumulative dynamic AUC: $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$

The setting of cumulative cases and dynamic controls may be regarded as the most natural choice for planning enrollment criteria in clinical trials or when specific evaluation times are of particular interest. It simply corresponds to defining cases at time t as subjects who experienced the event prior to time t, and controls at time t as patients who were still event-free at time t. In other words, it corresponds to setting $D_i(t) = \mathbb{1}(T_i \leq t)$. Cumulative true positive rates and dynamic false positive rates are then respectively defined as

$$\mathrm{TPR}^{\mathbb{C}}(c,t) = \mathbb{P}(X > c|T \leq t) \quad \text{and} \quad \mathrm{FPR}^{\mathbb{D}}(c,t) = \mathbb{P}(X > c|T > t). \tag{2}$$

The *cumulative*/*dynamic* AUC at time t is then obtained by using these definitions of true and false positive rates in (1). Usually, however, $\mathbb{1}(T \leq t)$ is not observed for all subjects due to the presence of censoring before time t and simple contingency tables can therefore not be used to return estimates of $\mathrm{TPR}^{\mathbb{C}}(c,t)$, $\mathrm{FPR}^{\mathbb{D}}(c,t)$ and $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$. To handle censoring, Bayes' theorem can be used to rewrite $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$ as a function of the conditional survival function $\mathbb{P}(T > t|X = x)$ (see paragraph 3.1.1 below). Other approaches rely on so-called Inverse Probability of Censoring

Weighted (IPCW) estimates (see paragraph 3.1.2 below). Before describing these two approaches in more details below, we shall add that Chambless and Diao [6] developed an alternative method – which will not be described here – based on an idea similar to the one used to derive the Kaplan-Meier estimator of the cumulative distribution function in the presence of censoring. Among all these methods, only those relying on primary estimates of $\mathbb{P}(T > t|X = x)$ (and a recent extension of IPCW estimates proposed in [2]) may account for the dependence between censoring and the marker (since they basically only assume that T and C are independent given X and not that T and C are independent). We refer the reader to [2, 19, 37] for empirical comparisons and illustrations of these various methods.

3.1.1 Methods based on primary estimates of $\mathbb{P}(T > t|X = x)$

Bayes' theorem yields the following expressions for $\mathrm{TPR}^{\mathbb{C}}(c,t)$ and $\mathrm{FPR}^{\mathbb{D}}(c,t)$

$$\mathrm{TPR}^{\mathbb{C}}(c,t) = \frac{\int_c^\infty \mathbb{P}(T \leq t|X = x)g(x)dx}{\mathbb{P}(T \leq t)}, \quad \mathrm{FPR}^{\mathbb{D}}(c,t) = \frac{\int_c^\infty \mathbb{P}(T > t|X = x)g(x)dx}{\mathbb{P}(T > t)}.$$

From (1), it readily follows that

$$\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t) = \int_{-\infty}^{\infty}\int_c^{\infty} \frac{\mathbb{P}(T \leq t|X = x)\mathbb{P}(T > t|X = c)}{\mathbb{P}(T \leq t)\mathbb{P}(T > t)} g(x)g(c)dxdc. \tag{3}$$

Since $\mathbb{P}(T > t) = \int_{-\infty}^{\infty} \mathbb{P}(T > t|X = x)g(x)dx$, any estimator $\widehat{S}_n(t|x)$ of the conditional survival function $\mathbb{P}(T > t|X = x)$ yields an estimator of $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$:

$$\widehat{\mathrm{AUC}}^{\mathbb{C},\mathbb{D}}(t) = \frac{\sum_{i=1}^n \sum_{j=1}^n \widehat{S}_n(t|X_j)[1 - \widehat{S}_n(t|X_i)]\,\mathbb{1}(X_i > X_j)}{\sum_{i=1}^n \sum_{j=1}^n \widehat{S}_n(t|X_j)[1 - \widehat{S}_n(t|X_i)]}.$$

In [6], the authors suggested to use a Cox model to derive estimates $\widehat{S}_n(t|x)$, while one of the methods described in Heagerty *et al.* [17] reduces to using the conditional Kaplan-Meier estimator as in [1]. Some theoretical results for these two methods can be found in [33] and [3, 7, 20] respectively.
We shall add that Viallon and Latouche [37] related $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$ to the quantity $\mathbb{P}(T \leq t|X = G^{-1}(q))$ – a time-dependent version of the predictiveness curve:

$$\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t) = \frac{\int_0^1 q\mathbb{P}(T \leq t|X = G^{-1}(q))dq - [\mathbb{P}(T \leq t)]^2/2}{\mathbb{P}(T > t)\mathbb{P}(T \leq t)}.$$

This confirms that most standard statistical summaries of predictability and discrimination can be derived from the predictiveness curve, as pointed out in [14, 15].

3.1.2 Methods based on IPCW estimators

In [19] and [36], the authors independently suggested to use IPCW-type estimates:

$$\widehat{\mathrm{TPR}}^{\mathbb{C}}(c,t) = \frac{\sum_{i=1}^{n} \mathbb{1}(X_i > c, Z_i \leq t)\frac{\delta_i}{n\hat{S}_C(Z_i)}}{\sum_{i=1}^{n} \mathbb{1}(Z_i \leq t)\frac{\delta_i}{n\hat{S}_C(Z_i)}}, \quad \widehat{\mathrm{FPR}}^{\mathbb{D}}(c,t) = \frac{\sum_{i=1}^{n} \mathbb{1}(X_i > c, Z_i > t)}{\sum_{i=1}^{n} \mathbb{1}(Z_i > t)},$$

where $\hat{S}_C(\cdot)$ is the Kaplan-Meier estimator of the survival function of the censoring time C. The expression of the false positive rate estimator is more compact because weights all equal $1/(n\hat{S}_C(t))$ under the assumption of independence between C and X, and then vanish. $\widehat{\mathrm{FPR}}^{\mathbb{D}}(c,t)$ corresponds to 1 minus the empirical distribution function of X among individuals for whom $Z_i > t$. In the absence of censoring before time t, $\widehat{\mathrm{TPR}}^{\mathbb{C}}(c,t)$ also reduces to the usual empirical version of $\mathrm{TPR}^{\mathbb{C}}(c,t)$, *i.e.*, 1 minus the empirical distribution function of X among individuals for whom $T_i \leq t$. It can be shown (see [19, 30]) that an estimator of $\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(t)$ is then given by

$$\widehat{\mathrm{AUC}}^{\mathbb{C},\mathbb{D}}(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}(Z_i \leq t)\mathbb{1}(Z_j > t)\mathbb{1}(X_i > X_j)\frac{\delta_i}{\hat{S}_C(Z_i)\hat{S}_C(t)}}{n^2\hat{S}(t)[1-\hat{S}(t)]},$$

where $\hat{S}(t)$ is the Kaplan-Meier estimator of $\mathbb{P}(T > t)$.

Theoretical guarantees for these estimators can be found in [19] and [36]. These estimators are in a sense more flexible than those presented in paragraph 3.1.1 above: they are model-free and they do not rely on any bandwidth selection (unlike the estimator of Heagerty *et al.* [17] for instance, which is based on a local version of the Kaplan-Meier estimator). However, when censoring may depend on marker X, quantities like the conditional survival function of C given the marker X, $S_C(\cdot|X)$, have to be estimated [2], which implies either to work under some (semi-)parametric model or the selection of some parameter if nonparametric estimation is prefered.

3.2 The incident static AUC: $\mathrm{AUC}^{\mathbb{I},\mathbb{S}}(t)$

Using the dynamic definition of controls, the control group varies with time, and so does the x-axis of the corresponding ROC curves: in situations where trends over time are of particular interest, this renders their interpretation more difficult (since such trends may be partly due to changing control groups). Moreover, the group of static controls is interesting in that it tries to mimic the group of individuals who never develop the disease, which can be seen as an ideal control group in some situations. In particular, patients with preclinical diseases are eliminated from the control group as far as possible, if τ is large enough.
Regarding cases, the incident definition has several advantages over the dynamic definition [26]. The cumulative TPR does not distinguish between events that occur

early versus late, and it shows redundant information over time (since early events are also included in the cumulative TPR for late evaluation times). Moreover, as pointed out by [4], the cumulative TPR can be computed from the incident TPR when the distribution of the event time is known.
Putting all this together, several authors have proposed estimators of the time-dependent ROC curve relying on the incident definition of cases and static definition of controls. Standard numerical integration techniques are then used to compute an estimate for $\mathrm{AUC}^{\mathbb{I},\mathbb{S}}(t)$ from the estimators of the ROC curve.

Incident true positive rates and static false positive rates are defined as

$$\mathrm{TPR}^{\mathbb{I}}(c,t) = \mathbb{P}(X > c | T = t) \quad \text{and} \quad \mathrm{FPR}^{\mathbb{S}}_{\tau}(c) = \mathbb{P}(X > c | T > \tau). \tag{4}$$

Applying Bayes' theorem, they can further be rewritten (see, *e.g.*, [33])

$$\mathrm{TPR}^{\mathbb{I}}(c,t) = \frac{\int_c^{\infty} f(t|x)g(x)dx}{\int_{-\infty}^{\infty} f(t|x)g(x)dx} \quad \text{and} \quad \mathrm{FPR}^{\mathbb{S}}_{\tau}(c) = \frac{\int_c^{\infty} \mathbb{P}(T > \tau | X = x)g(x)dx}{\int_{-\infty}^{\infty} \mathbb{P}(T > \tau | X = x)g(x)dx},$$

where $f(t|x) = \partial \mathbb{P}(T \leq t | X = x)/\partial t$ is the conditional density function of T given $X = x$. Under a standard Cox model of the form $\lambda(t;X) = \lambda_0(t)\exp(\beta X)$ – here $\lambda(t;X)$ stands for the conditional hazard rate of T given X while λ_0 is the unspecified baseline hazard rate – Song and Zhou [33] deduced that

$$\begin{aligned}
\mathrm{TPR}^{\mathbb{I}}(c,t) &= \frac{\int_c^{\infty} \exp(\beta x)\exp\{-\Lambda_0(t)\exp(\beta x)\}g(x)dx}{\int_{-\infty}^{\infty} \exp(\beta x)\exp\{-\Lambda_0(t)\exp(\beta x)\}g(x)dx} \\
\mathrm{FPR}^{\mathbb{S}}_{\tau}(c) &= \frac{\int_c^{\infty} \exp\{-\Lambda_0(\tau)\exp(\beta x)\}g(x)dx}{\int_{-\infty}^{\infty} \exp\{-\Lambda_0(\tau)\exp(\beta x)\}g(x)dx},
\end{aligned}$$

where $\Lambda_0(t) = \int_{-\infty}^{t} \lambda_0(u)du$ is the cumulative baseline hazard function. Estimation of $\mathrm{TPR}^{\mathbb{I}}(c,t)$ and $\mathrm{FPR}^{\mathbb{S}}_{\tau}(c)$ can then be achieved by plug-in methods. We shall add that Song and Zhou actually considered a slightly more general set-up where additional covariates can be accounted for.

In [18], Heagerty and Zheng adopted a slightly different approach. To estimate $\mathrm{TPR}^{\mathbb{I}}(c,t)$, they used a (possibly time-varying-coefficients) Cox model of the form $\lambda(t;X) = \lambda_0(t)\exp(\beta(t)X)$ in combination with the fact that the distribution of $X \cdot \exp(\beta X)$ for subjects in the risk set at time t is equal to the conditional distribution of X given $T = t$ (see, *e.g.,* [38]). Setting $R(t) = \{i : Z_i \geq t\}$, this leads to

$$\widehat{\mathrm{TPR}}^{\mathbb{I}}(c,t) = \frac{\sum_{i \in R(t)} \mathbb{I}(X_i > c)\exp\{\beta(t)X_i\}}{\sum_{i \in R(t)} \exp\{\beta(t)X_i\}}.$$

As for the estimation of $\mathrm{FPR}^{\mathbb{S}}_{\tau}(c)$, they proposed a model-free approach using the empirical distribution function for marker values among the control set $S_\tau := \{i : Z_i > \tau\}$. Namely, denoting by n_τ the cardinality of S_τ, they proposed

$$\widehat{\mathrm{FPR}}_{\tau}^{\mathbb{S}}(c) = \frac{1}{n_\tau} \sum_{i \in S_\tau} \mathbb{I}(X_i > c),$$

which is $\widehat{\mathrm{FPR}}^{\mathbb{D}}(c,t)$ of Section 3.1.2, except τ is used instead of t.
Cai *et al.* [4] proposed another approach in the context of longitudinal markers; it will be described in more details in Section 4. In addition, two non parametric approaches were recently proposed (see [32] and [29]).
Note also that estimators for the time-dependant incident/dynamic AUC, $\mathrm{AUC}^{\mathbb{I},\mathbb{D}}(t)$, can be obtained by simply replacing τ by t in the definitions of $\widehat{\mathrm{FPR}}_{\tau}^{\mathbb{S}}(c)$ above [18]. A global accuracy measure has further been derived from the definition of $\mathrm{AUC}^{\mathbb{I},\mathbb{D}}(t)$, which is particularly appealing when no a priori time t is identified and/or when trends over time are not of interest [18].

4 Time-dependent ROC curve and AUCs with longitudinal marker

In this section, we review extensions of the above estimators for longitudinally collected subject measurements. For instance some authors would assess the discrimination performance of CD4 counts repeatedly measured every week on time from seroconvertion to progression to AIDS [42]. Therefore, time-dependent sensitivities and specificities have been extended to deal with the fact that (i) the time at which marker X is measured can vary and (ii) marker can be repeatedly measured on the same subject. Let s denote the timing of marker measurement and $X(s)$ the marker value at time s. For $t \geq s$, Zheng and Heagerty [42] extended *cumulative/dynamic* definitions

$$\mathrm{TPR}^{\mathbb{C}}(c,s,t) = \mathbb{P}(X(s) > c | T \in [s,t]), \qquad \mathrm{FPR}^{\mathbb{D}}(c,s,t) = \mathbb{P}(X(s) > c | T > t).$$

For a fixed time $\tau \geq s$, Zheng and Heagerty [40] extended *incident/static* definitions:

$$\mathrm{TPR}^{\mathbb{I}}(c,s,t) = \mathbb{P}(X(s) > c | T = t) \quad \text{and} \quad \mathrm{FPR}^{\mathbb{S}}(c,s,\tau) = \mathbb{P}(X(s) > c | T > \tau).$$

Although others approaches have been proposed to estimate these quantities, we only review estimators that deal with censored data here. We should also mention that in this longitudinal context, estimators of the AUC are obtained by numerically integrating estimators of the ROC curve.

4.1 Cumulative dynamic estimators with longitudinal marker

Rizopoulos [27] recently proposed a joint modeling approach. The marker trajectory is modeled by usual linear mixed model for longitudinal data, and a parametric pro-

portional hazard is used to model the time-to-event given the marker trajectory. The two submodels are linked with shared random effects to capture the intra-subject correlation. Standard maximum likelihood estimation is used to fit the joint model. Then, $\text{TPR}^{\mathbb{C}}$ and $\text{FPR}^{\mathbb{D}}$ are computed from the estimated parameters and Monte Carlo simulations are used to make inference. As this approach is fully parametric, its main advantage is its efficiency. This approach also allows censoring to depend on the marker [35]. The counterpart is that the parametric model must be carefully chosen, and checking model fit is not straightforward.
A more flexible methodology was proposed in [42], with fewer parametric assumptions. Setting $T^* = T - s$, the "residual failure time", and $t^* = t - s$, they rewrote

$$\text{TPR}^{\mathbb{C}}(c,s,t^*) = \frac{1 - F_{X|s}(c) - S(c,t^*|s)}{1 - S(t^*|s)}, \qquad \text{FPR}^{\mathbb{D}}(c,s,t^*) = 1 - \frac{S(c,t^*|s)}{S(t^*|s)},$$

with $F_{X|s}(c) = \mathbb{P}(X(s) < c|s, T^* > 0)$ the conditional distribution of marker given measurement time, $S(t^*|s) = \mathbb{P}(T^* > t^*|s, T^* > 0)$ the survival probability for individuals who survived beyond s and $S(c,t^*|s) = \mathbb{P}(X(s) > c, T^* > t^*|s, T^* > 0)$. They proposed to estimate $F_{X|s}(c)$ with the semiparametric estimator proposed by Heagerty and Pepe [16]. Therefore, only the location and scale of the conditional distribution of marker given measurement time are parametrized. To estimate the survival terms $S(c,t^*|s)$ and $S(t^*|s)$, they proposed the use of a "partly conditional" hazard function to model the residual failure time $T^* = T - s$. For subject i at measurement time s_{ik}, this function is modeled by

$$\lambda_{ik}(t^*|X_i(s_{ik}), 0 \leq s_{ik} \leq T_i) = \lambda_0(t^*) \exp\left[\beta(t^*)X_i(s_{ik}) + \alpha^T f(s_{ik})\right]$$

where $f(s)$ are vectors of spline basis functions evaluated at measurement time s, and $\lambda_0(t^*)$ is left unspecified. Estimators of $\beta(\cdot)$ and α have been previously proposed [41]. As this approach is semiparametric, its main advantage is its flexibility. However, by contrast to the approach of Rizopoulos [27], this one is less efficient and does not allow marker-dependent censoring.

4.2 Incident static estimators with longitudinal marker

Several authors consider the incident/static definition of AUC [31, 12, 4, 34]. However, censored data are only accounted for by Cai *et al.* [4] who proposed to model

$$\begin{aligned} \text{TPR}^{\mathbb{I}}(c,s,t) &= g_D\left(\eta_\alpha(t,s) + h(c)\right), \quad t \leq \tau \\ \text{FPR}^{\mathbb{S}}(c,s,\tau) &= g_\tau\left(\xi_a(s) + d(c)\right) \end{aligned}$$

where g_D and g_τ are specified inverse link functions and $h(\cdot)$ and $d(\cdot)$ are unspecified baseline functions of threshold c. The dependence in time is parametrically modeled by $\eta_\alpha(t,s) = \alpha^T \eta(t,s)$ and $\xi_a(s) = a^T \xi(s)$ where $\eta(t,s)$ and $\xi(s)$ are vectors of polynomial or spline basis functions. These models are semiparametric with

respect to the marker distribution in cases and nonparametric in regards to controls. As pointed out in [26], this model is very flexible as it does not specify any distributional form for the distribution of the marker given the event-time, but only model the effect of time-to-event on the marker distribution with a parametric form. Model estimation is performed by solving some estimating equations and large sample theory was established allowing a resampling method to construct confidence bands and make inference [4]. Interestingly, the authors of [4] also showed that covariates can easily be included in $\mathrm{TPR}^{\mathbb{I}}$ and $\mathrm{FPR}^{\mathbb{S}}$, enabling to directly quantify how performances of the marker vary with these covariates.

5 Time-dependent AUC and Competing risks

We now consider the setting where a subject might experience multiple type of failures: in this section, we review extensions of time-dependent AUCs to competing risks. For example, we may want to assess the discrimination of a given score on death from prostate cancer with death from other causes acting as a competing event. For the sake of simplicity, we will assume there are only two competing events, and we let $\delta_i = j$ denote that subject i experienced the competing event of type j ($j = 1,2$, with $j = 1$ for the event of interest). The observed data consists of a failure time and a failure type (Z_i, δ_i) with $\delta_i = 0$ denoting a censored observation.
In their review paper [26] sketched most potential extensions and introduced event-specific sensitivity and specificity. They also highlighted that the crucial point was to determine whether patients experimenting a competing event should be treated as a control when evaluating the discrimination of the marker under study with respect to the event of interest. More precisely, two settings can be considered.

First, if marker X is potentially discriminatory for both the event of interest and the competing event, then both event specific AUCs should be considered simultaneously [28, 13]. For illustration, in the cumulative/dynamic setting, cases at time t can be stratified according to the event type, $\mathrm{Case}_1 = \{i : T_i \leq t, \delta_i = 1\}$ and $\mathrm{Case}_2 = \{i : T_i \leq t, \delta_i = 2\}$, and controls at time t are the event-free group at time t, Control=$\{i : T_i > t\}$. Following these lines Saha and Heargerty [28] proposed event specific versions of (2)

$$\mathrm{TPR}_j^{\mathbb{C}}(c,t) = \mathbb{P}(X > c | T \leq t, \delta = j), \ \mathrm{FPR}^{\mathbb{D}}(c,t) = \mathbb{P}(X > c | T > t, \delta \in \{1,2\}). \quad (5)$$

Estimation follows from [17], using the conditional cumulative incidence associated to competing event j, $\mathbb{P}(T \leq t, \delta = j | X)$, instead of the conditional survival function of $\mathbb{P}(T \leq t | X)$. In the context of renal transplantation, Foucher *et al.* [13] considered a slight modification of definitions (5), where controls can also be "event specific". In addition, an extension of the incident/dynamic AUC to the competing events setting was proposed by Saha and Heargerty [28].

The other option is to consider both event-free patients and patients with the competing event [39, 21] as controls. For instance, dynamic controls at time t can be defined as the group $\{i : T_i > t\} \cup \{i : T_i \leq i, \delta_i = 2\}$. This leads to the estimation of only one ROC curve, for the event of interest. In [39], Zheng *et al.* based their approach on initial estimates of the conditional cumulative incidence function for the event of interest. Their initial method provides consistent estimators if the proportional hazard assumption holds for each cause specific hazard. To relax this assumption a smooth estimator was also proposed. Another approach was described in [21], which follows the lines of DeLong *et al.* [9]. However, the suitability of this method to deal with censored data is not established.
We shall add that, as pointed out in [28, 39], employing a direct regression model for the conditional cumulative incidence would lead to a simpler estimation of the cumulative/dynamic AUC and a less convoluted interpretation of the marker effect. However, the extension to the setting of a longitudinal marker [8] as well as the evaluation of a risk score (which is usually built with a cause-specific hazard approach) would not be straightforward.

6 Discussion

While the AUC is uniquely defined in the context of the evaluation of diagnostic tests, its extension to prognostic markers has led to the development of a variety of definitions: these definitions vary according to the underlying definitions of cases (incident or cumulative) and controls (static or dynamic), and also depend on the study characteristics (the marker can be measured only once or repeatedly and competing events may be considered, or not). Regarding the choice of the retained definition for cases and controls, no clear guidance has really emerged in the literature. It seems however that the cumulative/dynamic definition may be more appropriate for clinical decisions making (enrollment in clinical trials for instance) while the incident/static definition may be more appropriate for "pure" evaluation of the marker (if interpretation of trends of AUC values over time is of particular interest for instance). Once this definition has been chosen, appropriate estimators are available, depending on various assumptions (independence of the marker and censoring, proportional hazards, ...), and we presented most of them in this review article.
For the sake of brevity, we were not able to cover some interesting extensions of time-dependent AUCs. In particular, covariate specific time-dependent ROC curves and AUCs have been studied in order to adjust the discrimination of a marker for external covariates (age, gender, ...). We refer the reader to [19, 33] for the standard setting, [4] for the longitudinal setting and [39] for the competing events setting. In addition, some authors advocate that not the entire ROC curve is of interest and the area under only a portion of it should be computed, leading to the so-called *partial AUC* [11]. In the context of prognostic markers, Hung and Chiang [20] proposed a nonparametric estimator of the cumulative/dynamic time-dependent version of the partial AUC. Other interesting extensions include diverse censoring patterns [22]

(only right-censoring was considered in this review) and the combination of results from multiple studies [3] which is particularly useful in genomic studies.
Another closely related topic is the evaluation of the added predictive ability of a new marker: for instance, we may wonder how better a risk score would be if we added some biological markers (SNPs, genes, ...). We refer the reader to the works in [5, 10, 23, 25] for some insights, noticing though that most of these works do not cover the right-censored setting considered in our review.

References

1. M. Akritas. Nearest neighbor estimation of a bivariate distribution under random censoring. *The Annals of Statistics*, 22:1299–1327, 1994.
2. P. Blanche, J. F. Dartigues, and H. Jacqmin-Gadda. Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Submitted*, 2012.
3. T. Cai, T.A. Gerds, Y. Zheng, and J. Chen. Robust prediction of t-year survival with data from multiple studies. *Biometrics*, 67:436–444, Jun 2011.
4. T. Cai, M.S. Pepe, Y. Zheng, T. Lumley, and N.S. Jenny. The sensitivity and specificity of markers for event times. *Biostatistics*, 7:182–197, Apr 2006.
5. L.E. Chambless, C.P. Cummiskey, and G. Cui. Several methods to assess improvement in risk prediction models: extension to survival analysis. *Statistics in medicine*, 30(1):22–38, 2011.
6. L.E. Chambless and G. Diao. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in Medicine*, 25:3474–3486, Oct 2006.
7. C.T. Chiang and H. Hung. Nonparametric estimation for time-dependent AUC. *Journal of Statistical Planning and Inference*, 140(5):1162 – 1174, 2010.
8. G. Cortese and P. K. Andersen. Competing risks and time-dependent covariates. *Biometrical Journal*, 52(1):138–158, 2010.
9. E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, 1988.
10. O.V. Demler, M.J. Pencina, and R.B. D'Agostino Sr. Equivalence of improvement in area under ROC curve and linear discriminant analysis coefficient under assumption of normality. *Statistics in Medicine*, 30(12):1410–1418, 2011.
11. L.E. Dodd and M.S. Pepe. Partial AUC estimation and regression. *Biometrics*, 59(3):614–623, 2003.
12. R. Etzioni, M. Pepe, G. Longton, C. Hu, and G. Goodman. Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. *Medical Decision Making*, 19(3):242–251, 1999.
13. Y. Foucher, M. Giral, J.P. Soulillou, and J.P. Daures. Time-dependent ROC analysis for a three-class prognostic with application to kidney transplantation. *Statistics in Medicine*, 29:3079–3087, Dec 2010.
14. M.H. Gail and R.M. Pfeiffer. On criteria for evaluating models of absolute risk. *Biostatistics*, 6(2):227–239, 2005.
15. W. Gu and M.S. Pepe. Measures to summarize and compare the predictive capacity of markers. *International Journal of Biostatistics*, 5:27–49, 2009.
16. P. J. Heagerty and M. S. Pepe. Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):533–551, 1999.
17. P.J. Heagerty, T. Lumley, and M.S. Pepe. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56:337–344, Apr 2000.
18. P.J. Heagerty and Y. Zheng. Survival model predictive accuracy and ROC curves. *Biometrics*, 61:92–105, 2005.

19. H. Hung and C.T. Chiang. Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics*, 38(1):8–26, 2010.
20. H. Hung and C.T. Chiang. Nonparametric methodology for the time-dependent partial area under the ROC curve. *Journal of Statistical Planning and Inference*, 141(12):3829 – 3838, 2011.
21. M. Lee, K. A. Cronin, M. H. Gail, and E. J. Feuer. Predicting the absolute risk of dying from colorectal cancer and from other causes using population-based cancer registry data. *Statistics in Medicine*, 31(5):489–500, 2012.
22. J. Li and S. Ma. Time-dependent ROC analysis under diverse censoring patterns. *Statistics in Medicine*, 30:1266–1277, 2011.
23. M.J. Pencina, R.B. D'Agostino Sr, R.B. D'Agostino Jr, and R.S. Vasan. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine*, 27(2):157–172, 2008.
24. M.S. Pepe. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA, 2004.
25. M.S. Pepe, K.F. Kerr, G. Longton, and Z. Wang. Testing for improvement in prediction model performance. *Preprint available at http://www.bepress.com/uwbiostat/paper379/*, 2011.
26. M.S. Pepe, Y. Zheng, Y. Jin, Y. Huang, C.R. Parikh, and W.C. Levy. Evaluating the ROC performance of markers for future events. *Lifetime Data Analysis*, 14:86–113, Mar 2008.
27. D. Rizopoulos. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67:819–829, Sep 2011.
28. P. Saha and P.J. Heagerty. Time-dependent predictive accuracy in the presence of competing risks. *Biometrics*, 66:999–1011, Dec 2010.
29. P. Saha-Chaudhuri and PJ Heagerty. Non-parametric estimation of a time-dependent predictive accuracy curve. *Biostatistics*, 2012.
30. G.A. Satten and S. Datta. The kaplan-meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, 55(3):207–210, 2001.
31. E.H. Slate and B.W. Turnbull. Statistical models for longitudinal biomarkers of disease onset. *Statistics in medicine*, 19(4):617–637, 2000.
32. X. Song, X. H. Zhou, and S. Ma. Nonparametric receiver operating characteristic-based evaluation for survival outcomes. *Statistics in Medicine*, 31(23):2660–2675, 2012.
33. X. Song and X.H. Zhou. A semi-parametric approach for the covariate specific roc curve with survival outcome. *Statistica Sinica*, 18:947–965, 2008.
34. F. Subtil, C. Pouteil-Noble, S. Toussaint, E. Villar, and M. Rabilloud. A Simple Modeling-free Method Provides Accurate Estimates of Sensitivity and Specificity of Longitudinal Disease Biomarkers. *Methods of Information in Medicine*, 48:299–305, 2009.
35. A.A. Tsiatis and M. Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834, 2004.
36. H. Uno, T. Cai, L. Tian, and L.J Wei. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537, 2007.
37. V. Viallon and A. Latouche. Discrimination measures for survival outcomes: connection between the AUC and the predictiveness curve. *Biometrical Journal*, 53:217–236, Mar 2011.
38. R. Xu and J. O'Quigley. Proportional hazards estimate of the conditional survival function. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):667–680, 2000.
39. Y. Zheng, T. Cai, Y. Jin, and Z. Feng. Evaluating Prognostic Accuracy of Biomarkers under Competing Risk. *Biometrics*, Dec 2011.
40. Y. Zheng and P.J. Heagerty. Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics*, 5(4):615–632, 2004.
41. Y. Zheng and P.J. Heagerty. Partly conditional survival models for longitudinal data. *Biometrics*, 61(2):379–391, 2005.
42. Y. Zheng and P.J. Heagerty. Prospective accuracy for longitudinal markers. *Biometrics*, 63:332–341, 2007.