UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Florent LE BORGNE

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
sous le sceau de l'Université Bretagne Loire

École doctorale : Biologie-Santé

Discipline : Recherche clinique, innovation technologique, santé publique
Spécialité : Biostatistique
Unité de recherche : EA4275

Soutenue le : 6 octobre 2016
Thèse N° :

Conception d'un outil simple d'utilisation pour réaliser des analyses statistiques ajustées valorisant les données de cohortes observationnelles de pathologies chroniques : application à la cohorte DIVAT.

JURY

Président du jury :	**Paul LANDAIS,** Professeur des Universités – Practicien Hospitalier, Université de Montpellier
Rapporteurs :	**Hélène JACQMIN-GADDA**, Directrice de recherche, Université de Bordeaux **Raphaël PORCHER,** Maître de Conférences des Universités – Practicien Hospitalier, Université Paris Descartes
Examinateur :	**Delphine MAUCORT-BOULCH**, Maître de Conférences des Universités – Practicien Hospitalier, Université Claude Bernard Lyon 1
Directeurs de Thèse :	**Yohann FOUCHER**, Maître de Conférences des Universités, Université de Nantes **Magali GIRAL**, Professeur des Universités – Practicien Hospitalier, Centre Hospitalier Universitaire de Nantes

Résumé

En recherche médicale, les cohortes permettent de mieux comprendre l'évolution d'une pathologie et d'améliorer la prise en charge des patients. La mise en évidence de liens de causalité entre certains facteurs de risque et l'évolution de l'état de santé des patients est possible grâce à des études étiologiques. L'analyse de cohortes permet aussi d'identifier des marqueurs pronostiques de l'évolution d'un état de santé. Cependant, les facteurs de confusion constituent souvent une source de biais importante dans l'interprétation des résultats des études étiologiques ou pronostiques.

Dans ce manuscrit, nous présentons deux travaux de recherche en Biostatistique dans la thématique des scores de propension. Dans le premier travail, nous comparons les performances de différents modèles permettant d'évaluer la causalité d'une exposition sur l'incidence d'un événement en présence de données censurées à droite. Dans le second travail, nous proposons un estimateur de courbes ROC dépendantes du temps standardisées et pondérées permettant d'estimer la capacité prédictive d'un marqueur en prenant en compte les facteurs de confusion potentiels. En cohérence avec l'objectif de fournir des outils statistiques adaptés, nous présentons également dans ce manuscrit une application nommée Plug-Stat®. En lien direct avec la base de données, elle permet de réaliser des analyses statistiques adaptées à la pathologie afin de faciliter la recherche épidémiologique et de mieux valoriser les données de cohortes observationnelles.

Mots clés : Scores de propension, Facteurs de confusion, Données censurées, Courbes ROC, Transplantation rénale, Cohorte.

Abstract

In medical research, cohorts help to better understand the evolution of a pathology and improve the care of patients. Causal associations between risk factors and outcomes are regularly studied through etiological studies. Cohorts analysis also allow the identification of new markers for the prediction of the patient evolution. However, confounding factors are often source of bias in the interpretation of the results of etiologic or prognostic studies.

In this manuscript, we presented two research works in Biostatistics, the common topic being propensity scores. In the first work, we compared the performances of different models allowing to evaluate the causality of an exposure on an outcome in the presence of right censored data. In the second work, we proposed an estimator of standardized and weighted time-dependent ROC curves. This estimator provides a measure of the prognostic capacities of a marker by taking into account the possible confounding factors. Consistent with our objective to provide adapted statistical tools, we also present in this manuscript an application, so-called Plug-Stat®. Directly linked with the database, it allows to perform statistical analyses adapted to the pathology in order to facilitate epidemiological studies and improve the valorization of data from observational cohorts.

Key words : Propensity score, Confounding factors, Censored data, ROC curves, Kidney transplantation, Cohort.

Remerciements

Cette thèse est le fruit d'un travail de recherche mené en collaboration avec l'équipe EA4275-SPHERE (MethodS for Patients-centered outcomes and HEalth REsearch) de l'Université de Nantes et la société IDBC/A2com spécialisée dans la conception de solutions informatiques pour le domaine médical et de la recherche. Aussi, j'aimerais remercier toutes les personnes qui, de près ou de loin, ont contribué à la réalisation de ce projet.

En premier lieu, je tiens à remercier mes deux directeurs de thèse Magali Giral et Yohann Foucher, pour m'avoir donné l'opportunité de faire cette thèse, pour votre sympathie et pour votre disponibilité permanente. Yohann, tu es le principal initiateur des questions traitées dans ce document, pour ton soutien, ton œil critique et ta rigueur scientifique, je te suis particulièrement reconnaissant. Magali, ta bonne humeur, ton enthousiasme et ta bienveillance m'ont fait passer trois formidables années. Je souhaite associer à ces remerciements Cyrille Loncle, pour ta disponibilité, ta gentillesse et ton implication. Tu as soutenu ce projet et tu continues de soutenir nos projets futurs, pour tout cela je te remercie.

Je désire grandement remercier Isabelle Delaune-Minard et Noël Minard pour avoir cru en ce projet et pris le risque de le porter financièrement. Vous m'avez permis de m'épanouir professionnellement en m'accordant votre confiance dans la gestion de ce projet. Vous continuez à soutenir notre projet "LabCom", pour cela aussi je vous suis reconnaissant.

J'exprime toute ma gratitude envers les membres de mon jury qui ont accepté de juger cette thèse. Au Pr Jacqmin-Gadda et au Dr Porcher qui m'ont fait l'honneur d'être rapporteurs de ce travail de recherche. Je remercie également le Dr Maucort-Boulch et le Pr Landais pour leur participation au jury comme examinateurs.

Mes remerciements vont également à Sahar Bayat et Bruno Giraudeau, les membres de mon comité de suivi de thèse, dont les remarques pertinentes et les suggestions ont contribué à l'amélioration de ce travail. Je n'oublie pas Christophe Combescure, ton aide sur les courbes ROC dépendantes du temps combinée à tes connaissances et ton expérience dans ce domaine ont été très précieuses.

Mes vifs remerciements vont à Pascal Daguin, quel bonheur ce fut de partager ton bureau pendant ces trois années ! Tu as soutenu ce projet et contribué à le rendre possible. Pour tes précieux conseils, ton aide et ta bonne humeur quotidienne je ne saurais suffisamment te remercier.

Un merci tout particulier à Florence pour sa relecture minutieuse et ses nombreux conseils.

Bien sûr mes remerciements vont à Véronique Sébille, directrice de l'EA4275-SPHERE, pour m'avoir accueilli au sein de son équipe et pour son soutien à notre projet de Laboratoire Commun. Je désire en outre remercier l'ensemble de l'équipe DIVAT-SPHERE ainsi que tous les membres de l'EA4275-SPHERE, du service de Néphrologie et de l'équipe d'IDBC. A tous, je vous adresse mes sentiments les plus chaleureux.

Je remercie particulièrement mon amie, Stella, pour sa patience et pour tout le réconfort qu'elle a su m'apporter.

Enfin, mes derniers mots sont réservés à mes parents. Je mesure la chance que j'ai eu de faire de longues études. J'y ai appris beaucoup. C'est grâce à vous.

Valorisations scientifiques

Publications

Publications principales issues de la thèse

- Le Borgne F., Giraudeau B., Querard A.H., Giral M., Foucher Y. Comparisons of the performance of different statistical tests for time-to-event analysis with confounding factors : practical illustrations in kidney transplantation. *Statistics in Medicine*, Oct. 2015.
- Le Borgne F., Combescure C., Gillaizeau F., Giral M., Chapal M., Giraudeau B., Foucher Y. Standardized and weighted time-dependent ROC curves to evaluate the intrinsic prognostic capacities of a marker by taking into account confounding factors. *Statistical Methods in Medical Research* (en révision)
- Le Borgne F., Loncle C., Querard A.H., Foucher Y. Plug-stat® : a user friendly web application to query, extract and analyze observational data in clinical epidemiology. (en préparation)

Autres publications sur la thématique de la thèse

- Le Borgne F., Gillaizeau F., Rousseau C., Leyrat C., Giral M., Chapal M., Barbin L., Laplaud D., Giraudeau B., Foucher Y. The traditional multivariable logistic regression compared to propensity score : a study to put an end to preconceived ideas. (en préparation)
- Gillaizeau F., Senage T., Leffondre K., Le Borgne F., Le Tourneau T., Roussel J.C., Giraudeau B., Dantan E., Foucher Y. Inverse Probability Weighting to control confounding in an illness-death model for interval-censored data. (soumis)
- Chapal M., Le Borgne F., Legendre C., Kreis H., Mourad G., Garrigue V., More- lon E., Buron F., Rostaing L., Kamar N., Kessler M., Ladrière M., Soulillou J.-P., Launay K., Daguin P., Offredo L., Giral M., and Foucher Y. A useful scoring system for the prediction and management of delayed graft function following kidney transplantation from cadaveric donors. *Kidney Int*, 86(6) :1130–1139, 2014.
- Chapal M., Neel M., Le Borgne F., Meffray E., Carceles O., Hourmant M., Giral M., Foucher Y., Moreau A., and Fakhouri F. Increased Soluble Flt-1 Correlates With Delayed Graft Function and Early Loss of Peritubular Capillaries in the Kidney Graft. *Transplantation*, 96(8) :739–744, Oct. 2013.

Communications orales

- Le Borgne F., Combescure C., Gillaizeau F., Giral M., Chapal M., Giraudeau B., Foucher Y.Courbes ROC standardisées dépendantes du temps pour évaluer les capacités pronostiques inhérentes à un marqueur en tenant compte des facteurs de confusion. 9^{10} Conférence Francophone d'Epidémiologie Clinique (EPICLIN), Strasbourg, France, 25-27 mai 2016.
- Le Borgne F., Foucher Y. Adjusted survival curves by using inverse probability of treatment weighting : the comparison of three adapted log-rank tests. 35^{th} Annual Conference of International Society for Clinical Biostatistics (ISCB), Vienne, Autriche, Août 2014.

- Le Borgne F., Foucher Y. Tests du Log-rank ajusté : étude de simulations. $46^{èmes}$ Journées de Statistique de la Société Française de Statistique (SFdS), Rennes, France, juin 2014.
- Le Borgne F., Daguin P., Giral M., Querard A.H., Foucher Y. Courbes de survie ajustées : Intérêt, application en transplantation rénale et implémentation dans le portail de gestion des cohortes Intégr@lis®. Journée "Méthodologie & Biostatistique" du GIRCI Grand Ouest, Angers, France, juin 2014.

Table des matières

Notations

ANR Agence Nationale de la Recherche. 130, 135, 137

ARC assistant de recherche clinique. 5, 7

ATE effet moyen de l'exposition dans l'ensemble de la population, ou *the average treatment effect in the entire population*. 15–18, 21, 65, 126–128, 133–135

ATT effet moyen de l'exposition dans la population des exposés, ou *the average treatment effect on the treated*. 15–17, 19, 21, 65, 126–128, 133–135

AUC aire sous la courbe, ou *area under curve*. 31, 32, 104, 105, 108, 134

CHU centre hospitalier universitaire. 5, 7, 8

CIFRE convention industrielle de formation par la recherche. 8

CKD-EPI *chronic kidney disease epidemiology collaboration*. 4

CNIL Commission Nationale de l'Informatique et des Libertés. 7

CRB centre de ressource biologique. 6

DFG débit de filtration glomérulaire. 3, 4

DIVAT Données Informatisées et VAlidées en Transplantation, www.divat.fr. 5–7, 12, 36, 68, 107, 110, 126, 129, 131, 132, 134, 135

EA4275-SPHERE MethodS for Patients-centered outcomes and HEalth REsearch. 6, 135

ECD donneur à critères élargis, ou *expanded criteria donor*. 65

EDTA éthylène diamine tétra-acétique. 4

ERA épisode de rejet aigu. 6

ERC essai randomisé contrôlé. 2, 9, 14, 16, 18, 20, 23, 132

HAS Haute Autorité de Santé. 4

HR rapport de risques, ou *hazard ratio*. 14, 17, 28, 29, 64, 65, 131, 135

HTA hypertension artérielle. 6

IC95% intervalle de confiance à 95%. 136

IPCW *inverse probability of censoring weighting*. 30

IPW *inverse probability weighting*. 10, 15–18, 20, 22, 26, 34, 36, 65, 68, 104, 105, 107, 127, 128, 132–134

IRC insuffisance rénale chronique. 4

IRCT insuffisance rénale chronique terminale. 4, 5, 106, 133

ITUN Institut de Transplantation Urologie Néphrologie. 5–8

KDRI *kidney donor risk index*. 106, 107, 133, 134

LabCom Laboratoire Commun. 130, 131, 135–137

MDRD *modification of diet in renal disease*. 4

MRC maladie rénale chronique. 3, 4

OMS Organisation Mondiale de la Santé. 3

OR rapport de cotes, ou *odds ratio*. 14, 16, 17, 19, 64, 127, 128, 134

PV valeur de placement, ou *placement value*. 32, 33

REIN Réseau Epidémiologie et Information en Néphrologie. 4

RISCA Recherche en Informatique et en Statistique pour l'Analyse de Cohortes. 131

ROC *Receiver Operating Characteristic*. 12, 29, 31–34, 104, 105, 107, 108, 132, 134

SI système d'information. 7

SVD dégénérescence valvulaire, ou *structural valve deterioration*. 136

sw poids stabilisés, ou *stabilized weights*. 18

TER travail d'étude et de recherche. 130

Chapitre 1

Introduction

Sommaire

1.1 Etudes observationnelles

Les études épidémiologiques sont utilisées en recherche médicale pour identifier les causes d'une pathologie et établir des liens de causalité entre des expositions et des critères de jugement. Les deux principaux composants de toute étude épidémiologique sont : l'exposition et le critère de jugement. L'exposition peut être un traitement, un facteur de risque ou un test diagnostic, par exemple. Le critère de jugement dépend de l'objectif de l'étude, il correspond généralement à un état de santé tel que la présence ou l'absence d'un état pathologique ou le décès (Jepsen et al., 2004). Les études épidémiologiques peuvent se diviser en deux grandes catégories : les études expérimentales et les études observationnelles. Dans les études expérimentales, l'investigateur contrôle l'exposition qu'il étudie : de façon aléatoire ou non il choisit les sujets qui vont recevoir l'exposition et ceux qui ne la recevront pas. Au contraire, dans les études observationnelles l'investigateur peut uniquement observer l'effet de l'exposition sur les sujets de l'étude, il ne joue aucun rôle dans l'allocation de l'exposition. Les études observationnelles sont donc plus "vulnérables" aux problèmes méthodologiques que les essais randomisés contrôlés (ERC) généralement considérés comme apportant le plus fort pouvoir de preuve. Cependant, les ERC présentent aussi des inconvénients : leur manque de validité externe a notamment été largement débattu dans la littérature (Van Spall et al., 2007; Rothwell, 2005; Black, 1996; Zwarenstein et Treweek, 2009). En cause, les conditions optimales dans lesquelles ils sont réalisés, peu représentatives du monde réel : des patients plus adhérents aux traitements, des suivis renforcés, l'exclusion des patients les plus à risque (enfants, personnes âgés, femmes enceintes, etc.). De plus, les ERC sont très coûteux tant sur le plan humain que financier et la randomisation est parfois éthiquement impossible (Sanson-Fisher et al., 2007). Les études observationnelles évitent certains de ces écueils car elles peuvent être considérées comme des expériences naturelles dans lesquelles l'exposition et les critères de jugement sont mesurés dans le monde réel (Rochon et al., 2005). Cependant, elles sont soumises à un manque de validité interne avec la présence de biais de sélection et de confusion. En effet, l'attribution non aléatoire de l'exposition ne garantit plus la comparabilité des groupes qui doit alors être prise en considération lors de l'analyse statistique.

Trois types d'études observationnelles peuvent être distinguées : les études de cohorte, les études cas-témoins et les études transversales. Dans cette thèse, nous nous sommes intéressés aux études de cohortes. Elles sont considérées comme les études observationnelles apportant le meilleur niveau de preuve scientifique pour juger du rôle causal d'un facteur de risque sur l'apparition d'un événement. Plus globalement, le terme cohorte désigne un groupe de sujets suivis longitudinalement. Les cohortes peuvent être prospectives ou rétrospectives. Dans le premier cas, les sujets sont inclus dans la cohorte puis suivis afin de recueillir la survenue éventuelle d'un ou plusieurs événements d'intérêt ainsi qu'un maximum d'informations concernant le sujet. Dans le second cas, les patients sont inclus après avoir subi l'exposition et parfois une partie des événements d'intérêt, leur historique doit alors être reconstitué. Qu'elles soient prospectives ou rétrospectives, on distingue deux types de cohortes (Goldberg et Zins, 2013) :

- Les cohortes en population générale qui s'intéressent aux causes des maladies, aux déterminants environnementaux, génétiques, médico-économiques, sociétaux, etc. Ces cohortes peuvent inclure jusqu'à plusieurs centaines de milliers de sujets indemnes de la pathologie d'intérêt et les suivre pendant des décennies.

- Les cohortes de malades d'une taille plus restreinte, incluant de quelques milliers à quelques dizaines de milliers de sujets recrutés en milieu médical et partageant une pathologie particulière (insuffisants cardiaques, transplantés rénaux, etc.).

1.2 Pathologies chroniques

1.2.1 Présentation générale

L'Organisation Mondiale de la Santé (OMS) définit les maladies chroniques comme des affections de longue durée qui évoluent lentement avec le temps et nécessitent une prise en charge sur une longue période (de plusieurs années à plusieurs décennies). Ces pathologies affectant fortement la vie professionnelle, sociale et familiale des malades représentent environ 60% des décès dans le monde (World Health Organization, 2014). En France, on estime à 15 millions le nombre de personnes atteintes d'une maladie chronique soit environ 20% de la population (Ministère des Affaires sociales et de la Santé, 2007) dont plus de 9 millions sont inscrites en affection de longue durée. Ces maladies comprennent des maladies non transmissibles comme le diabète, le cancer, la maladie d'Alzheimer, l'asthme ou l'insuffisance rénale chronique ; des maladies transmissibles persistantes, comme le sida ou l'hépatite C ; des troubles mentaux de longue durée comme la dépression ou la schizophrénie ; des maladies rares ou inflammatoires comme la mucoviscidose ou la polyarthrite rhumatoïde ; des maladies neurologiques ou dégénératives comme la myopathie ou la maladie de Parkinson, etc.
Les pathologies chroniques représentent un défi majeur de santé publique, le paradoxe étant que les progrès médicaux et l'allongement de l'espérance de vie, contribuent à leur augmentation. En 2005, l'OMS a fait de l'amélioration de la qualité de vie des personnes atteintes de maladies chroniques une priorité (World Health Organization, 2005). Cette priorité a été reprise en France dans un plan national du Ministère des Affaires sociales et de la Santé (2007). Mieux connaitre les déterminants de l'apparition et de l'évolution des maladies chroniques est donc primordial que ce soit pour en limiter l'incidence, améliorer la prise en charge et la qualité de vie des patients mais aussi prévenir leurs complications. Dans cette thèse, nous nous intéressons plus particulièrement à l'insuffisance rénale chronique terminale.

1.2.2 Insuffisance rénale chronique terminale

La maladie rénale chronique (MRC) est un syndrome défini indépendamment de sa cause, par la présence pendant plus de 3 mois d'une baisse du débit de filtration glomérulaire (DFG) comportant également des anomalies hydroélectriques et endocriniennes. Elle est due à la réduction permanente et définitive du nombre de néphrons fonctionnels (ce qui la différencie de l'insuffisance rénale aiguë ou fonctionnelle). L'importance de la MRC est mesurée par le DFG : il représente la capacité des reins à filtrer une certaine quantité de sang par unité de temps. On parle d'insuffisance rénale chronique (IRC) dès lors que le patient présente un DFG inférieur à $60mL/min/1.73m^2$ et d'insuffisance rénale chronique terminale (IRCT) lorsque le DFG est inférieur à $15mL/min/1.73m^2$ (Agence Nationale d'Accréditation et d'Évaluation en Santé, 2002). Les méthodes de mesure du DFG étant coûteuses et complexes (inulin, iohexol et éthylène diamine tétra-acétique (EDTA), par exemple), ce dernier est en pratique estimé à partir de la concentration plasmatique d'une molécule dont la sécrétion dans le sang est connue et stable telle que la créatinine. Plusieurs équations ont été proposées pour estimer le DFG, les dernières en date étant la formule de la MDRD (*modification of diet in renal disease*) recommandée par le Groupe de travail de la Société de Néphrologie (2009) et la formule CKD-EPI (*chronic kidney disease epidemiology collaboration*) (Levey et al., 2009) dont l'usage est recommandé par la HAS (Haute Autorité de Santé) depuis 2011.
L'IRCT représente le dernier stade de la MRC. Les reins étant des organes vitaux, seules des techniques de suppléance (dialyse chronique ou transplantation rénale) peuvent assurer la survie des patients qui en sont atteints. En France, le Réseau Epidémiologie et Information en Néphrologie (REIN) estimait en 2013, à 160 par million d'habitants le taux d'incidence moyen de l'IRCT (Lassalle et al., 2013). En

termes de prévalence, 76 187 malades recevaient un traitement de suppléance au 31 décembre 2013 dont 56% étaient traités par dialyse et 44% étaient porteurs d'un greffon rénal fonctionnel (Briançon et al., 2013).

La dialyse est une technique d'épuration extra-rénale qui permet d'éliminer les toxines urémiques et de rétablir les équilibres ioniques qui ne sont plus assurés par les reins. Deux techniques de dialyses existent, elles sont au choix du patient et sur critères médicaux (en l'absence de contre-indication à l'une ou l'autre des méthodes).

- L'hémodialyse (ou dialyse extra-corporelle) qui repose sur des échanges entre un bain de dialyse et le sang des patients au travers d'une membrane artificielle (rein artificiel).
- La dialyse péritonéale (ou dialyse intra-corporelle) qui repose sur des échanges entre un bain de dialyse et le sang des patients au travers de la membrane péritonéale du malade.

La transplantation est le traitement de choix de l'IRCT en termes de contraintes, de réinsertion socio-professionnelle, de qualité de vie et de survie des patients mais aussi de coût pour la société (Wong et al., 2012; Winkelmayer et al., 2002; Tonelli et al., 2011; Cameron et al., 2000; Wolfe et al., 1999). D'après Blotière et al. (2010) le coût global de la prise en charge des 60 900 patients atteints IRCT en 2007 était de 4 milliards d'euros dont 77% pour l'hémodialyse. Le coût annuel moyen pour un patient en dialyse était de 89 000 euros pour l'hémodialyse et 64 000 euros pour la dialyse péritonéale, pour un patient greffé le coût est proche lors de la première année (86 000 euros) puis bien inférieur les années suivantes (20 000 euros). La transplantation nécessite cependant que le patient prenne un traitement, dit immunosuppresseur, afin de diminuer l'activité du système immunitaire et ainsi prévenir le rejet du greffon.

L'accès à la transplantation est difficile, due à une pénurie de greffon. En 2013, 10 451 nouveaux malades ont débuté un traitement de suppléance (dialyse ou greffe préemptive) pour IRCT (Lassalle et al., 2013). Le nombre total de patients en attente d'une greffe (nouveaux inscrits et malades restant en attente) a atteint 14 336, représentant une augmentation de 7% en 1 an et de 40% en 5 ans (Hourmant et al., 2013). Cette même année, 3 074 greffes rénales ont été réalisées, dont 12% étaient des greffes préemptives (chez des non dialysés) et 16% étaient des retransplantations. Même si l'écart entre le nombre de candidats à une transplantation et le nombre de greffons disponibles doit être relativisé par les 4 226 malades inscrits sur liste d'attente mais en contre-indication temporaire, il apparait clairement que le nombre de greffons disponible est insuffisant pour couvrir les besoins des patients souffrant d'IRCT.

Dans ce contexte, il est primordial d'optimiser la survie du greffon afin de minimiser le nombre de patients à réinscrire en liste d'attente après la perte d'un premier greffon et d'améliorer la qualité de vie des patients et leur survie. En 2013, les retours en dialyse représentaient 9% des malades débutant un traitement de suppléance (Hourmant et al., 2013).

1.2.3 La cohorte DIVAT

Dans ce travail de thèse nous avons réalisé nos développements dans le cadre du réseau DIVAT (Données Informatisées et VAlidées en Transplantation, www.divat.fr). Ce réseau, coordonné par l'Institut de Transplantation Urologie Néphrologie (ITUN) du centre hospitalier universitaire (CHU) de Nantes, regroupe sept centres transplanteurs français des CHU de : Nantes, Nancy, Paris-Necker, Lyon, Montpellier, Paris-Saint-Louis et Nice. Les objectifs du réseau DIVAT sont multiples et formalisés par un contrat de consortium entre les hôpitaux. Son objectif principal est de réaliser des études épidémiologiques permettant d'améliorer la connaissance du suivi des patients transplantés rénaux pour améliorer leur prise en charge.

La cohorte DIVAT est une cohorte prospective de patients transplantés du rein et/ou du pancréas. La saisie des données est réalisée selon un thésaurus commun par des assistants de recherche clinique (ARC) spécifiquement formés et dédiés à chaque centre. La saisie se fait à partir de formulaires mis au point en accord avec les centres et annuellement révisés pour suivre l'évolution du thésaurus. Les données pré-greffe sont saisies au moment de la greffe puis les données de suivi sont collectées prospectivement après chaque visite de suivi du patient transplanté. Les visites de suivi ont lieu à 3 mois et 6 mois pendant la première année de greffe, puis elles sont espacées à une visite par an (à chaque date anniversaire de greffe). Toutes les complications qui surviennent entre deux visites sont enregistrées dans la base. La qualité des données du réseau DIVAT est validée par un audit annuel des données informatiques par rapport aux données sources (dossiers papiers) réalisé par un ARC d'un autre centre que le centre audité. L'audit est complété par des contrôles automatiques de cohérence portant sur les données à remplissage obligatoire.
La cohorte DIVAT contient aujourd'hui les données de plus de 15 000 patients. Selon les centres, les données sont exploitables depuis 1996 pour Nantes, Paris-Necker, Nancy et Montpellier, depuis 2006 pour Lyon et depuis 2013 pour Paris-Saint-Louis et Nice. Si le nombre total de patients informatisés dans la base de données est important, le nombre de patients inclus dans les études est beaucoup plus faible, généralement inférieur à 5 000 patients, selon les critères d'inclusion.
Avec plus de 300 variables, le champ des données recueillies dans DIVAT est large, il comporte :

- Les données du donneur : âge, sexe, statut vital, cause du décès, antécédent d'hypertension artérielle (HTA), pression artérielle, sérologies virales et données biologiques (urée, protéinurie et créatinémie).
- Les données du receveur : âge, sexe, sérologies virales, poids, taille, antécédents médicaux (cancer, cardiovasculaire, etc.), maladie initiale, technique de dialyse, greffes précédentes, date de première dialyse, etc.
- Les données de la transplantation : date, temps d'ischémie froide, nombre d'incompatibilités HLA-A.B.DR, cross-match lymphocytaire, immunisation anti-HLA de classe I et II, etc.
- Les données de suivi : poids, taille, pression artérielle, créatinémie, protéinurie des 24 heures, cholestérolémie, taux d'hémoglobine, traitements immunosuppresseurs et antihypertenseurs.
- D'autres paramètres post-greffe : épisode de rejet aigu (ERA), infections, complications, retours en dialyse, décès, etc.

Si le suivi médical n'est pas interrompu par le patient, la collecte des informations dans la base est poursuivie jusqu'au moment où le patient retourne en dialyse suite à la perte définitive de son greffon ou jusqu'à son décès s'il décède avec un greffon fonctionnel. Ces données sont également enrichies par une biocollection qui a été récemment couplée avec la base DIVAT (DIVAT-Biocoll). Elle rassemble les urines, le sérum, le plasma et les cellules des donneurs et des receveurs au sein des centres de ressources biologiques (CRB) dans les CHU de Lyon, Paris-Necker et Nantes (réseau CENTAURE). Cette biocollection a pour objectifs d'identifier et de valider de nouveaux biomarqueurs pour prédire des événements post-greffe importants tels que les épisodes de rejet aigu ou la perte du greffon. Un consentement spécifique est demandé aux patients pour les données propres à la biocollection.
Autour des données de la cohorte DIVAT, un groupe de travail pluridisciplinaire en Épidémiologie et Biostatistique en transplantation rénale s'est développé, intitulé DIVAT-SPHERE, il a trois objectifs : i) la gestion de la qualité des données, ii) la garantie de l'analyse statistique des données issues du réseau et iii) proposer des outils pour une aide à la prise de décision médicale basés sur des développements méthodologiques novateurs. Ce groupe est constitué de chercheurs des équipes MethodS for Patients-centered outcomes and HEalth REsearch (EA4275-SPHERE) et ITUN.

Pour la saisie, le stockage et l'exploitation de ses données, le réseau DIVAT, en partenariat avec la société IDBC/A2com et l'ITUN, a mis au point deux applications web accessibles via une connexion internet sécurisée :

- Intégr@lis® qui permet aux ARC de saisir les données des patients grâce à des formulaires. Lors de la saisie, les données sont instantanément intégrées dans une base de données informatique appelée base de production. Les données de chaque centre sont séparées dans des bases de production différentes. Ces bases peuvent être hébergées soit au sein du CHU propriétaire des données soit au sein d'IDBC/A2com qui est agréé hébergeur de données de santé à caractère personnel par le Ministère des Affaires sociales et de la Santé. Ce choix est fait par les systèmes d'information (SI) de chaque CHU. Intégr@lis® peut aussi être utilisée comme dossier de spécialité et de soins, c'est le cas dans le CHU de Nantes où l'ensemble du personnel du service de néphrologie (médecins, infirmiers, ARC, etc.) utilise Intégr@lis® pour accéder aux dossiers des patients, gérer les rendez-vous, etc. Dans ce cas, des informations complémentaires à celles saisies par les ARC sont intégrées aux dossiers (courriers, observation médicales, rendez-vous, etc.).
- Plug-Stat® qui permet de requêter, d'analyser et d'extraire les données de la base de données consolidée. La base de données consolidée regroupe les données des bases de production de chaque centre en une seule base de données multicentrique anonymisée, hébergée au sein d'IDBC/A2com. La consolidation des données dans cette base unique permet : i) d'anonymiser les données (un numéro informatique unique par patient est créé, permettant si besoin les échanges d'informations entre les bases de production et la base consolidée en conservant l'anonymat des patients), ii) de mutualiser les données des différents centres, iii) de ne conserver que les données saisies par les ARC et celles transférées automatiquement des biocollections pour les centres concernés (les éventuelles données du dossier de soins (courriers, rendez-vous, etc.) ne sont pas incluses dans la base consolidée), iv) de procéder au recodage de certaines variables et v) d'effectuer des contrôles de cohérence automatiques et d'en envoyer un rapport aux ARC concernés en cas de données manquantes, suspectes ou incohérentes.

Ces bases de données font l'objet d'une déclaration à la (CNIL, accord n° 891735).

1.3 IDBC/A2com

Il y a plus de dix ans, le Pr. Magali Giral et Pascal Daguin de l'ITUN du CHU de Nantes, ont fait le projet de créer une entreprise de service informatique dévolue à la création de bases de données médicales. En partenariat avec l'institut de commerce de Nantes et grâce au soutien d'Oseo-ANVAR est née la société IDBC qui a reçu la même année le prix Microsoft pour la création d'entreprises innovantes. Ce savoir-faire s'est développé autour de la cohorte DIVAT. Cette structure a permis la création d'outils de gestion de cohortes quel que soit le champ d'application (biologique, médical, chirurgical, génétique). Pascal Daguin, en collaboration avec l'ITUN, a ensuite développé dans le cadre de la société IDBC l'application Intégr@lis®. Cette application est couramment utilisée comme dossier de spécialité et de soins dans plusieurs unités médicales du CHU de Nantes, mais aussi de Nancy ou de Marseille. Elle est également utilisée comme simple outil de cohorte dans plusieurs autres unités médicales de différents CHU. Un module complémentaire permettant de requêter et extraire les données a ensuite été développé par Pascal Daguin, toujours en collaboration avec l'ITUN. En 2006, la société IDBC a été rachetée par A2com dont le siège social se situe à Pacé (35740, France). Ce rapprochement a permis une amélioration dans la gestion des réseaux et la sécurité informatique. La société IDBC/A2com fait partie du pôle de compétitivité Atlantique Biothérapie depuis 2008 et compte actuellement 50

employés. Quelques années plus tard, IDBC/A2com a obtenu le marché pour la création et la gestion de cohortes du CHU de Nantes. Suite au gain de ce marché, IDBC/A2com a pris la décision de centrer son activité de recherche et développement sur la refonte de son outil d'extraction des données afin de le moderniser. Naît alors un nouvel outil basé sur des technologies actuelles (JavaScript, VB.net), le rendant plus intuitif, ergonomique et performant que l'ancienne version. Cependant, cet outil restait insuffisant pour complètement répondre aux attentes des cliniciens et des chercheurs qui souhaitent valoriser leurs données par des travaux scientifiques d'audience internationale. Récemment, IDBC/A2com a décidé de poursuivre le développement de ce dernier afin d'en faire un outil innovant permettant de réaliser des analyses statistiques descriptives mais également des analyses statistiques ajustées prenant en compte les facteurs de confusion toujours en lien direct avec la base de données (i.e. sans étape d'extraction des données). C'est l'objet de cette thèse réalisée dans le cadre d'une convention industrielle de formation par la recherche (CIFRE) et formalisant un partenariat historique entre IDBC/A2com, l'Université de Nantes et le CHU de Nantes. Enfin notons que la société IDBC/A2com a récemment été agréé hébergeur de données de santé à caractère personnel par le Ministère des Affaires sociales et de la Santé. Ainsi, IDBC/A2com est une société investie depuis sa création dans la recherche et le développement d'outils innovants au service des malades, des équipes soignantes et des équipes de recherche associées.

1.4 Valorisation des données de cohortes

1.4.1 Enjeux et problématique

La constitution et le maintien d'une cohorte a un coût élevé et demande un fort investissement de la part de l'investigateur, qui doit trouver les financements nécessaires et la dynamique pour maintenir la qualité et l'exhaustivité des données saisies. En effet, les frais engendrés par une cohorte sont nombreux : stockage des données, logiciel de recueil et mise à jour du logiciel, personnel, exploitation des données (data management et biostatistique), juridique. Etant donné cet investissement, la valorisation des données devient cruciale. Or il existe un réel risque que le temps passé par les chercheurs à la gestion et à la logistique du projet empiète trop largement sur celui de l'investigation scientifique et nuise à la production scientifique (Goehrs et al., 2012).

L'objectif principal d'une cohorte de malade est d'étudier l'évolution d'une maladie, dans le but notamment d'améliorer la prise en charge et/ou le traitement des patients. Les cohortes de malades permettent aussi de mieux connaitre les déterminants de santé et leurs effets. La valorisation d'une cohorte passe donc par la réalisation d'études épidémiologiques afin de mettre en évidence des liens de causalité entre certains facteurs de risque (l'exposition) et l'évolution de l'état de santé des patients. De telles études peuvent également être menées afin d'identifier des marqueurs permettant de mieux pronostiquer l'évolution de l'état de santé. La problématique majeure rencontrée dans les études de cohortes est la non comparabilité des groupes d'études (exposés/non-exposés). Au contraire des essais randomisés contrôlés (ERC), où la comparabilité des groupes d'études est assurée par l'attribution aléatoire de l'exposition (à condition que le nombre de sujets inclus soit suffisamment important), les études observationnelles sont souvent associées à des sujets exposés différents des sujets non-exposés. Si ces facteurs associés à l'exposition sont également associés à l'événement, on parle alors de facteurs de confusion. Pour être exact une troisième condition complète la définition d'un facteur de confusion, précisant que le tiers facteur ne doit pas être une conséquence de l'exposition (c'est à dire sur le chemin causal) (Bouyer, 2009; Jager et al., 2008). Ces facteurs de confusion engendrent des biais de confusion lors de l'estimation de la causalité de l'effet de l'exposition sur l'événement et doivent donc être pris en compte lors de l'analyse statistique.

1.4.2 Prise en compte des facteurs de confusion

Quelle que soit la nature du critère de jugement d'intérêt (variable quantitative, variable qualitative, variable censurée : délai de survenue d'un événement), plusieurs méthodes permettent de prendre en compte les facteurs de confusion au moment de l'analyse (Bouyer, 2009; Normand et al., 2005). Parmi ces méthodes, les plus couramment utilisées en épidémiologie sont les suivantes :

- La stratification consiste à diviser l'échantillon d'étude en sous-groupes (aussi appelés strates), chaque strate représente un profil défini par les combinaisons possibles des niveaux des facteurs de confusion. L'objectif étant de créer des sous-groupes dans lesquels les facteurs de confusion sont identiques entre les sujets exposés et les sujets non-exposés.
- L'appariement individuel propose de choisir pour chaque sujet exposé un ou plusieurs sujet(s) non-exposé(s) ayant des caractéristiques identiques (ou très proches) pour les facteurs confondants. L'objectif est d'obtenir deux groupes comparables. L'appariement individuel peut être vu comme un cas particulier de la stratification avec autant de strates que de paires formées.
- La modélisation multivariée utilise l'ensemble des observations pour estimer les associations entre les facteurs de confusion et l'événement et fournir une estimation ajustée de l'effet causal individuel de l'exposition.
- La pondération IPW (*inverse probability weighting*) a été introduite par Rosenbaum et Rubin (1983). Le score de propension résume l'information contenue dans les différents facteurs de confusion à l'intérieur d'une seule variable : le score de propension. L'idée de la méthode IPW est de pondérer la contribution de chaque sujet par l'inverse de la probabilité qu'il avait de recevoir l'exposition qu'il a vraiment reçu. L'objectif étant d'obtenir un pseudo-échantillon (ou échantillon pondéré) dans lequel la distribution des covariables est similaire entre exposés et non-exposés.
- L'appariement sur le score de propension consiste à apparier chaque sujet exposé à un ou plusieurs sujets non-exposés ayant une valeur proche du score de propension.

Les deux premières méthodes ont pour principaux inconvénients de limiter l'analyse à un nombre restreint de facteurs de confusion et d'être associées à une moins bonne puissance statistique (McNamee, 2005; Bouyer, 2009). De plus, la stratification ne peut fonctionner que si elle est suffisamment fine pour éliminer l'association entre l'événement et les facteurs de confusion à l'intérieur des strates. Or, plus le nombre de strates augmente, plus les effectifs dans chaque strate diminuent rendant l'estimation de l'effet au sein des strates instable voire impossible (Normand et al., 2005). La méthode d'appariement est elle associée à une perte de puissance statistique du fait de l'exclusion de sujets. En effet, les sujets exposés ne trouvant pas de paire sont exclus ainsi qu'un certains nombre de sujets non-exposés non utilisés (par exemple si 40% des sujets sont exposés, dans le cas d'un appariement 1 :1 un maximum de 80% de l'échantillon total sera utilisé). Cette exclusion de sujets peut aussi être une source de biais de sélection (les sujets inclus ne sont pas représentatifs de l'échantillon initial). La modélisation multivariée est la méthode la plus couramment utilisée. Elle permet de prendre en compte simultanément un nombre important de facteurs de confusion et ne nécessite pas de catégoriser les facteurs quantitatifs. Elle présente cependant l'inconvénient d'être basée sur de nombreuses hypothèses, notamment sur la forme des relations entre les facteurs de confusion et l'événement. De plus, son interprétation par les professionnels de santé est difficile. La méthode IPW est de plus en plus fréquemment utilisée dans la littérature médicale. C'est une alternative intéressante aux trois méthodes précédentes car, comme la modélisation multivariée, elle permet de prendre en compte un nombre important de facteurs de confusion. De plus, l'échantillon pondéré est plus facile à expliquer à des non-spécialistes de la Biostatistique de part son mimétisme avec l'essai randomisé. Notons cependant qu'il existe des différences dans la nature des effets estimés selon la méthode utilisée. Cette notion sera développée dans les chapitres suivants.

1.5 Objectifs de la thèse

1.5.1 Motivations

Historiquement, la production scientifique issue de DIVAT est toujours venue de collaborations étroites entre cliniciens d'une part et biostatisticiens d'autre part. De manière générale, l'exploitation et l'analyse des données de cohortes observationnelles nécessitent des logiciels appropriés, des connaissances en data management ainsi qu'en Biostatistique. Or, peu de cohortes disposent de leur propre équipe de Biostatistique capable d'accompagner la réalisation des études épidémiologiques. Couramment, les cliniciens impliqués dans la cohorte doivent faire appel à des groupes de travail extérieurs pour mener à bien leurs études épidémiologiques. De plus, nous constatons que pour chaque étude, un certain nombre d'étapes sont répétées. Ainsi, le schéma standard après définition de l'étude est le suivant : i) le data-manager procède à l'extraction des données nécessaires à l'étude, ii) il réalise les différentes étapes de data-management permettant d'obtenir des données exploitables (recodage, contrôle de cohérence, etc.), iii) il transmet les données au biostatisticien, iv) ce dernier importe les données dans un logiciel de statistique, v) il réalise les analyses statistiques adaptées, et vi) il rédige son rapport contenant les méthodes et résultats. Nous pensons que ces étapes longues, coûteuses et sources d'erreurs expliquent en partie l'actuelle sous-valorisation des bases de données de cohortes observationnelles.
Dans ce contexte, il semble primordial de simplifier ces étapes afin de faciliter l'exploitation et la valorisation des données de cohortes observationnelles. Pour cela, il est crucial que les biostatisticiens, épidémiologistes et cliniciens-chercheurs ayant une formation en Biostatistique aient un accès autonome aux données. Il est également nécessaire qu'ils puissent à partir d'un seul outil, requêter les données, réaliser des analyses statistiques ajustées ainsi qu'extraire, si besoin, les données dans un format adapté aux analyses.

1.5.2 Objectifs

L'objectif de ce travail de thèse est d'améliorer l'outil d'extraction et d'analyses statistiques désormais nommé Plug-Stat®, proposé par la société IDBC/A2com afin qu'il réponde aux attentes des chercheurs qui souhaitent valoriser leurs données de cohortes par des travaux scientifiques robustes et méthodologiquement fiables pour d'éventuelles applications en clinique.
Plus précisément, l'objectif principal est de proposer des outils biostatistiques en lien direct avec la base de données *via* des interfaces ergonomiques et intuitives. Les modèles statistiques proposés doivent être adaptés à la pathologie (pour permettre de modéliser les principaux critères de jugement identifiés) ainsi qu'aux données disponibles dans la cohorte. L'ensemble des solutions proposées par IDBC/A2com étant dévouées à la gestion de cohortes observationnelles, il est primordial d'intégrer des méthodes spécifiques à l'analyse de données observationnelles afin notamment de minimiser les biais de confusion.
En cohérence avec l'objectif principal qui est de fournir des outils statistiques adaptés, ce travail de thèse a deux objectifs secondaires. Le premier est de comparer les performances en termes d'erreurs de première et de seconde espèce de différents modèles permettant d'évaluer la causalité d'une exposition sur l'incidence d'un événement en présence de données censurées à droite. Ce travail nous guidera dans le choix des modèles à implémenter dans Plug-Stat®. Le second est de proposer un nouvel estimateur des courbes ROC (*Receiver Operating Characteristic*) dépendantes du temps permettant de prendre en compte les potentiels facteurs de confusion afin d'estimer les capacités pronostiques intrinsèques au marqueur. L'implémentation des méthodes pour évaluer les capacités pronostiques d'un marqueur dans l'application Plug-Stat® ne fait pas partie des objectifs de cette thèse mais est une perspective envisagée.
Bien que l'ensemble des développements effectués pendant cette thèse aient été appliqués à la cohorte

DIVAT, ce travail est avant tout un travail conceptuel adaptable à toutes les bases de données de cohortes observationnelles de pathologies chroniques.

1.5.3 Plan du mémoire

Le manuscrit est composé de quatre parties. Dans le chapitre suivant, nous essayons de résumer les connaissances déjà existantes sur les concepts statistiques qui seront repris dans les développements ultérieurs. Les trois chapitres suivants sont présentés sous la forme d'articles suivis de compléments. Le chapitre 3 tente de comparer les performances en termes d'erreurs de première et de seconde espèce de différents modèles permettant d'évaluer la causalité d'une exposition sur un événement en présence de données censurées à droite. Dans le chapitre 4, nous proposons un estimateur de courbes ROC dépendantes du temps standardisées et pondérées. Cet estimateur fournit une estimation des capacités pronostiques intrinsèques à un marqueur en tenant compte des facteurs potentiellement confondants. Le chapitre 5 présente Plug-Stat®. Le dernier chapitre est consacré à une discussion qui résume les travaux et ouvre vers nos nouvelles perspectives de travail.

Chapitre 2

Etat de l'art

Sommaire

2.1 Score de propension

2.1.1 Contexte et définition

Dans les études observationnelles, l'estimation de l'effet causal d'une exposition sur un événement peut-être biaisée par un ou plusieurs facteur(s) de confusion. La définition couramment trouvée dans la littérature d'un facteur de confusion est la suivante : tiers facteur qui est à la fois associé à l'exposition étudiée et l'événement d'intérêt sans être sur le chemin causal (c'est à dire sans être une conséquence de l'exposition). Il est possible de minimiser les biais de confusion de deux manières (Gayat et al., 2012) : i) en se focalisant sur la relation entre les potentiels facteurs de confusion et le critère de jugement, et ii) en se focalisant sur la relation entre les potentiels facteurs de confusion et l'exposition. La première consiste à modéliser le critère de jugement conditionnellement aux potentiels facteurs de confusion. Il s'agit de l'approche la plus fréquemment utilisée en épidémiologie clinique. La deuxième consiste à reconstruire *a posteriori* une situation similaire à celle obtenue par randomisation. Cette deuxième approche est par exemple, obtenue avec certaines méthodes basées sur le score de propension. Ces dernières sont de plus en plus utilisées. Dans une revue systématique, Gayat et al. (2010) montrent que le nombre d'articles basés sur le score de propension est en constante augmentation : passant de moins de 10 en 1998 à plus de 350 en 2009. L'attrait grandissant pour ces méthodes est principalement dû au mimétisme de ces méthodes avec le design des ERC qui leur confèrent un aspect plus intuitif. Ces deux stratégies peuvent également être combinées (Gayat et al., 2012). Selon l'approche utilisée, l'effet estimé peut être différent. En effet, les méthodes de régression multiple classiques estiment des effets conditionnels qui peuvent être différents des effets marginaux estimés par certaines méthodes basées sur le score de propension. L'effet conditionnel est l'effet moyen de l'exposition au niveau individuel, c'est à dire l'effet moyen associé au passage d'un individu du statut d'exposé à celui de non-exposé ou *vice versa*. Au contraire, l'effet marginal est l'effet moyen de l'exposition au niveau populationnel, c'est à dire l'effet moyen associé au passage d'une population entière, du statut d'exposée à non-exposée ou *vice versa*, dans la situation hypothétique où tous les sujets pourraient être à la fois exposés et non-exposés (Austin, 2011a). Lorsque le lien entre l'effet de l'exposition et l'espérance de l'événement conditionnellement aux covariables est linéaire (des différences de moyennes, par exemple), les effets marginaux et conditionnels coïncident. Ils coïncident également lorsque l'effet est nul. Au contraire, lorsque l'événement est binaire ou censuré et que l'effet est exprimé sous forme d'indicateurs non-linéaires (rapport de cotes, ou *odds ratio* (OR) ou rapport de risques, ou *hazard ratio* (HR)) les effets marginaux et conditionnels diffèrent (Austin et al., 2007b).

2.1.2 Différentes méthodes

Le score de propension p_i est défini comme la probabilité (ou propension) qu'un individu i soit exposé sachant les facteurs de confusion potentiels notés Z_i (Rosenbaum et Rubin, 1983) : $p_i = P(X_i = 1|Z_i = z_i)$, où X est le facteur d'exposition (avec $X = 0$ pour non-exposé et $X = 1$ pour exposé). Ce score, qui est généralement estimé par un modèle de régression logistique, est une mesure de la vraisemblance qu'a un sujet de présenter l'exposition d'intérêt en fonction de ses caractéristiques observées.
Dans leurs travaux, Rosenbaum et Rubin (1983) ont démontré que conditionnellement au score de propension, la distribution des covariables observées à la baseline est indépendante de l'exposition. Quatre méthodes utilisant le score de propension existent pour estimer l'effet d'une exposition sur un critère de jugement.

- La stratification sur le score de propension propose de comparer le critère de jugement entre les sujets exposés et les non-exposés à l'intérieur de strates définies par le score de propension. Le

plus fréquent est d'utiliser cinq strates de tailles équivalentes basées sur les quintiles du score de propension (Austin, 2011c, 2010c). Cette méthode estime l'effet conditionnel.

- L'appariement sur le score de propension consiste à former des paires de sujets exposés et non-exposés ayant des valeurs proches du score de propension. Bien qu'il existe différentes méthodes d'appariement, la plus fréquente dans la littérature médicale est d'apparier sans remise chaque sujet exposé avec son plus proche voisin en termes de scores de propension, en respectant une distance maximale nommée *caliper*. Austin (2011b) recommande l'utilisation d'un *caliper* égal à 0,2 fois l'écart-type du logit du score de propension. L'appariement sur le score de propension estime l'effet marginal, excepté lorsqu'un modèle stratifié sur les paires est utilisé.
- L'ajustement sur le score de propension modélise l'événement d'intérêt en fonction du facteur d'exposition et du score de propension estimé. Cette méthode est la plus proche de la méthode de régression multiple classique, elle propose d'utiliser le score de propension comme une covariable résumant l'information des facteurs de confusion au lieu des facteurs de confusion eux-mêmes. Elle estime l'effet conditionnel.
- La pondération IPW consiste à pondérer la contribution des individus par des poids qui dépendent du score de propension. Cette méthode permet d'estimer l'effet marginal, elle est détaillée dans la section suivante.

Précisons que deux effets marginaux de nature différentes se distinguent (Pirracchio et al., 2013; Imbens, 2004). L'effet moyen de l'exposition dans l'ensemble de la population (en anglais, *the average treatment effect in the entire population*, ATE) et l'effet moyen de l'exposition dans la population des exposés (en anglais, *the average treatment effect on the treated*, ATT). L'ATE estime l'effet moyen associé au passage d'une population entière d'un statut d'exposé à celui de non-exposé ou *vice versa*. L'ATT estime l'effet moyen associé au passage de l'ensemble de la population exposée à un statut de non-exposé. L'ATE et l'ATT ciblent donc des populations différentes. Dans les ERC, ces deux mesures coïncident car la randomisation assure sous réserve d'un nombre de sujets suffisant que la population des exposés soit comparable à la population globale. Dans les études observationnelles, ces deux effets vont le plus souvent différer. Le choix d'estimer l'ATE ou l'ATT doit principalement dépendre de la question clinique (Austin, 2011a). Par exemple, si l'objectif est d'évaluer le bénéfice d'un traitement lourd habituellement réservé aux patients les plus à risque et qu'il n'est pas question d'administrer ce traitement à l'ensemble des patients, alors seul l'ATT est adapté. Dans un autre contexte, si l'objectif est d'évaluer le bénéfice d'une stratégie de prévention dans la population globale afin de décider de la réalisation ou non de cette campagne de prévention sur l'ensemble de la population, alors l'ATE est le plus adapté. Les données disponibles peuvent aussi influencer le choix entre l'une des deux mesures. En effet, nous verrons dans la section 2.1.6 que l'ATT requière une hypothèse de positivité beaucoup plus faible. Parmi les deux méthodes basées sur le score de propension permettant d'estimer un effet marginal décrites précédemment, l'appariement estime l'ATT alors que la pondération permet d'estimer l'ATE et l'ATT selon la pondération utilisée.

Les performances de ces quatre méthodes ont été comparées dans la littérature pour tout type de critères de jugement et selon l'effet souhaité (i.e. l'ATE ou l'ATT). Premièrement, Austin et al. (2007b) ont montré que les méthodes basées sur le score de propension (l'appariement, la stratification et l'ajustement sur le score de propension), fournissent une estimation biaisée des OR ou des HR conditionnels contrairement aux méthodes de régression conventionnelles. Ensuite, Austin (2007a) a comparé les performances de trois des quatre méthodes basées sur le score de propension (appariement, stratification et ajustement) pour l'estimation d'OR marginaux. Les résultats montrent que la méthode appariée est la plus performante en termes de biais, d'erreurs standards et de taux de couverture. Il ressort de ces

deux travaux que la stratification et l'ajustement sur le score de propension fournissent à la fois une estimation biaisée de l'effet conditionnel et de l'effet marginal. Dans un travail plus récent, Austin (2013) a confirmé ce résultat.

En 2010, Austin a comparé les performances des quatre méthodes pour estimer une différence de proportions (Austin, 2010c). La différence de proportions a la particularité d'être *collapsible*, c'est à dire que la différence de proportions marginale est égale à la différence de proportions conditionnelle (la différence de la moyenne des proportions est égale à la moyenne des différences de proportions). Les résultats montrent que la méthode de pondération IPW est la plus performante, elle présente à la fois une estimation non biaisée des différences de risques, les plus faibles erreurs standards, des intervalles de confiance avec des taux de couverture corrects et un risque de première espèce correct. Dans sa discussion, l'auteur indique que toutes les méthodes ont été comparées à des valeurs théoriques correspondantes à l'ATE alors que l'appariement estime l'ATT. Il précise que le biais obtenu avec la méthode appariée est peut-être principalement dû à cette différence d'effets estimés.

En présence de données censurées, les performances des méthodes d'appariement et d'ajustement sur le score de propension pour estimer des HR conditionnels et marginaux ont été étudiées par Gayat et al. (2012). Notons qu'au contraire des différences de proportions, les HR sont *non-collapsibles*. Dans cette étude, les auteurs montrent que le modèle de Cox marginal apparié fournit une estimation non biaisée des effets marginaux alors que le modèle de Cox stratifié sur les paires formées par le score de propension a de mauvaises performances. De plus, ils précisent qu'il est nécessaire d'utiliser un estimateur robuste de la variance pour les modèles appariés afin de prendre en compte la structure appariée des données. Ils indiquent également que le modèle de Cox ajusté sur le score de propension estime de manière biaisé à la fois les effets marginaux et les effets conditionnels. Austin (2013) a également comparé les performances des différentes méthodes basées sur le score de propension pour estimer des HR marginaux. Les résultats de son étude de simulations indiquaient que deux méthodes estimaient des HR marginaux sans biais : l'appariement et la pondération IPW, avec un avantage pour la pondération qui permet d'estimer sans biais à la fois l'ATT et l'ATE selon le système de pondération utilisé alors que l'appariement permet uniquement d'estimer l'ATT. De ces deux approches, la méthode IPW présentait une plus faible variance lors de l'estimation de l'ATT. Enfin, l'auteur indique que l'ajustement et la stratification sur le score de propension estiment de manière biaisée les HR marginaux mais aussi les HR conditionnels. Une limite majeure de cette étude de simulations est que l'auteur n'a pas induit de processus de censure dans ses simulations.

L'estimation des effets absolus d'un traitement en termes de survie (délai moyen de survie, délai médian de survie, probabilité d'occurrence d'un évènement à un temps de suivi fixé) par des méthodes basées sur le score de propension a été étudiée par Austin et Schuster (2014). La méthode stratifiée était associée au plus fort biais et à des taux d'erreurs de première espèce supérieurs à 5%, que ce soit lors de l'estimation de l'ATE ou de l'ATT. Les méthodes IPW et appariées étaient associées aux meilleurs résultats. Là encore, les auteurs n'ont pas simulé de censure.

Pour résumé, selon la nature du critère de jugement, les effets marginaux et conditionnels peuvent coïncider (différences de moyennes ou de proportions, par exemple) ou différer (OR ou HR, par exemple). Cependant, quelle que soit la nature du critère de jugement, les résultats de la littérature montrent que parmi les méthodes basées sur le score de propension, l'appariement et la pondération IPW sont les plus performantes pour obtenir une estimation marginale de l'effet. Un avantage en faveur de la méthode pondérée est sa capacité à estimer l'ATT et l'ATE selon les poids utilisés alors que l'appariement permet uniquement d'estimer l'ATT.

2.1.3 Méthode IPW

La méthode de pondération basée sur le score de propension (IPW) consiste à pondérer la contribution de chaque individu de l'échantillon par un poids inversement proportionnel à la probabilité qu'a l'individu d'appartenir à son groupe d'exposition sachant les covariables observées. Les individus ayant un profil sous-représenté vont se voir attribuer un poids élevé et les individus ayant un profil surreprésenté vont se voir attribuer un poids faible. L'objectif de la méthode est de créer un pseudo-échantillon dans lequel la distribution des covariables observées est la même dans les deux groupes (exposés et non-exposés). Ainsi, une situation de quasi-randomisation est créée artificiellement. La différence majeure avec un ERC étant que ce dernier assure, sous réserve d'une taille d'échantillon suffisante, l'équilibre de toutes les covariables alors que la méthode IPW ne garantit l'équilibre que pour les covariables observées.

Plusieurs pondérations ont été proposées, la première l'a été par Horvitz et Thompson (1952) et Rosenbaum (1987). Elle pondère la contribution de chaque individu par un poids w_i égal à l'inverse probabilité qu'a l'individu i d'appartenir à son groupe d'exposition sachant les covariables observées :

$$w_i = X_i/p_i + (1 - X_i)/(1 - p_i) \tag{2.1}$$

Ces poids présentent deux inconvénients. Le premier est que la somme des poids des individus est égale au double de la taille de l'échantillon initial. Xu et al. (2010) indiquent que cette augmentation de la taille de l'échantillon peut conduire à une sous-estimation de la variance et donc produire des intervalles de confiance trop restrictifs et une augmentation des rejets à tort des hypothèses nulles. Le second inconvénient concerne la possible obtention de poids très importants pour quelques individus ayant un profil très rare, ces individus ont alors des contributions trop importantes dans l'estimation du modèle.

Pour contrer ces écueils, des auteurs ont proposé des poids dits stabilisés (en anglais, *stabilized weights*, sw) (Robins, 1998; Robins et al., 2000; Cole et Hernán, 2004; Xu et al., 2010). Ces poids incluent au numérateur la probabilité marginale qu'a l'individu i d'appartenir à son groupe d'exposition :

$$sw_i = X_i P(X_i = 1)/p_i + (1 - X_i)P(X_i = 0)/(1 - p_i) \tag{2.2}$$

L'utilisation des poids stabilisés permet d'obtenir un pseudo-échantillon de taille similaire à celle de l'échantillon initial, de plus Xu et al. (2010) ont montré qu'ils permettent d'obtenir une estimation correcte de la variance de l'effet estimé et du taux d'erreur de première espèce. D'autres auteurs ont proposé de limiter la contribution maximale d'un individu à l'analyse. Liu et al. (2007) ont par exemple proposé d'utiliser au dénominateur le maximum entre le score de propension de l'individu i et un multiple du score de propension moyen. Ainsi, le poids des individus ayant une très faible probabilité d'appartenir à leur groupe d'exposition est borné. Enfin, pour prendre en compte la nature pondérée du pseudo-échantillon, plusieurs auteurs recommandent d'utiliser un estimateur robuste de la variance (Lin et Wei, 1989; Cole et Hernán, 2004; Joffe et al., 2004; Austin et Stuart, 2015).

Les poids décrits ci-dessus permettent d'estimer l'ATE, parce qu'ils sont conçus pour utiliser l'échantillon global comme population de référence. Autrement dit, dans l'échantillon pondéré la distribution des covariables dans chaque groupe d'exposition est la même que la distribution des covariables dans l'échantillon non pondéré global. Il est possible d'estimer l'ATT en utilisant des poids égaux à $X_i + \frac{(1-X_i)p_i}{1-p_i}$ (Morgan et Todd, 2008; Austin, 2014), les individus exposés se voient alors tous attribuer un poids égal à 1, alors que les individus non-exposés se voient attribuer un poids égal à $\frac{p_i}{1-p_i}$. Ainsi, la population de référence est l'échantillon des individus exposés. Dans l'échantillon pondéré, la distribution des covariables dans chaque groupe d'exposition est alors identique à la distribution des covariables chez les individus exposés dans l'échantillon original (i.e. non pondéré).

2.1.4 Sélection des variables d'ajustement

Quatre ensembles de variables peuvent être inclus dans le modèle du score de propension : i) toutes les variables associées à la fois à l'exposition et au critère de jugement d'intérêt (vrais facteurs de confusion), ii) toutes les variables associées à l'exposition (indépendantes ou non du critère de jugement), iii) toutes les variables associées au critère de jugement (indépendantes ou non de l'exposition), et iv) toutes les variables observées. Intuitivement, il semblerait préférable de choisir les méthodes i) ou ii), la première car d'après la définition d'un facteur de confusion ce sont les seules variables pouvant biaiser l'estimation causale de l'effet, la seconde pour équilibrer la distribution des covariables entre les groupes d'exposition et créer artificiellement une situation de quasi-randomisation. Plusieurs auteurs ce sont intéressés à cette problématique. Brookhart et al. (2006) indiquent qu'il est préférable d'inclure dans le score de propension toutes les variables associées au critère de jugement. Ces résultats concordent avec ceux de Austin et al. (2007a) et de Austin et al. (2007b) en faveur de l'inclusion des vrais facteurs de confusion ou de toutes les variables associées au critère de jugement, mais sont discordants des résultats d'une étude de simulations s'intéressant à l'estimation d'OR marginaux et concluant que les plus faibles biais et les plus faibles variances sont obtenus en incluant soit les variables liées à l'exposition, soit l'ensemble des variables observées (Austin, 2007a). Notons que ces études s'intéressaient aux méthodes appariée, stratifiée et ajustée sur le score de propension, mais aucune n'incluait la méthode pondérée.

Une explication aux résultats assez contre-intuitifs en faveur de l'inclusion des vrais facteurs de confusion ou de toutes les variables associées au critère de jugement est fourni par Austin (2011a). Il explique que les variables associées à l'exposition et indépendantes du critère de jugement ne sont pas source de biais et qu'il n'est donc pas nécessaire de les inclure dans le score de propension. De plus, les inclure dans le score de propension peut augmenter la variance de l'estimation de l'effet de l'exposition.

2.1.5 Avantages des méthodes basées sur le score de propension

Une question fréquemment posée concerne l'avantage des méthodes basées sur le score de propension par rapport aux méthodes de régression plus conventionnelles. Premièrement les méthodes basées sur le score de propension permettent d'estimer des effets marginaux, autrement dit l'effet moyen associé au passage de la population entière du statut d'exposé au statut de non-exposé. Cet effet marginal est celui estimé par l'effet brut (non ajusté) dans les ERC. Notons cependant qu'il n'est pas impossible à partir de modèles de régression multiples d'estimer indirectement l'effet marginal (Bender et Blettner, 2002; Austin, 2010a,b). Le choix de la méthode doit donc avant tout dépendre de la question clinique.

Un autre avantage des méthodes basées sur le score de propension (à l'exception de l'ajustement sur le score de propension) est que ces méthodes permettent aux études observationnelles d'avoir une présentation des résultats apparaissant semblable à celle des ERC, argument convaincant pour les relecteurs et les éditeurs des journaux biomédicaux. Enfin, le mimétisme de ces méthodes avec l'ERC les rend facilement compréhensibles pour des non-spécialistes de la Biostatistique, et leur confère un avantage sur les aspects d'interprétation non négligeable.

Concernant l'appariement et la stratification, le score de propension présente l'avantage de synthétiser l'information de l'ensemble des facteurs de confusion en une seule variable et offre ainsi une mise en œuvre simplifiée de la stratification et de l'appariement. Dans le cadre des modèles multi-états un travail en cours dans notre équipe montre l'avantage de la méthode IPW en termes de simplicité de mise en œuvre et de gain de temps (Gillaizeau, 2015).

2.1.6 Hypothèses

Les méthodes d'inférence causale basées sur le score de propension requièrent quatre hypothèses : la consistance, l'échangeabilité, la positivité et l'absence d'erreur de spécification du score de propension (Cole et Hernán, 2008; Austin et Stuart, 2015).

Consistance

La consistance se définit comme le fait que l'événement potentiel d'un individu sous l'exposition qu'il a vraiment reçu est égal à l'événement réellement observé pour l'individu. Cette définition fait appel à la notion d'événement potentiel aussi connue sous le nom d'événement contrefactuel qui attribue à chaque individu i une paire d'événements potentiels : $Y_i(1)$ et $Y_i(0)$, respectivement l'événement qui aurait été observé si l'individu avait reçu l'exposition et l'événement qui aurait été observé si l'individu n'avait pas reçu l'exposition. La consistance se définit alors comme : $Y_i(x) = Y_i$ si $X_i = x$ (Cole et Hernán, 2008). Le score de propension est le plus souvent estimé par un modèle de régression logistique, la consistance du score de propension repose alors sur la correcte spécification du modèle.

Echangeabilité

L'hypothèse d'échangeabilité des individus exposés et non-exposés est garantie dans les ERC par le tirage au sort. Dans les études observationnelles, elle se traduit par l'absence de facteurs de confusion non mesurés. Cette hypothèse n'est pas vérifiable empiriquement et seule la collecte de l'ensemble des variables qui affectent l'exposition et/ou le critère de jugement permettrait de s'assurer de cette hypothèse.

Positivité

Les méthodes basées sur le score de propension estiment des effets marginaux indiquant que tous les individus auraient pu être exposés ou non-exposés. Autrement dit, tous les individus doivent avoir une probabilité non nulle de recevoir chaque exposition : $0 < P(X = 1|Z = z) < 1$. Cette hypothèse est appelée hypothèse de positivité. Elle se vérifie par la présence de patients exposés et non-exposés pour chacune des combinaisons possibles des facteurs de confusion observés dans l'échantillon d'étude (Westreich et Cole, 2010; Cole et Hernán, 2008). Pirracchio et al. (2013) précisent que lors de l'estimation de l'ATT, l'hypothèse de positivité est plus faible, réduite au fait que tous les individus exposés auraient pu être non-exposés. Elle se traduit formellement par : $P(X = 0|Z = z) > 0$ pour chaque valeur z observée parmi les exposés.
Dans les études observationnelles, la violation de cette hypothèse peut être déterministe ou aléatoire. Elle est déterministe si pour au moins une combinaison des facteurs de confusion observés, les sujets ne peuvent pas recevoir l'ensemble des expositions étudiées. C'est le cas par exemple lorsqu'une contre-indication à un traitement empêche certains patients d'être observés dans un des groupes. En l'absence de violation déterministe, l'hypothèse de positivité peut quand même être violée de manière aléatoire, notamment en présence d'un échantillon de petite taille et/ou d'un grand nombre de covariables observées. En présence de covariables continues, il existe un nombre infini de combinaisons possibles entrainant une violation quasi-certaine de l'hypothèse de positivité. Dans de tel cas, l'utilisation de modèles paramétriques permet de modéliser le score de propension en lissant l'information à partir des patients ayant des profils proches de ceux ayant une probabilité égale à zéro de recevoir l'une des expositions (Westreich et Cole, 2010).
Pour détecter une violation de cette hypothèse, Cole et Hernán (2008) préconisent de regarder la distri-

bution des poids stabilisés. Si la moyenne des poids est éloignée de 1, ou s'il existe des valeurs extrêmes (i.e. des poids élevés comparativement à la taille de l'échantillon), alors ceci peut indiquer une violation de l'hypothèse de positivité ou une mauvaise spécification du modèle. En cas de violation déterministe de l'hypothèse de positivité et si l'ATE est la mesure choisie, la question qui se pose est de savoir si l'ATT peut être estimé plus facilement et si ce n'est pas la mesure de l'effet la plus pertinente (Stuart, 2010). Une autre solution pour corriger une violation déterministe de l'hypothèse de positivité est de restreindre la population d'étude en excluant les patients ayant un profil ne pouvant pas recevoir l'une des deux expositions, partant du constat que la population cible de l'étude avait été mal choisie. Dans le cas d'une violation aléatoire de l'hypothèse, Westreich et Cole (2010) recommandent d'investiguer le choix des covariables afin de trouver un compromis entre la réduction des biais de confusion et l'augmentation des biais et de la variance due à un non respect de l'hypothèse de positivité.

Bonne spécification du score de propension

Le score de propension est un score d'équilibre : conditionnellement au vrai score de propension, les sujets exposés et non exposés ont la même distribution des covariables observées. Par conséquent, la bonne spécification du score de propension peut être évaluée en comparant la distribution des covariables mesurées dans les échantillons pondérés, appariés ou stratifiés entre les exposés et des non-exposés (Austin et Stuart, 2015). Différentes méthodes ont été proposées pour examiner l'équilibre de la distribution des covariables entre les exposés et les non-exposés (Ali et al., 2015). L'approche recommandée et la plus fréquemment utilisée dans la littérature est l'examen des différences standardisées pour chacune des covariables. Par exemple, pour la méthode IPW il s'agit de calculer la différence moyenne entre les groupes des exposés et des non-exposés divisée par l'écart type dans l'échantillon pondéré (Austin et Stuart, 2015). Pour les variables binaires, la différence standardisée est ainsi définie par :

$$d = \frac{\hat{p}_{X=1} - \hat{p}_{X=0}}{\sqrt{\frac{\hat{p}_{X=1}(1-\hat{p}_{X=1})+\hat{p}_{X=0}(1-\hat{p}_{X=0})}{2}}} \tag{2.3}$$

où $\hat{p}_{X=1}$ et $\hat{p}_{X=0}$ correspondent aux proportions de sujets dans l'échantillon pondéré présentant la caractéristique d'intérêt respectivement chez les exposés et les non-exposés. Pour les variables continues, la différence standardisée est définie par :

$$d = \frac{\bar{z}_{X=1} - \bar{z}_{X=0}}{\sqrt{\frac{s^2_{X=1}+s^2_{X=0}}{2}}} \tag{2.4}$$

où $\bar{z}_{X=1}$ et $\bar{z}_{X=0}$ correspondent aux moyennes pondérées de la variable d'intérêt dans respectivement les groupes des exposés et des non-exposés, et $s^2_{X=1}$ et $s^2_{X=0}$ sont les variances pondérées correspondantes : $s^2_{X=x} = \frac{\sum \mathbb{1}(X=x)w_i}{\left(\sum \mathbb{1}(X=x)w_i\right)^2 - \sum \mathbb{1}(X=x)w_i^2} \sum \mathbb{1}(X=x)w_i(z_i-\bar{z})^2$ (Austin et Stuart, 2015).
Le déséquilibre des facteurs de confusion entre les groupes d'exposition est considéré comme négligeable lorsque la différence standardisée est inférieure à 10% (Austin, 2011a; Normand et al., 2001). Ali et al. (2015) décrivent cet indicateur comme facile à calculer, simple d'interprétation et insensible à la taille de l'échantillon, la seule limitation étant le choix arbitraire du seuil.
Austin (2011a) indique que la différence standardisée permet de comparer les moyennes ou proportions des covariables entre les groupes d'exposition dans l'échantillon pondéré, apparié ou stratifié. Cependant, le score de propension doit assurer l'équilibre de l'entière distribution des covariables et pas seulement de la moyenne. Par conséquent, il recommande de compléter par une approche graphique pour les facteurs continus et de comparer les interactions entre les covariables selon les groupes d'exposition.

En l'absence d'équilibre dans les échantillons pondérés, appariés ou stratifiés, l'estimation finale de l'effet peut être sujette à des biais résiduels de confusion. Il est donc nécessaire de rechercher un meilleur score de propension. Parmi les solutions possibles, notons l'inclusion de facteurs supplémentaires, d'interactions entre les facteurs ou d'effets non linéaires pour les facteurs continus.
Dans la littérature, il ne semble pas clairement indiqué si l'équilibre doit être vérifié pour l'ensemble des covariables observées ou uniquement pour celles incluses dans le score de propension. Cependant, dans la section 2.1.4 relative à la sélection des variables à inclure dans le score de propension, nous avons vu qu'il était préférable de ne pas inclure l'ensemble des variables associées à l'exposition, mais de se limiter aux variables associées au critère de jugement. En effet, les variables non associées au critère de jugement (même si elles sont déséquilibrées entre les groupes d'exposition) ne biaisent pas l'estimation de l'effet d'intérêt et ne pas les inclure permet une meilleure estimation de la variance. Ainsi, choisir une telle stratégie de sélection de variables implique d'accepter que certains facteurs ne soient pas équilibrés entre les groupes d'exposition (contrairement à ce qui est observé dans un ERC) et la vérification de l'hypothèse d'équilibre devrait être appliquée uniquement aux variables incluses dans le score de propension.

2.2 Analyses de survie

L'analyse de survie est l'étude du délai de la survenue d'un événement, analyses couramment réalisées à partir des cohortes. Une des caractéristiques des données de survie est l'existence d'observations incomplètes, induites par exemple par des sorties d'études ou des temps de suivi insuffisamment longs pour observer l'événement chez tous les individus.
La variable aléatoire étudiée est le délai de survenue de l'événement d'intérêt à partir d'une date d'origine. Nous noterons $\tilde{T}$ cette variable dans la suite de ce manuscrit. En présence de données censurées à droite, le temps de survie $\tilde{T}$ n'est pas observé pour tous les individus, seul le couple (T,δ) l'est :

$$(T,\delta) \text{ avec } \begin{cases} T = \min(\tilde{T},C) \\ \delta = \mathbb{1}(\tilde{T} \leq C) \end{cases} \tag{2.5}$$

où C est le temps de censure et T est le temps d'événement ou le temps de censure. Plusieurs fonctions liées entre elles et définies sur $\mathbb{R}+$ définissent $\tilde{T}$:

- **La fonction de densité de probabilité notée** $f(t)$. Elle représente la probabilité de connaître l'événement dans un petit intervalle Δt suivant t :

$$f(t) = \lim_{\Delta t \to 0^+} P(t \leq \tilde{T} < t + \Delta t)/\Delta t \tag{2.6}$$

- **La fonction de répartition** $F(t)$. Elle représente la probabilité que l'événement se produise entre 0 et t :

$$F(t) = P(\tilde{T} \leq t) = \int_0^t f(u)du \tag{2.7}$$

- **La fonction de survie** $S(t)$. Elle mesure la probabilité que l'évènement ne se produise pas avant t :

$$S(t) = P(\tilde{T} > t) = 1 - F(t) = \int_t^{+\infty} f(u)du \tag{2.8}$$

- **La fonction de risque instantané** $\lambda(t)$. Elle représente la probabilité que l'événement se produise dans un petit intervalle Δt juste après t, conditionnellement au fait que l'événement n'ai pas

eu lieu jusqu'à t :

$$\lambda(t) = \lim_{\Delta t \to 0^+} P(t \leq \tilde{T} < t + \Delta t | \tilde{T} \geq t)/\Delta(t) = f(t)/S(t) \tag{2.9}$$

- **La fonction de risque cumulée** $\Lambda(t)$, est définie par :

$$\Lambda(t) = \int_0^t \lambda(u)du \tag{2.10}$$

A partir des définitions précédentes, on retrouve les relations suivantes :

$$f(t) = -dS(t)/d(t) \tag{2.11}$$

$$S(t) = \exp(-\int_0^t \lambda(u)du) = \exp(-\Lambda(t)) \tag{2.12}$$

2.2.1 Estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier, est l'estimateur non paramétrique le plus fréquemment utilisé pour estimer la fonction de survie lorsque l'on étudie le délai de survenue d'un évènement pour un ou plusieurs groupe d'individus en présence de données censurées à droite (Kaplan et Meier, 1958). L'estimateur de Kaplan-Meier s'appuie sur le théorème des probabilités conditionnelles et découle du raisonnement suivant : ne pas encore avoir subi l'évènement à l'instant t c'est ne pas l'avoir subi juste avant t et ne pas le subir en t. Cet estimateur nécessite une indépendance entre l'événement et la censure.
Soient $t_1 < \cdots < t_k$ les différents temps ordonnés auxquels sont observés des évènements, on peut définir :

- d_j : le nombre de sujets subissant l'évènement au temps t_j
- c_j : le nombre de sujets censurés dans $[t_j,t_{j+1}[$
- Y_j : le nombre de sujets à risque au temps t_j (effectif à risque)

Comme ne pas encore avoir subi l'évènement en t_j est équivalent à ne pas encore avoir subi l'évènement en t_{j-1} et ne pas le subir en t_j, on peut calculer l'estimateur de Kaplan-Meier $\hat{S}(t)$ de la fonction de survie d'après les probabilités conditionnelles :

$$\hat{S}(t_j) = \hat{P}(\tilde{T} > t_j) \tag{2.13}$$

$$= \hat{P}(\tilde{T} > t_j | \tilde{T} > t_{j-1})\hat{P}(\tilde{T} > t_{j-1}) \tag{2.14}$$

$$= \hat{P}(\tilde{T} > t_j | \tilde{T} > t_{j-1}) \dots \hat{P}(\tilde{T} > t_2 | T > t_1)\hat{P}(\tilde{T} > t_1) \tag{2.15}$$

On estime $P(\tilde{T} > t_j | \tilde{T} > t_{j-1})$ par $(Y_j - d_j)/Y_j$. La probabilité de survivre après t est donnée par le produit de ces probabilités estimées pour tous les temps d'évènements antérieures à t.

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{Y_j}\right) \tag{2.16}$$

La fonction de survie obtenue $\hat{S}(t)$ est donc une fonction en escalier décroissante, constante entre deux temps d'évènements consécutifs. Elle n'est pas définie au-delà de la dernière observation si celle-ci est censurée (la fonction n'est pas nulle pour cette dernière observation mais au-delà l'effectif à risque étant nul, on ne peut plus la calculer).

2.2.2 Test du Log-rank

Soient d_{jx} et Y_{jx} respectivement le nombre de sujets subissant l'évènement et le nombre de sujets à risque au temps t_j dans le groupe d'exposition $X = x$, avec $X = \{0,1\}$. Soient $d_j = d_{j0} + d_{j1}$ le nombre total de sujets subissant l'évènement au temps t_j et $Y_j = Y_{j0} + Y_{j1}$ le nombre total de sujets à risque au temps t_j. L'hypothèse nulle H_0 testée est l'égalité des fonctions de survie dans les deux groupes. La statistique du Log-rank propose de comparer à chaque temps d'événement le nombre observé de sujets subissant l'événement au nombre attendu de sujets subissant l'événement sous l'hypothèse nulle. Elle est définie par (Mantel, 1966; Peto et Peto, 1972) :

$$G = \frac{\sum_{j=1}^{k}(d_{j1} - Y_{j1}\frac{d_j}{Y_j})}{\sqrt{\sum_{j=1}^{k} v_j}} \tag{2.17}$$

avec $v_j = [Y_{j0}Y_{j1}d_j(Y_j - d_j)]/[(Y_j)^2(Y_j - 1)]$. La statistique de test suit asymptotiquement une loi normale centrée réduite sous H_0.

2.2.3 Estimateur de Kaplan-Meier pondéré

Dans les articles biomédicaux, les courbes de survie estimées à l'aide de l'estimateur de Kaplan-Meier sont fréquemment présentées dans une première partie descriptive avant d'estimer l'effet ajusté de l'exposition par un modèle de Cox multiple. En effet, les courbes de survie sont très appréciées car elles permettent d'illustrer précisément les différences de survie entre les groupes au cours du temps. Cependant, dans toutes les études observationnelles où la présence de facteurs de confusion est commune, l'estimateur de Kaplan-Meier est inadéquat et fournit une estimation biaisée de l'effet propre de l'exposition sur la survie. Pour obtenir des courbes de survie non biaisées prenant en compte les facteurs de confusion, Cole et Hernán (2004) et Xie et Liu (2005) ont séparément proposé une méthode basée sur la pondération IPW. Soient $d_j^w = \sum_{i:t_i=t_j} sw_i\delta_i$ le nombre pondéré de sujets subissant l'événement au temps t_j et $Y_j^w = \sum_{i:t_i\geq t_j} sw_i$ le nombre pondéré de sujets à risque au temps t_j, en utilisant les poids stabilisés sw. L'estimateur pondéré de Kaplan-Meier est simplement une adaptation de l'estimateur usuel, tel que :

$$\hat{S}^w(t) = \prod_{j:t_j\leq t}\left(1 - \frac{d_j^w}{Y_j^w}\right) \tag{2.18}$$

2.2.4 Tests du Log-rank pondérés

Identiquement à l'estimateur de Kaplan-Meier, le test du Log-rank devient inadéquat en présence de facteurs de confusion. Pour répondre à cette problématique, deux tests du Log-rank pondérés ont été proposés dans la littérature, le premier par Xie et Liu (2005) et le second par Sugihara (2010).

Xie et Liu (2005) proposent de préalablement ajuster les poids de chaque individu à chaque temps d'événement t_j selon les individus encore à risque dans le groupe d'exposition tel que $sw_{ij}^{'} = sw_iY_{j1}/Y_{j1}^w$ pour un individu i dans le groupe $X = 1$ et $sw_{ij}^{'} = sw_iY_{j0}/Y_{j0}^w$ pour un individu i dans le groupe $X = 0$. A partir de ces poids on peut définir $d_{jx}^{w'} = \sum_{i:t_i=t_j} sw_{ij}^{'}\delta_i\mathbb{1}(X_i = x)$ et $Y_{jx}^{w'} = \sum_{i:t_i\geq t_j} sw_{ij}^{'}\mathbb{1}(X_i = x)$ respectivement le nombre pondéré de sujets subissant l'événement au temps t_j et le nombre pondéré de sujets à risque au temps t_j dans le groupe $X = x$. La statistique de test proposée par Xie et Liu (2005) est : $G^{w'}/\sqrt{\mathrm{Var}(G^{w'})}$ avec :

$$G^{w'} = \sum_{j=1}^{k} d_{j1}^{w'} - Y_{j1}^{w'}\frac{d_j^{w'}}{Y_j^{w'}} \tag{2.19}$$

$$\mathrm{Var}(G^{w'}) = \sum_{j=1}^{k} \left\{ \frac{d_j(Y_j - d_j)}{Y_j(Y_j-1)} \sum_{i=1}^{Y_j} \left[\left(\frac{Y_{j0}^{w'}}{Y_j^{w'}} \right)^2 sw_{ij}^{'2} X_i + \left(\frac{Y_{j1}^{w'}}{Y_j^{w'}} \right)^2 sw_{ij}^{'2} (1 - X_i) \right] \right\} \tag{2.20}$$

où $d_j^{w'} = d_{j0}^{w'} + d_{j1}^{w'}$ est le nombre pondéré de sujets subissant l'événement au temps t_j et $Y_j^{w'} = Y_{j0}^{w'} + Y_{j1}^{w'}$ est le nombre pondéré de sujets à risque au temps t_j. La statistique de test suit asymptotiquement sous H_0 une loi normale centrée réduite.

Le test du Log-rank pondéré proposé par Sugihara (2010) est équivalent excepté l'étape d'ajustement des poids au cours du temps qui n'est pas réalisée. Ainsi le nombre pondéré de sujets subissant l'événement au temps t_j et le nombre pondéré de sujets à risque au temps t_j dans le groupe $X = x$ sont respectivement définis par $d_{jx}^w = \sum_{i:t_i=t_j} sw_i \delta_i \mathbb{1}(X_i = x)$ et $Y_{jw}^w = \sum_{i:t_i \geq t_j} sw_i \mathbb{1}(X_i = x)$. La statistique de test correspondante est : $G^w / \sqrt{\mathrm{Var}(G^w)}$ avec :

$$G^w = \sum_{j=1}^{k} d_{j1}^w - Y_{j1}^w \frac{d_j^w}{Y_j^w} \tag{2.21}$$

$$\mathrm{Var}(G^w) = \sum_{j=1}^{k} \left\{ \frac{d_j(Y_j - d_j)}{Y_j(Y_j-1)} \sum_{i=1}^{Y_j} \left[\left(\frac{Y_{j0}^w}{Y_j^w} \right)^2 sw_i^2 X_i + \left(\frac{Y_{j1}^w}{Y_j^w} \right)^2 sw_i^2 (1 - X_i) \right] \right\} \tag{2.22}$$

2.2.5 Modèle de Cox

Proposé par Cox (1972), le modèle semi-paramétrique à risques proportionnels permet d'établir une relation entre les facteurs de risque de la survenue de l'événement et la distribution des durées de survie, sans définir une forme paramétrique particulière à cette dernière. C'est le modèle le plus utilisé en analyse de cohortes pour étudier l'effet de variables explicatives sur le risque de survenue d'un événement.

Soit $Z_i = (Z_{1i}, \ldots, Z_{pi})$ le vecteur des p covariables potentielles facteurs de confusion caractérisant l'individu i et $\beta = (\beta_1, \ldots, \beta_p)$ le vecteur des coefficients de régression associés. Le modèle de Cox s'écrit :

$$\lambda(t|x,z) = \lambda_0(t) \exp(\alpha x + \beta' z) \tag{2.23}$$

où α est le coefficient de régression associé à l'exposition et $\lambda_0(t)$ est la fonction de risque (instantané) de base. La fonction exponentielle a été choisie pour obtenir une fonction de risque positive sans contrainte sur les valeurs des coefficients.

La particularité du modèle semi-paramétrique est qu'il ne nécessite pas l'estimation de la fonction de risque de base. En effet, la vraisemblance partielle, notée $\mathcal{VP}$, correspond au produit des probabilités conditionnelles calculées à chaque temps d'événement et sa maximisation permet d'estimer les coefficients de régression. Plus précisément, en considérant qu'il n'y a qu'un seul événement en chaque temps t_j, elle s'écrit :

$$\mathcal{VP} = \prod_{i=1}^{n} \left\{ \frac{\lambda_0(t_i) \exp(\alpha x_i + \beta' z_i)}{\sum_{l:t_l \geq t_i} \lambda_0(t_i) \exp(\alpha x_l + \beta' z_l)} \right\}^{\delta_i} = \prod_{i=1}^{n} \left\{ \frac{\exp(\alpha x_i + \beta' z_i)}{\sum_{l:t_l \geq t_i} \exp(\alpha x_l + \beta' z_l)} \right\}^{\delta_i} \tag{2.24}$$

où l indique les individus encore à risque au temps t_j. Il est généralement plus agréable de considérer la log-vraisemblance, elle s'écrit :

$$\log \mathcal{VP} = \sum_{i=1}^{n} \delta_i \left\{ \alpha x_i + \beta' z_i - \log \Big(\sum_{l:t_l \geq t_i} \exp(\alpha x_l + \beta' z_l) \Big) \right\} \tag{2.25}$$

Le modèle de Cox est basé sur deux hypothèses : la proportionnalité des risques et la log-linéarité.

Hypothèse de proportionnalité des risques

La validité de l'effet estimé repose sur l'hypothèse que les fonctions de risque pour les différentes modalités de X sont proportionnelles et leur rapport est constant au cours du temps $\text{HR} = \exp(\alpha)$.
Pour vérifier cette hypothèse, de nombreuses méthodes graphiques et analytiques existent, chacune avec ses avantages et inconvénients. Parmi les méthodes graphiques, on peut par exemple tracer pour chaque groupe les courbes $log[-log(\hat{S}(t))]$ en fonction du temps et vérifier que les courbes présentent un écart constant au cours du temps t. Les méthodes graphiques permettent une première approche visuelle mais ont pour défauts leur subjectivité car dépendantes de l'interprétation qu'en fait l'analyste et le fait que l'on peut difficilement prendre en compte les autres variables d'ajustement. La méthode graphique peut donc être complétée par une approche analytique, telle que l'analyse des résidus de Schoenfeld (Schoenfeld, 1982).

Hypothèse de log-linéarité

Dans le modèle de Cox (comme dans un modèle logistique), la relation entre les variables explicatives et la fonction de risque est log-linéaire. Cela signifie que pour une variable explicative continue introduite dans le modèle, une augmentation d'une unité de cette variable est associée à un HR égal à $\exp(\beta)$, quelque soit sa valeur initiale. Pour vérifier cette hypothèse une solution consiste à transformer la variable quantitative en variable catégorielle à plusieurs classes (déterminées habituellement selon les quantiles) puis à représenter graphiquement les logit du modèle incluant la variable transformée en fonction des niveaux de la variable. L'hypothèse de log-linéarité sera acceptable si la relation entre ces points est linéaire. Une autre solution consiste à représenter sur le même graphique les valeurs prédites par le modèle incluant la variable quantitative classiquement avec celles prédites par le modèle lorsqu'une fonction spline est appliquée à la variable quantitative. La proximité des courbes obtenues et la forme linéaire de la courbe obtenue lorsqu'une fonction spline est appliquée à la variable quantitative indiqueront le respect de l'hypothèse de log-linéarité.
Si l'hypothèse de log-linéarité semble acceptable alors la variable pourra être gardée dans le modèle sous sa forme quantitative, dans le cas contraire il faudra par exemple, introduire la variable sous une forme catégorielle ou appliquer une fonction spline à la variable. La dernière solution rendant l'effet de la variable sur l'événement étudié difficilement interprétable.

2.2.6 Modèle de Cox pondéré (IPW)

Comme le modèle de Cox multiple, le modèle de Cox pondéré permet de prendre en compte des facteurs de confusion lors de l'estimation de l'effet d'un exposition sur la survenue d'un événement afin d'estimer sans biais un effet causal. La principale différence entre ces deux méthodologies est la nature de l'effet estimé, conditionnelle pour le premier et marginale pour le second (voir section 2.1.1). Le modèle de Cox pondéré, utilisant des poids stabilisés et un estimateur robuste de la variance, a été suggéré par Cole et Hernán (2004). Il consiste en un modèle univarié dans lequel seul le facteur d'exposition d'intérêt est inclus. La vraisemblance partielle pondérée correspondante s'écrit alors (Binder, 1992; Lumley et Scott, 2013) :

$$\log \mathcal{VP} = \sum_{i=1}^{n} sw_i \delta_i \left\{ \alpha x_i + \beta' z_i - \log \Big(\sum_{l:t_l \geq t_i} sw_l \exp(\alpha x_l + \beta' z_l) \Big) \right\} \tag{2.26}$$

2.2.7 Problématique

Plusieurs auteurs ont comparé les performances des méthodes basées sur le score de propension pour estimer des HR en présence de données censurées (Gayat et al., 2012; Austin, 2013; Austin et Schuster, 2014). Dans ces études les méthodes sont principalement comparées en termes de biais. Austin et Schuster (2014) ont comparé les performances des tests statistiques correspondants à chacune des méthodes en termes de risque de première espèce mais sans présenter le risque de seconde espèce. Pourtant, les conclusions fondées sur les taux d'erreurs de première espèce sans être contrebalancées par les taux d'erreurs de seconde espèce peuvent conduire au choix d'un test associé à une faible puissance statistique. Cette littérature incomplète ne conduit pas à un consensus sur la méthode la plus appropriée pour estimer l'effet marginal d'une exposition en prenant en compte les facteurs de confusion. Cela explique probablement une partie de l'importante variabilité des pratiques observées dans les articles biomédicaux. Afin de compléter cette littérature, nous proposons dans le chapitre 3 une étude de simulations comparant les performances en termes d'erreurs de première et seconde espèces de différents modèles permettant d'évaluer l'effet marginal d'une exposition sur un événement en présence de données censurées à droite.

2.3 Analyses pronostiques

Deux axes d'analyses pronostiques peuvent être distingués : i) la construction d'un score ou l'identification d'un marqueur permettant de prédire le statut du patient, et ii) l'évaluation des capacités pronostiques du score ou du marqueur. Dans cette thèse, nous nous sommes intéressés à l'étude des capacités d'un marqueur X à discriminer une population d'individus ayant subi l'événement d'intérêt au temps τ d'une population d'individus n'ayant pas subi l'événement d'intérêt à ce même temps.
La courbe ROC (*Receiver Operating Characteristics*) dépendante du temps construite à partir des sensibilités et spécificités dépendantes du temps, est la méthode la plus couramment utilisée pour évaluer les capacités discriminantes d'un marqueur. A l'origine développée pour évaluer des tests diagnostiques (en l'absence de censure), la courbe ROC usuelle a ensuite été adaptée aux données censurées par plusieurs auteurs. Heagerty et al. (2000) ont proposé des estimateurs de la sensibilité et de la spécificité dépendants du temps basés sur le théorème de Bayes, et sur les estimateurs de Kaplan-Meier et d'Akritas (Akritas, 1994). Chambless et Diao (2006) ont ensuite proposé deux méthodes alternatives basées sur l'estimateur de Kaplan-Meier et un algorithme récursif pour la première et sur le théorème de Bayes et le modèle de Cox pour la seconde. Uno et al. (2007) et Hung et Chiang (2010) ont indépendamment proposé des estimateurs pondérés par l'inverse de la probabilité d'être censuré (en anglais, *inverse probability of censoring weighting*, IPCW). Une revue récente a été proposée par Blanche et al. (2013). Seuls les estimateurs IPCW seront présentés dans les sections suivantes car ce sont ceux qui seront adaptés dans le chapitre 4 afin de prendre en compte des covariables.

2.3.1 Sensibilité et spécificité dépendantes du temps

La qualité pronostique d'un marqueur se définit par sa capacité à correctement discriminer les individus en deux groupes : ceux qui subiront l'événement et ceux qui ne subiront pas l'événement. Notons X le marqueur pronostique d'intérêt inversement corrélé à $\tilde{T}$, v une valeur seuil de X et τ le temps pronostique. Le test pronostique est dit positif si $X > v$ et négatif sinon. A partir des travaux d'Heagerty et al. (2000), on définit :

- **la sensibilité dépendante du temps** qui représente la proportion de sujets avec un marqueur supérieur au seuil v parmi les patients qui subissent l'événement avant τ. Elle s'écrit :

$$se(\tau,v) = P(X > v|\tilde{T} \leq \tau) \tag{2.27}$$

- **la spécificité dépendante du temps** qui représente la proportion de sujets avec un marqueur inférieur ou égal au seuil v parmi les patients qui n'ont pas subi l'événement au temps τ. Elle s'écrit :

$$sp(\tau,v) = P(X \leq v|\tilde{T} > \tau) \tag{2.28}$$

A l'aide du théorème des probabilités conditionnelles ces deux probabilités peuvent être développées pour obtenir :

$$se(\tau,v) = P(X > v,\tilde{T} \leq \tau)/P(\tilde{T} \leq \tau) \tag{2.29}$$

$$sp(\tau,v) = P(X \leq v,\tilde{T} > \tau)/P(\tilde{T} > \tau) \tag{2.30}$$

Pour prendre en compte l'information apportée par les individus censurés, les deux estimateurs IPCW de la sensibilité et de la spécificité dépendantes du temps (Blanche et al., 2013) :

$$\widehat{se}(\tau,v) = \frac{\sum_{i=1}^{n} \delta_i \mathbb{1}(T_i \leq \tau,X_i > v)/\hat{S}_C(T_i)}{\sum_{i=1}^{n} \delta_i \mathbb{1}(T_i \leq \tau)/\hat{S}_C(T_i)} \tag{2.31}$$

$$\widehat{sp}(\tau,v) = \frac{\sum_{i=1}^{n} \mathbb{1}(T_i > \tau,X_i \leq v)}{\sum_{i=1}^{n} \mathbb{1}(T_i > \tau)} \tag{2.32}$$

où $\hat{S}_C(T_i)$ est la probabilité estimée par l'estimateur de Kaplan-Meier que le temps de censure soit supérieur à T_i. L'absence de terme de pondération dans la spécificité est due au fait que les termes s'annulent, chaque individu se voyant attribuer un poids égal à $\hat{S}_C(\tau)$.
Notons que ces mesures ne dépendent pas de l'incidence de l'événement jusqu'au temps τ. C'est une des raisons principales de leur popularité. On parle de mesure intrinsèque des capacités discriminantes, ces résultats étant indépendants de la fragilité de la population à l'événement.

2.3.2 Courbe ROC dépendante du temps

La courbe ROC dépendante du temps, notée ROC(τ), permet de résumer sur un graphique l'ensemble de l'information apportée par les sensibilités et spécificités calculées pour chaque valeur seuil v possible et un temps pronostique τ fixé. La courbe ROC dépendante du temps est définie par l'ensemble des points $\big\{(se(\tau,v),\ 1 - sp(\tau,v)),\ v \in \mathbb{R}\big\}$. Monotone croissante sur [0,1], la courbe ROC(τ) est située dans un rectangle [0,1] $\times$ [0,1], et passe par les points [0,0] et [1,1]. Lorsque le marqueur ne présente aucune capacité discriminante alors la courbe ROC(τ) est égale à la première bissectrice, au contraire si le marqueur discrimine parfaitement les individus qui subissent l'événement de ceux qui ne le subissent pas avant τ, alors la courbe passe par le coin supérieur gauche du rectangle.
L'aire sous la courbe ROC(τ), notée AUC(τ), (en anglais, *area under curve*, AUC), est un indicateur très souvent utilisé. Elle correspond à la probabilité qu'un individu ayant subi l'événement avant le temps τ ait une valeur du marqueur plus élevée qu'un individu n'ayant pas encore subi l'événement au temps τ. Formellement :

$$\text{AUC}(\tau) = P(X_m > X_n|\tilde{T_m} \leq \tau,\tilde{T_n} > \tau) \tag{2.33}$$

Un marqueur dépourvu de toute capacité pronostique sera associé à une AUC(τ) égale à 0,5. Au

contraire, un marqueur discriminant parfaitement les sujets qui vont subir l'événement avant τ de ceux qui ne vont pas subir l'événement est associé à une AUC(τ) égale à 1.

2.3.3 Approche pronostique et facteurs de confusion

La majorité des études pronostiques visant à évaluer les capacités discriminantes d'un marqueur ne prennent pas en compte les potentiels facteurs de confusion. Cela s'explique par la volonté d'évaluer la qualité discriminante du marqueur sans intention de distinguer les capacités propres au marqueur de celles pouvant résulter de tiers facteurs. Mais par exemple, si l'événement est le décès et si le marqueur augmente avec l'âge, alors une partie des capacités discriminantes attribuées au marqueur peuvent être dues à l'âge. Les travaux de Janes et Pepe (2008), repris et complétés par Pardo-Fernandez (2014), ont montré l'intérêt de prendre en compte les facteurs de confusion dans l'étude des capacités pronostiques d'un marqueur. Tout d'abord, ils démontrent que lorsqu'une covariable est associée au marqueur, la prendre en compte dans l'analyse peut modifier les résultats. Plus précisément, dans le cas où la covariable est indépendante de l'événement, les courbes ROC stratifiées estimées pour chaque modalité de la covariable sont toutes supérieures à la courbe ROC globale. La différence entre ces courbes reflète l'amélioration dans la qualité de prédiction pouvant être obtenue par l'utilisation d'une valeur seuil spécifique à chaque modalité de la covariable. Dans la cas où la covariable est aussi associée à l'événement, les courbes ROC stratifiées enlèvent de l'estimation des capacités pronostiques du marqueur celles liées à la covariable. Au final, ces courbes ROC stratifiées peuvent se situer au-dessus ou au-dessous de la courbe ROC globale, selon la force des associations covariable-marqueur et covariable-événement.

Les auteurs recommandent de prendre en compte les potentiels facteurs de confusion dans les études pronostiques visant à évaluer les capacités discriminantes d'un marqueur pour plusieurs raisons. Premièrement, pour identifier des sous-populations où les capacités du marqueur sont optimales ou au contraire identifier des sous-populations où le marqueur est peu susceptible d'être intéressant. Deuxièmement, pour définir des valeurs seuils indiquant la positivité du test différentes selon les strates des covariables afin d'améliorer la qualité de la prédiction. Par exemple, si la technique de mesure du marqueur est différente selon les centres et entraine un biais de mesure, alors la prédiction de l'événement à partir du marqueur peut être améliorée en définissant une valeur seuil propre à chaque centre. Troisièmement, prendre en compte les covariables dans l'analyse permet d'estimer le vrai potentiel discriminant intrinsèque au marqueur.

Une part croissante de la littérature s'intéresse à l'évaluation des capacités pronostiques en tenant compte des facteurs de confusion potentiels. Ce constat est en partie dû au développement des techniques récentes de la génomique et de la protéomique telles que le screening à haut débit où des milliers de gènes sont profilés en même temps. L'objectif est souvent d'identifier des biomarqueurs ayant de bonnes capacités pronostiques intrinsèques en évitant l'identification de faux positifs dus à des tiers facteurs. Yu et al. (2012) ont ainsi proposé une approche pour classer les biomarqueurs selon leur capacité pronostique ajustée sur les facteurs de confusion, l'inconvénient étant que les trois méthodes qu'ils proposent ne reposent sur aucun rationnel.

En l'absence de censure, Dodd et Pepe (2003) ont proposé des modèles de régression semi-paramétriques pour estimer les courbes ROC et l'AUC. Janes et Pepe (2009) ont proposé des courbes ROC ajustées sur les covariables basées sur la théorie des valeurs de placement (en anglais, *placement values*, PV) (voir section 2.3.4). En présence de données censurées, différents estimateurs semi-paramétriques de courbes ROC(τ) estimées pour chaque valeur des covariables (en anglais, *covariate-specific*) ont été proposé par Cai et al. (2006), Song et Zhou (2008) et Zheng et al. (2012). Des estimateurs non-paramétriques complémentaires ont été proposés par Rodríguez-Álvarez et al. (2015) alors que Zheng et al. (2010) ont

proposé des estimateurs semi-paramétriques des valeurs prédictives dépendantes du temps.

2.3.4 Estimateur basé sur la théorie des valeurs de placement

Le principe des PV est de considérer la courbe ROC comme une fonction de répartition du marqueur chez les cas, après que le marqueur ait été standardisé par rapport à sa distribution chez les témoins. Cet estimateur a été proposé par Pepe et Cai (2004) pour prendre en compte des covariables lors de l'estimation d'une courbe ROC en présence de données non censurées. Soit $F_{\tilde{T}_j>\tau}(x) = P(X \leq x|\tilde{T}_j > \tau)$ la fonction de répartition de X chez les témoins. La PV de X est :

$$u = 1 - F_{\tilde{T}_j>\tau}(v) \tag{2.34}$$

Elle représente la proportion de témoins ayant une valeur du marqueur supérieure à v. La courbe ROC est une représentation graphique de la sensibilité en fonction de 1 moins la spécificité, i.e. le tracé de $1 - F_{\tilde{T}_j\leq\tau}(v) = P(X > v|\tilde{T}_j \leq \tau)$ en fonction de u, où $F_{\tilde{T}_j\leq\tau}(v)$ est la fonction de répartition de X en v chez les cas. On a alors, $\text{ROC}(u) = 1 - F_{\tilde{T}_j\leq\tau}(v)$. Avec $v = F^{-1}_{\tilde{T}_j>\tau}(1-u)$ on obtient la relation suivante :

$$\text{ROC}(u) = 1 - F_{\tilde{T}_j\leq\tau}(F^{-1}_{\tilde{T}_j>\tau}(1-u)) \tag{2.35}$$

L'estimateur des courbes ROC ajustées basé sur la théorie des PV s'écrit (Janes et Pepe, 2009) :

$$\text{AROC}(u) = 1 - F_{\tilde{T}_j\leq\tau}(F^{-1}_{\tilde{T}_j>\tau}(1-u|Z_{\tilde{T}_j\leq\tau})) \tag{2.36}$$

où $Z_{\tilde{T}_j\leq\tau}$ sont les variables associées à la distribution de X chez les cas et $F_{\tilde{T}_j>\tau}(x|Z_{\tilde{T}_j\leq\tau})$ est égal à $P(X \leq x|\tilde{T}_j > \tau, Z_{\tilde{T}_j\leq\tau})$, c'est à dire la fonction de répartition chez les cas sachant les caractéristiques $Z_{\tilde{T}_j\leq\tau}$. La courbe ROC *covariate-specific* s'écrit $\text{ROC}(u|z) = 1 - F_{\tilde{T}_j\leq\tau}(F^{-1}_{\tilde{T}_j>\tau}(1-u|z)|z)$, il est intéressant de noter que la courbe ROC ajustée peut s'écrire comme une moyenne pondérée des courbes ROC *covariate-specific* :

$$\text{AROC}(u) = \int 1 - F_{\tilde{T}_j\leq\tau}(F^{-1}_{\tilde{T}_j>\tau}(1-u|Z_{\tilde{T}_j\leq\tau})|Z_{\tilde{T}_j\leq\tau} = z)\ dP(Z_{\tilde{T}_j\leq\tau} = z) \tag{2.37}$$

$$\text{AROC}(u) = \int \text{ROC}(u|z)\ dP(Z_{\tilde{T}_j\leq\tau} = z) \tag{2.38}$$

2.3.5 Problématique

L'estimateur basé sur les valeurs de placement permet d'obtenir une courbe ROC ajustée prenant en compte les facteurs de confusion potentiels. Il nécessite de connaître pour chaque individu son statut (i.e. cas ou témoins) et ne peut donc pas être appliqué en présence de données censurées. Nous avons vu dans la section 2.3.3 qu'en présence de données censurées des estimateurs permettant d'obtenir des courbes ROC(τ) *covariate-specific* ont été proposés mais aucun d'entre eux ne permet d'obtenir une courbe ROC(τ) globale en tenant compte de multiples facteurs de confusion catégoriels ou continus. Pour répondre à cette problématique, nous proposerons dans le chapitre 4 de nouveaux estimateurs basés sur une approche IPW, de sensibilités et spécificités dépendantes du temps.

Chapitre 3

Comparaisons des performances du modèle de Cox multiple par rapport aux méthodes basées sur le score de propension

Sommaire

3.1 Publication dans *Statistics in Medicine*

Résumé de l'article

Dans les études observationnelles, la présence de facteurs de confusion est fréquente et la comparaison de différents groupes de sujets nécessite alors un ajustement. En présence de données de survie, cet ajustement est dans la littérature majoritairement réalisé à l'aide d'un modèle de régression multiple (modèle de Cox le plus souvent). Cependant, les méthodes basées sur le score de propension sont de plus en plus utilisées. Dans cet article, nous comparons les performances, en termes de risques de première et de seconde espèces, du test de Wald associé au modèle de Cox multiple à celles des tests associés à quatre méthodes basées sur le score de propension (deux variantes du test du Log-rank ajusté, le modèle de Cox pondéré par le score de propension et l'appariement sur le score de propension). Nous proposons d'abord de comparer les performances des tests associés à ces méthodes par une étude de simulations, en se consacrant particulièrement sur les risques de première et seconde espèces (voir sous-section 2.2.7). Puis nous comparons ces méthodes à partir de données réelles extraites de la cohorte DIVAT.
Les résultats montrent que le test du Log-rank IPW proposé par Xie et Liu (2005) peut être utilisé en tant que première approche descriptive car il permet de présenter des courbes de survie ajustées et est associé à des erreurs de première et seconde espèces acceptables. Cependant, de meilleurs résultats sont obtenus avec le test de Wald associé au modèle de Cox multiple.
Dans ce travail, une confusion existe entre effet conditionnel et marginal. Nous discutons cette limite dans la section 3.3 qui suit l'article. Nous verrons que l'impact de cette confusion sur nos résultats doit être minime.

Special Issue Paper

Statistics in Medicine

Received 14 November 2014, Accepted 5 October 2015 Published online 29 October 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6777

Comparisons of the performance of different statistical tests for time-to-event analysis with confounding factors: practical illustrations in kidney transplantation

Florent Le Borgne,[a,b,c] Bruno Giraudeau,[d,e,f,g] Anne Héléne Querard,[a,c,h] Magali Giral[c] and Yohann Foucher[a,c*†]

Confounding factors are commonly encountered in observational studies. Several confounder-adjusted tests to compare survival between differently exposed subjects were proposed. However, only few studies have compared their performances regarding type I error rates, and no study exists evaluating their type II error rates. In this paper, we performed a comparative simulation study based on two different applications in kidney transplantation research. Our results showed that the propensity score-based inverse probability weighting (IPW) log-rank test proposed by Xie and Liu (2005) can be recommended as a first descriptive approach as it provides adjusted survival curves and has acceptable type I and II error rates. Even better performance was observed for the Wald test of the parameter corresponding to the exposure variable in a multivariable-adjusted Cox model. This last result is of primary interest regarding the exponentially increasing use of propensity score-based methods in the literature. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: survival data; adjusted Kaplan–Meier estimator; adjusted log-rank test; simulation study; inverse probability weighting; propensity score

1. Introduction

In randomized clinical trials, the log-rank test is generally suitable to compare time-to-event distributions. In contrast, for observational studies, subject characteristics may be imbalanced between the exposure groups of interest. Causality evaluations between the exposure and the time-to-event then require to take into account possible confounding factors. Often, a Wald test related to the regression coefficient of the exposure variable in a Cox proportional hazards (PH) model [1] is used to perform such a confounder-adjusted comparison. Alternative methods that were proposed for this purpose include stratification, matching and inverse probability weighting (IPW) [2–4] and are often based on propensity score theory, especially when there are several confounding factors. Austin and Schuster [5] compared the performance of the corresponding statistical tests regarding type I error rates. Major limitations of their simulation

[a]*SPHERE (EA 4275): bioStatistics, Pharmacoepidemiology & Human sciEnces REsearch, University of Nantes, Nantes, France*
[b]*IDBC/A2com, Espace Antrium Parc de la Teillais,35740 PACE, France*
[c]*Transplantation, Urology and Nephrology Institute (ITUN), Nantes Hospital and University, INSERM U1064, Nantes, France*
[d]*Centre de recherche Epidémiologie et Biostatistique, INSERM U1153, Paris, France*
[e]*Centre d'Investigation clinique INSERM , CIC 1415, Tours, France*
[f]*Université François Rabelais de Tours, PRES Centre-Val de Loire Université, Tours, France*
[g]*CHRU de Tours, Tours, France*
[h]*Médecine néphrologie - Hémodialyse, Centre Hospitalier Départemental Vendée Site de La Roche sur Yon, France*
[*]*Correspondence to: Yohann Foucher, SPHERE (EA 4275): bioStatistics, Pharmacoepidemiology & Human sciEnces REsearch, University of Nantes, Nantes, France.*
[†]*E-mail: Yohann.Foucher@univ-nantes.fr*

study were the absence of censoring and the lack of a type II error rate comparison. Conclusions based on type I error rates without counterbalancing with type II error rates may lead to choosing a test associated with a significant loss of statistical power, because statistical power decreases with an increasing censoring rate.

Regarding this incomplete literature, there is no consensus on the most appropriate confounder-adjusted statistical test. This may explain the high variability in practices. Should we always consider the usual *p*-values obtained from the Cox PH model as a reference regarding the increased use of other tests based on propensity scores? If not, which method should we choose? The principal aim of our paper is to clarify these questions.

More precisely, we wished to evaluate the performance in terms of type I and II error rates of the tests based on propensity score and on the Cox PH model. In the second section, we define the different confounder-adjusted statistical tests for the comparison of time-to-event distributions between two exposure groups. In the third and fourth section, we propose simulations and applications based on two examples in kidney transplantation. The complete R code [6] used to generate and analyse the simulated data and to analyse the real data of the two illustrations is available in the Supporting Information.

2. Methods

2.1. Confounder-adjusted log-rank statistic

Let $(T_i, \delta_i, X_i, Z_i)$ be the random variables associated with the subject i ($i = 1, \dots, n$). n is the sample size. T_i is the participating time. δ_i is the censoring indicator: $\delta_i = 0$ if T_i is a right censoring and $\delta_i = 1$ if T_i corresponds to the time-to-event. X_i is the explanatory variable representing the exposure factor of interest made up of K groups ($k = 1, \dots, K$). $Z_i = (Z_{1i}, \dots, Z_{pi})$ is the vector of covariates (p possible confounders). Let $p_{ik} = P(X_i = k|Z_i)$ be the probability for the subject i to be in group k depending on its characteristics Z_i. Let $S_k(t)$ be the survival function of group k at time t. Suppose D_k the number of different times for which events are observed in the group k. At time t_j ($j = 1, \dots, D_k$), the number of events d_{jk} and at risk subjects Y_{jk} can be defined as $d_{jk} = \sum_{i:t_i=t_j} \delta_i I(X_i = k)$ and $Y_{jk} = \sum_{i:t_i \geqslant t_j} I(X_i = k)$, respectively. The corresponding number of events and at risk subjects at time t_j in the whole sample can be noted $d_j = \sum_{k=1}^{K} d_{jk}$ and $Y_j = \sum_{k=1}^{K} Y_{jk}$. The IPW method makes it possible to reduce confounding bias when exposure groups are compared by correcting the contribution of each subject i by a weight $sw_{ik} = P(X_i = k)/p_{ik}$. This stabilized weight has been shown to produce a suitable estimate of the variance and maintains appropriate type I error rates even when there are subjects with extremely large weights [7, 8]. Then, the weighted number of events d_{jk}^w and at risk subjects Y_{jk}^w at time t_j are defined as $d_{jk}^w = \sum_{i:t_i=t_j} sw_{ik}\delta_i I(X_i = k)$ and $Y_{jk}^w = \sum_{i:t_i \geqslant t_j} sw_{ik} I(X_i = k)$. The corresponding weighted number of events and at risk subjects at time t_j in the whole sample can be noted $d_j^w = \sum_{k=1}^{K} d_{jk}^w$ and $Y_j^w = \sum_{k=1}^{K} Y_{jk}^w$. The resulting confounder-adjusted Kaplan–Meier estimator is simply the adaptation of the usual one as $\hat{S}_k(t) = \prod_{t_j \leqslant t} [1 - d_{jk}^w / Y_{jk}^w]$.

Liu *et al*. [9] proposed an alternative definition of the weights because p_{ik} close to 0 may result in unstable estimators. Their proposal is to use $\tilde{p}_{ik} = \max\{\epsilon \bar{p}_k, p_{ik}\}$ instead of p_{ik}, where $\bar{p}_k$ is the average of p_{ik} over all subjects i in the group k and $\epsilon \in (0, 1)$ is the adjustment threshold. They recommended choosing ϵ between 0.1 and 0.33. The weights are then equalled to $1/\tilde{p}_{ik}$.

By considering only two groups ($K = 2$), denoted $X = 0$ and $X = 1$, Xie and Liu [10] proposed a log-rank test by adjusting the weights of each subject over time. We give the acronym ALRT (Adjusted Log-Rank test with Time-dependent weights) to this test. At time t_j ($j = 1, \dots, D_k$), the weight for a subject i in the group k is reassigned as $sw'_{ijk} = sw_{ik} Y_{jk} / Y_{jk}^w$. The weighted number of events becomes $d_j^{w'} = \sum_{k=1}^{K} d_{jk}^{w'}$ with $d_{jk}^{w'} = \sum_{i:t_i=t_j} sw'_{ijk} \delta_i I(X_i = k)$, and the weighted number of at risk subjects becomes $Y_j^{w'} = \sum_{k=1}^{K} Y_{jk}^{w'}$ with $Y_{jk}^{w'} = \sum_{i:t_i \geqslant t_j} sw'_{ijk} I(X_i = k)$. The resulting statistic is $G^{w'} / \sqrt{\mathrm{Var}(G^{w'})}$, with

$$G^{w'} = \sum_{j=1}^{D} d_{j1}^{w'} - Y_{j1}^{w'} \frac{d_j^{w'}}{Y_j^{w'}} \tag{1}$$

$$\mathrm{Var}\left(G^{w'}\right) = \sum_{j=1}^{D}\left\{\frac{d_j(Y_j-d_j)}{Y_j(Y_j-1)}\sum_{i=1}^{Y_j}\left[\left(\frac{Y_{j0}^{w'}}{Y_j^{w'}}\right)^2 sw_{ij}^{\prime 2}X_i + \left(\frac{Y_{j1}^{w'}}{Y_j^{w'}}\right)^2 sw_{ij}^{\prime 2}(1-X_i)\right]\right\} \quad (2)$$

The statistic has a standard normal distribution for a large sample under the null hypothesis of survival equality in both exposure groups $\{H_0 : S_{X=0}(t) = S_{X=1}(t), \text{ for all } t > 0\}$. An alternative confounder-adjusted log-rank statistic was proposed by Sugihara [11], using the same formula as Xie and Liu except the re-assignment of weights. We give the acronym ALRB (Adjusted Log-Rank test with Baseline weights) to this test. The corresponding statistic is thus $G^w/\sqrt{\mathrm{Var}(G^w)}$ with

$$G^w = \sum_{j=1}^{D} d_{j1}^w - Y_{j1}^w\frac{d_j^w}{Y_j^w} \quad (3)$$

$$\mathrm{Var}\left(G^w\right) = \sum_{j=1}^{D}\left\{\frac{d_j\left(Y_j-d_j\right)}{Y_j\left(Y_j-1\right)}\sum_{i=1}^{Y_j}\left[\left(\frac{Y_{j0}^w}{Y_j^w}\right)^2 sw_i^2X_i + \left(\frac{Y_{j1}^w}{Y_j^w}\right)^2 sw_i^2\left(1-X_i\right)\right]\right\}. \quad (4)$$

2.2. Multivariable Cox model

To test the difference in survival between different exposure groups, we estimated the usual Cox PH model and the associated Wald test by using the robust covariance proposed by Lin and Wei [12]. This method has been shown to be robust to several types of misspecification of the Cox model, in particular to non-PH. The Wald statistic from this Multivariable Cox model (WMC) is $\beta_X/\sqrt{\mathrm{Var}(\beta_X)}$, which follows a standard normal distribution under the null hypothesis of survival equality in the two groups, i.e. $\{H_0 : \beta_X = 0\}$.

2.3. Weighted Cox model

Cole and Hernán [13] suggested using a univariate Cox PH regression model weighted by sw_{ik} in which only the variable corresponding to the exposure of interest is included. The corresponding weighted partial log-likelihood is $\sum_{i=1}^{n} sw_{ik}\delta_i\left\{\beta X_i - \log\left[\sum_{j:t_j\geqslant t_i} sw_{jk}Y_{jk}\exp(\beta X_j)\right]\right\}$ where β is the logarithm of the hazard ratio associated with X. The robust variance of Lin and Wei [12] is used to obtain a variance estimate that is valid under H_0, i.e. $\beta = 0$. We give the acronym WWC to the corresponding Wald test of the Weighted Cox model.

2.4. Caliper matching

Each exposed subject was matched with an unexposed subject on the logit of p_{ik}, using calipers of width equal to 0.2 standard deviations [14–16]. Two types of matching were used: 1:1 caliper matching in which each exposed subject was matched with one unexposed subject, and 1:3 incomplete caliper matching where each exposed subject was matched with up to three unexposed subjects when possible (and 1:2 or even 1:1 in case of impossible 1:3). We preferred an incomplete 1:3 matching to a strict 1:3 matching in order to increase the number of included subjects. In both cases, the matching was performed without replacement. In contrast with the previous weighted methods, subjects may therefore be removed from the analysis. The package 'MatchIt' was used to generate the pairs [17]. A Stratified Log-Rank test (SLR) on the matched pairs was used for testing statistical significance.

3. Illustration 1: expanded criteria donors in kidney transplantation

3.1. Context

Renal transplants are currently classified into two categories: the expanded criteria donor (ECD) and the standard criteria donor (SCD). ECD are defined by donors older than 60 or 50–59 years with two of the following characteristics: history of hypertension and cerebrovascular accident as the cause of death or terminal serum creatinine greater than 1.5 mg/dL [18]. Transplantations from ECD have been described with a substantial worse patient and graft survival than SCD [19–21]. We based our following application on 1912 adults receiving a first or second transplant from a deceased heart beating donor between January 1996 and December 2013 and followed in the prospective DIVAT (Données Informatisées et VAlidées

Table I. Baseline characteristics of ECD and SCD transplantations (related to the first illustration).

	All grafts n=1912		ECD n=729		SCD n=1183		p-value
	Mean	SD	Mean	SD	Mean	SD	
Recipient age (years)	50.68	14.17	61.42	9.61	44.07	12.36	<0.0001
Recipient BMI (kg.m^{-2})	23.89	6.28	24.97	4.23	23.22	7.18	<0.0001
Cold ischemia time (hours)	23.23	8.92	22.32	8.23	23.79	9.28	0.0003
	n	%	n	%	n	%	
Male recipient	1198	62.7	447	61.3	751	63.5	0.3416
Recurrent nephropathy	612	32.0	171	23.5	441	37.3	<0.0001
Dialysis prior transplantation	211	11.0	93	12.8	118	10.0	0.0602
History of diabetes	242	12.7	127	17.4	115	9.7	<0.0001
History of hypertension	1679	87.8	655	89.8	1024	86.6	0.0327
History of cardiovascular disease	851	44.5	396	54.3	455	38.5	<0.0001
History of dyslipemia	717	37.5	365	50.1	352	29.8	<0.0001
History of malignancy	195	10.2	120	16.5	75	6.3	<0.0001
History of B or C hepatitis	130	6.8	35	4.8	95	8.0	0.0064
Positive recipient CMV serology	941	49.4	403	55.4	538	45.6	<0.0001
Positive recipient EBV serology	1745	96.0	689	96.8	1056	95.5	0.1713
Positive anti-class I immunization	691	36.5	265	36.8	426	36.4	0.8800
Positive anti-class II immunization	705	37.8	229	32.0	476	41.4	<0.0001
Retransplantation	421	22.0	142	19.5	279	23.6	0.0354
Recipient blood group							0.0018
A	854	44.8	510	43.3	344	47.3	
B	181	9.5	126	10.7	55	7.6	
AB	80	4.2	62	5.3	18	2.5	
O	792	41.5	481	40.8	311	42.7	
Donor blood group							0.0043
A	835	43.8	496	42.1	339	46.5	
B	171	9.0	118	10.0	53	7.3	
AB	74	3.9	57	4.8	17	2.3	
O	827	43.4	507	43.0	320	43.9	
Male donor	1186	62.3	401	55.2	785	66.7	<0.0001
Positive donor CMV serology	794	41.6	347	47.7	447	37.8	<0.0001
Positive donor EBV serology	1480	92.8	637	95.4	843	91.0	0.0010
HLA-A-B-DR incompatibilities>4	349	18.3	145	19.9	204	17.3	0.1481
Depleting induction	764	40.0	256	35.2	508	42.9	0.0008

ECD, expanded criteria donor; SCD, standard criteria donor; BMI, body mass index; CMV, cytomegalovirus; EBV, epstein-barr virus; HLA, human leukocyte antigen.

en Transplantation, www.divat.fr) cohort from Nantes University hospital. The outcome was the time between transplantation and graft failure, defined as return to dialysis or death, whichever occurred earlier. The demographic and baseline characteristics at the time of transplantation are described in Table I. Among the 1912 kidney transplants, 729 (38.1%) were from ECD. Note that the use of kidneys from ECD is constantly increasing due to the ageing of both donor and recipient populations combined with the increasing proportion of deceased donors [22–24]. Thus, studying the increased risk of ECD versus SCD remains of heightened importance.

3.2. Simulations

3.2.1. Design. The simulated data were defined according to the observed data. One binary exposure variable (X) and four covariates (Z_1, Z_2, Z_3 and Z_4) were assumed, all respecting the PH assumption. The covariates and the exposure variable of interest were simulated successively by starting an unconditional draw of Z_1, then drawing Z_2 conditional on Z_1, Z_3 conditional on Z_1 and Z_2, Z_4 conditional on Z_1, Z_2 and Z_3 and finally X conditional on Z_1, Z_2, Z_3 and Z_4. The hierarchy of these successive simulations was decided according to clinical considerations, and the distributions were estimated from the real data set (details in Table II, first part). The intercept β_0 in the model used for drawing the exposure variable of interest X

Table II. Parameters, distributions and corresponding variables in the application of the simulated covariates.

	Simulated variables	Distributions	Parameters
ECD	Z_1: recipient age*	Trunc. normal	mean = 23.56; var = 16.96; min = 3.24
	Z_2: retransplantation	Bernoulli	$\text{logit}(p_i) = -0.83 - 0.02Z_{1i}$
	Z_3: PRA anti-class I	Bernoulli	$\text{logit}(p_i) = 0.54 - 0.01Z_{1i} + 2.59Z_{2i}$
	Z_4: HLA incomp. >4	Bernoulli	$\text{logit}(p_i) = -1.26 + 0.002Z_{1i} - 1.92Z_{2i} + 0.07Z_{3i}$
	X: ECD	Bernoulli	$\text{logit}(p_i) = \beta_0 + 0.15Z_{1i} + 0.25Z_{2i} + 0.13Z_{3i} + 0.005Z_{4i}$
	C_i: censoring times	Uniform	minimum = 0; maximum = $b^{\dagger}$
	E_i: times-to-event	Weibull	$F(t\|x,z) = 1 - \exp[-0.04\exp(\beta_1 x + 0.01z_1 + 0.18z_2 + 0.37z_3 + 0.19z_4)t^{0.76}]$
KWR	Z_1: recipient age	Trunc. normal	mean= 44.64; var= 14.17; min= 18
	Z_2: CIT $\geqslant$ 36 hours	Bernoulli	$\text{logit}(p_i) = -2.00 - 0.01Z_{1i}$
	Z_3: creat. $\geqslant$ 120 $\mu mol/L$	Bernoulli	$\text{logit}(p_i) = -1.92 + 0.01Z_{1i} + 0.38Z_{2i}$
	Z_4: retransplantation	Bernoulli	$\text{logit}(p_i) = -1.05 + -0.02Z_{1i} + 0.06Z_{2i} - 0.23Z_{3i}$
	X: KWR < 2.3 g/kg	Bernoulli	$\text{logit}(p_i) = \beta_0 - 0.01Z_{1i} - 0.29Z_{2i} + 0.50Z_{3i} + 0.27Z_{4i}$
	C_i: censoring times	Uniform	minimum = 0; maximum = 14.3
	E_i: times-to-event	Weibull	$F(t\|x,z) = 1 - \exp[-\exp(0.02z_1 + 0.30z_2 + 0.53z_3 + 0.39z_4)(0.0026t^{0.86}x)(0.0030t^{0.77}(1-x))]$

*Recipient age variable was transformed to respect the linearity assumption (recipient age squared divided by 100).
† b varies from 17.2 to 99.0 in order to obtained the adequate censoring rates.
First part: According to the application on expanded criteria donor. Second part: According to the application on the ratio between the weights of the kidney and the recipient.
CIT, cold ischemia time; creat, creatinine; HLA incomp., human leukocyte antigen incompatibilities; KWR, kidney weight ratio; PRA, panel-reactive antibody.

varied in order to study different proportions of exposed subjects (5%, 20% and 40%: observed value in the application). The times-to-event were generated from a PH model with baseline Weibull distribution. The regression coefficient associated with the exposure variable of interest (β_X) varied according to the different scenarios (0, 0.250, 0.365: value estimated in the application and 0.500). The censoring times were generated from a Uniform distribution: the minimum was 0, and the maximum time varied to obtain different censoring rates (0.25 and 0.68: observed in the application). More details are available in Table II (first part).

We compared the performance of the different models for various sample sizes (100, 250, 500 and 1500: approximately the size of the sample in the application). For each scenario, we simulated 40 000 samples. We calculated the percentage of rejection of H_0 when $\beta_X = 0$ (type I error rate) and the percentage of non-rejection of H_0 when $\beta_X \neq 0$ (type II error rate). The null hypothesis was rejected when the p-value was lower than 0.05. We considered that a number of 15 subjects in each group ($X = 1$ and $X = 0$) was a minimum size for using such models in practice. Thus, the scenario in which $n = 100$ with 5% of exposed subjects has been excluded.

3.2.2. Results. Table III and Supporting Information Table S1 describe the simulation results. We observed type I error rates close to 5% for the WMC in all scenarios as well as for both SLR methods (i.e. 1:1 and 1:3). This observation also held for ALRT and ALRB except when the percentage of exposed subjects was very low (5%). The WWC was associated with higher type I error rates, especially for small sample sizes and imbalanced percentages between exposed and unexposed subjects.

As expected, type II error rates decreased with decreasing censoring rates, increasing baseline sample sizes and increasing effect sizes. More importantly, the same type II error rates were obtained for the ALRT and ALRB, also equivalent to rates obtained by using the WMC when the sample sizes were lower than or equal to 250 subjects. In contrast, the type II error rates for the two confounder-adjusted log-rank tests were greater than those obtained from the WMC when the sample sizes were larger. The WWC was associated with lower type II error rates compared with ALRT and ALRB, consistent with the previous results in terms of type I error rates. Comparing the two SLR variants, 1:3 matching seems preferable especially when the percentage of exposed subjects is low. Compared with ALRT and ALRB,

Table III. Error rates (x100) obtained from data simulated according to the application related to the expanded criteria donor.

% of exposed				Censoring rate ≈ 0.68						Censoring rate ≈ 0.25					
subjects	β_X	n	% matched	WMC	ALRT	ALRB	WWC	SLR 1:1	SLR 1:3	WMC	ALRT	ALRB	WWC	SLR 1:1	SLR 1:3
		100	46	6.0	6.2	7.0	11.1	4.4	4.9	6.7	6.3	7.1	10.6	4.8	4.8
	(a)	250	50	5.4	5.9	6.3	8.8	5.3	5.2	5.6	5.8	6.4	8.2	4.7	4.9
	0.000	500	52	5.1	5.6	6.0	7.3	4.9	5.0	5.4	5.5	5.8	7.1	5.1	5.1
		1500	54	5.0	5.3	5.5	5.9	5.1	5.3	5.2	5.2	5.4	5.8	5.2	5.2
		100	46	90.1	88.9	88.1	83.6	94.3	92.6	84.7	83.6	82.8	77.3	92.8	90.7
	(b)	250	50	85.3	86.3	85.8	82.3	90.0	88.8	73.8	77.3	77.0	72.5	86.9	83.9
	0.250	500	52	75.3	81.5	81.2	78.3	85.1	81.6	54.7	68.3	68.3	64.2	77.2	70.7
40%		1500	54	40.6	63.9	63.9	60.2	62.9	53.8	11.2	40.2	40.5	39.3	41.6	29.4
		100	46	85.3	84.7	83.9	78.3	92.8	89.6	75.2	75.6	75.2	68.3	89.4	86.1
	(b)	250	50	73.0	78.6	78.3	73.3	84.6	81.1	51.6	62.8	62.8	57.0	77.9	70.8
	0.365	500	52	53.0	68.4	68.4	63.6	73.8	66.3	23.3	47.6	47.8	43.7	58.2	47.6
		1500	54	9.8	36.9	37.2	33.7	34.5	23.0	0.3	13.9	14.4	14.8	12.6	5.3
		100	46	77.3	78.0	77.5	70.6	89.9	85.3	60.0	64.1	63.9	55.7	84.5	78.8
	(b)	250	50	53.5	65.5	65.5	59.1	74.6	68.5	25.1	45.6	45.4	39.6	63.1	51.9
	0.500	500	52	25.1	48.7	49.0	43.3	55.4	44.3	3.9	26.5	26.9	24.4	33.9	21.5
		1500	54	0.5	13.0	13.3	12.1	9.9	4.1	0.0	2.7	3.1	3.8	1.3	0.2
		100	55	6.3	7.6	9.3	18.9	4.6	4.9	7.3	8.5	10.3	19.6	4.4	5.0
	(a)	250	64	5.2	6.8	8.0	15.1	4.4	5.1	6.0	7.7	9.1	15.5	5.2	5.0
	0.000	500	67	5.2	6.2	7.0	12.1	5.1	5.0	5.3	7.3	8.3	12.5	4.8	4.9
		1500	70	4.9	5.3	5.8	8.6	5.2	5.0	5.0	6.2	6.8	9.2	4.9	5.2
		100	55	90.0	86.7	84.6	77.0	94.4	93.3	85.5	82.0	80.4	68.8	94.3	92.0
	(b)	250	64	86.6	85.5	84.0	78.6	92.6	90.0	78.1	79.2	78.0	68.1	90.0	87.2
	0.250	500	67	79.4	84.0	82.8	79.1	88.5	85.2	63.3	75.0	73.6	65.2	84.1	78.1
20%		1500	70	51.3	78.6	77.8	75.4	74.5	65.0	21.3	65.6	64.4	57.5	60.3	45.2
		100	55	86.5	82.9	81.1	73.5	93.6	91.5	77.7	75.9	74.3	61.5	92.8	89.1
	(b)	250	64	77.0	80.0	78.6	72.7	89.0	84.6	60.2	70.2	68.7	56.9	84.0	78.2
	0.365	500	67	61.2	76.5	75.2	70.3	81.3	73.9	35.1	64.1	62.6	52.6	72.1	60.3
		1500	70	18.5	65.7	64.9	60.2	52.4	36.9	1.9	47.1	45.9	39.4	31.8	15.3
		100	55	79.7	77.4	75.6	67.3	91.9	88.2	65.8	67.2	65.7	51.3	90.1	83.8
	(b)	250	64	61.8	72.2	70.9	63.6	83.5	76.0	36.8	58.9	57.5	44.8	74.5	64.3
	0.500	500	67	36.3	66.0	65.1	58.4	68.7	56.7	10.0	49.1	47.6	37.5	53.6	36.2
		1500	70	2.2	47.2	46.9	40.0	25.4	11.6	0.0	28.4	27.7	23.6	8.7	1.7

Six approaches were compared: the multivariable Cox model (WMC), two confounder-adjusted log-rank tests proposed by Xie and Liu (ALRT) and by Sugihara (ALRB), the weighted Cox model (WWC) proposed by Cole and Hernán and two caliper matching-based stratified log rank tests (SLR 1:1 and 1:3). β_X is the regression coefficient associated with the variable of interest.
(a) Type I error rates.
(b) Type II error rates.
% of matched indicates the percentage of exposed subjects included in SLR 1:1 and SLR 1:3 analyses. The area under the ROC curve from the logistic model between the exposure and the confounders equaled 0.87 when 40% of the sample is exposed and 0.89 when 20% of the sample is exposed.

the SLR attained higher type II error rates for small samples but lower rates for large sample sizes. The percentages of exposed subjects included in the analysis vary from 46% to 80%. Note that 1:1 and 1:3 matching led to identical proportions of matched exposed subjects.

The same simulations were also carried out by correcting the weights as proposed by Liu *et al*. [9] with a threshold value ϵ equal to 0.1. The corresponding results are detailed in the Supporting Information Table S2. This correction improved the type I error rates of the WWC even if the rates remained higher than the ones obtained from the WMC or the confounder-adjusted log-rank tests proposed by Xie and Liu or by Sugihara. For the two confounder-adjusted log-rank tests, we observed a slight increase in the type I error rates and a slight decrease in type II error rates for large sample sizes. Note that we have also explored the impact of the hierarchy in the covariates simulations by reversing the order (from Z_4 to Z_1), the results were unchanged (data not shown).

3.3. Application

The unadjusted hazard ratio (HR) of ECD versus SCD equalled 1.97 (95% confidence interval (CI): 1.67–2.33, p <0.0001). Inspecting Schoenfeld residuals, the PH assumption seemed respected for all variables. For the sake of comparability, we used the same set of adjustment covariates for all compared methods. To select these covariates, univariate Cox models were first estimated. If p-values were lower than 0.20, then the variable was included in the WMC. A backward selection procedure was then carried out to keep only variables that were significant at the 5% level. If the exclusion of an insignificant variable changed the regression coefficient corresponding to the donor criteria by more than 20% that variable was not excluded.

The adjustment variables retained in the final WMC were the recipient age, the history of cardiovascular disease, the body mass index and the anti class I immunization. The adjusted HR was estimated at 1.52 for recipients of ECD versus SCD (95% CI: 1.26–1.84, p <0.0001). Figure 1 describes the corresponding crude and confounder-adjusted survival curves. As for the Cox model, the difference between confounder-adjusted curves was less significant than between unadjusted curves. The following p-values were obtained: 0.0237 for the ALRT, 0.0638 for the ALRB and 0.0742 for the WWC. For the 1:1 SLR, we have re-performed the generation of pairs four times. The following p-values were obtained: 0.3086, 0.0160, 0.0577 and 0.0097. We also repeated the 1:3 SLR four times, and we obtained the following p-values: 0.0086, 0.0031, 0.0306 and 0.0043. According to the p-value threshold at 0.05 for rejecting H_0,

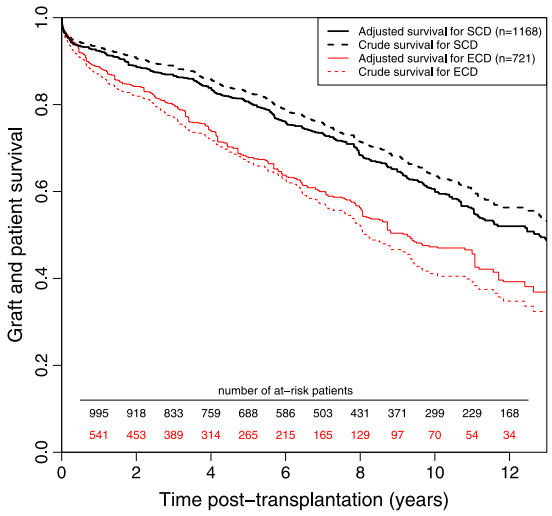


Figure 1. Patient and graft survival curves estimated by inverse probability weighting Kaplan–Meier estimator (continuous lines) and standard Kaplan–Meier estimator (dashed lines) for kidney transplant recipients according to the type of donor (standard criteria donor (SCD): dark lines, expanded criteria donor (ECD): light lines). The major steps observed on the confounder-adjusted curves appear when a subject with a high weight undergoes an event. To avoid major steps in confounder-adjusted survival curves, the correction of the weights proposed by Liu *et al*. [9] was used with $\epsilon = 0.33$.

the conclusions are obviously heterogeneous and discordant among the different tests, and also between the repetitions of the matching procedure. Only ALRT and 1:3 SLR were in accordance with the conclusion obtained from the WMC. Note that the two other variable selection strategies were performed: the LASSO penalization of the partial likelihood of the Cox model [25] and the consideration of all the covariates associated with the outcome as recommended in the literature related to propensity scores [4, 26, 27]. These two strategies included different numbers of covariates (6 and 12, respectively). The results obtained for each of these strategies are presented in the Supporting Information Table S3. The *p*-values obtained were all very close, except for the SLR test, for which the p-values partly differed substantially between the different strategies and also between the repetitions of the matching procedure.

4. Illustration 2: ratio of kidney weight to recipient weight

4.1. Context

We used the data from a previous analysis [28], which showed that a ratio of kidney weight to recipient weight (KWR) less than 2.3 g/kg is an independent risk factor for graft failure after 2 years post-transplantation (Figure 2). These data are especially useful to our study because the PH assumption does not hold. Among the 1060 patients transplanted between April 1995 and January 2006 included in the analysis, 223 (21,0%) had a KWR less than 2.3g/kg. The outcome was the time between transplantation and graft failure. The demographic characteristics according to the KWR levels are described in Table IV.

4.2. Simulations

4.2.1. Design. In section 3 (when PH assumption holds), the null hypothesis is defined in terms of identical survival curves ($\beta_X = 0$). By definition, when the hazards are not proportional, this definition is not adapted: the survival curves are different. This null hypothesis can be reformulated in terms of restricted mean survival time [29], i.e. $\{H_0 : E\{min(T, t^*)|X = 0\} = E\{min(T, t^*)|X = 1\}\}$, where t^* is the maximum follow-up time. Figure 3 illustrates two survival curves under H_0: the areas under the curves are equalled. Even if the methods we compared are not the most adapted in that context, it is interesting to identify the most robust one in this particular situation.

Four covariates (Z_1, Z_2, Z_3 and Z_4) and one binary exposure (X) were simulated successively by using the same principle as presented in Section 3.2.1. Distributions used for these simulations were principally estimated from the real data set as described in Table II (second part). The times-to-event were generated from a stratified PH model with different baseline Weibull distributions depending on KWR level. The

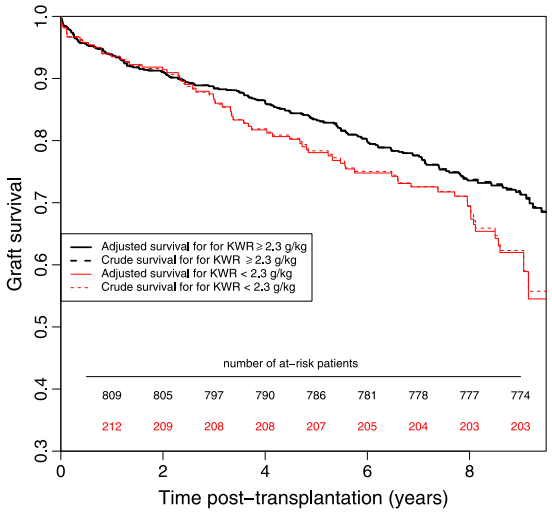


Figure 2. Graft survival curves estimated by inverse probability weighting Kaplan–Meier estimator (continuous lines) and standard Kaplan-Meier estimator (dashed lines) for kidney transplant recipients according to the kidney weight ratio (KWR; KWR $\geqslant 2.3g/kg$: dark lines, KWR $< 2.3g/kg$: light lines). The correction of the weights proposed by Liu *et al*. [9] was not performed.

Table IV. Demographic characteristics according to the kidney weight ratio level (related to the second illustration).

	All grafts n=1060		KWR $\geqslant 2.3g/kg$ n=837		KWR $< 2.3g/kg$ n=223		p-value
	Mean	SD	Mean	SD	Mean	SD	
Recipient age (years)	45.60	13.14	45.29	13.19	46.76	12.88	0.1329
Donor age (years)	39.78	15.29	40.79	14.41	36.01	17.76	0.0002
	n	%	n	%	n	%	
Male recipient	659	62.2	490	58.5	169	75.8	<0.0001
Male donor	719	68.0	608	72.8	111	49.8	<0.0001
Donor creatinine ($\geqslant$ 120 μmol/L)	214	20.6	181	22.0	33	15.3	0.0304
Cold ischemia time ($\geqslant$ 36 hours)	178	16.9	134	16.1	44	19.7	0.2019
DGF (>6 days)	296	28.5	240	29.3	56	25.6	0.2815
Retransplantation	122	11.5	101	12.1	21	9.4	0.2705
HLA-A-B-DR incompatibilities >4	131	12.6	100	12.2	31	14.1	0.4480
Positive anti-class I immunization	242	23.0	197	23.8	45	20.3	0.2722
Acute rejection >0	277	26.1	221	26.4	56	25.1	0.6964

DGF, delayed graft function; HLA, human leukocyte antigen.

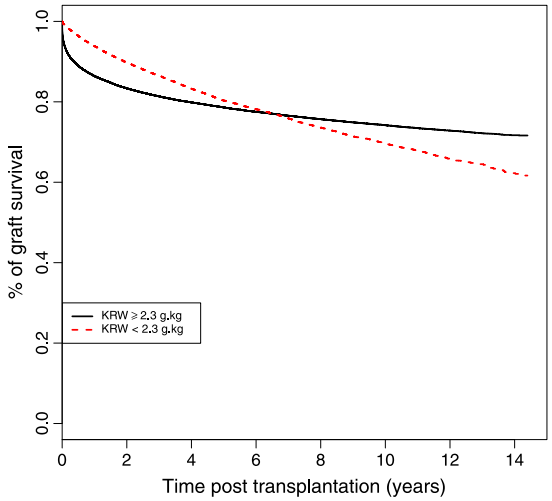


Figure 3. Graft survival curves based on the kidney weight ratio (KWR) application obtained after modification of the Weibull parameters in the group with a low KWR level in order to obtain two survival curves with identical restricted mean survival time. Areas under both survival curves were 11.0 years.

regression coefficients associated with the other covariates were identical regardless of the KWR level. To simulate two different restricted mean survival time (under $H_1 : E(T|T < t^*, X = 0) \neq E(T|T < t^*, X = 1)$), the parameters of the two baseline Weibull distributions were estimated from the application. Under H_0, the Weibull parameters in the group with a high KWR level were defined from the application, while the parameters in the other group were arbitrarily chosen to obtain identical areas under the two confounder-adjusted survival curves. The censoring times were generated from a Uniform distribution: the minimum was 0, and the maximum was 14.3. The coefficient β_0 of the logistic regression with the exposure as the outcome varied between 1.71 and 3.42 to keep a percentage of exposed subjects at 21.0%.

We compared the performance of the different models for various sample sizes at 100, 250, 500 and 1500, the same as in the previous simulations. In contrast, we did not change the censoring rate or the percentage of exposed subjects in order to avoid too many scenarios. As illustrated by Figure 2, the confounding in the relationship between the exposure variable of interest and the time-to-event was low, mainly due to the weak association between exposure and covariates (Table IV). To explore the performance of the different models when substantial confounding factors are present, we increased the

Table V. Percentages of rejection of the null hypothesis obtained from data simulated according to the application on kidney weight ratio after modification of the Weibull parameters in the group with a high kidney weight ratio level.

			Censoring rate ≈ 0.75					
Confounding strength	*n*	% matched	WMC	ALRT	ALRB	WWC	SLR 1:1	SLR 1:3
weak	100	94.4	5.3	4.1	4.1	5.7	5.3	3.6
	250	98.5	5.0	4.1	4.0	4.9	7.0	4.2
	500	99.4	4.9	4.1	4.1	4.8	8.8	5.0
	1500	99.9	5.0	4.2	4.1	4.9	17.3	8.6
strong	100	78.3	6.1	4.8	5.0	10.1	4.9	3.5
	250	85.3	5.1	5.7	5.9	9.3	5.5	4.2
	500	88.5	5.2	5.8	5.9	8.6	7.4	4.6
	1500	91.0	5.8	5.9	5.7	7.4	12.3	7.4

The first four lines correspond to the confounding strength as observed in the application, the second four lines to results obtained with an assumed increased confounding strength.
Six approaches were compared: the multivariable Cox model (WMC), two confounder-adjusted log-rank tests proposed by Xie and Liu (ALRT) and by Sugihara (ALRB), the weighted Cox model (WWC) proposed by Cole and Hernán and two caliper matching-based stratified log rank tests (SLR 1:1 and 1:3).
% matched indicates the percentage of exposed subjects included in SLR 1:1 and SLR 1:3 analyses.

perturbation in the relationship by multiplying by (i) the five regression coefficients related to the logistic model with the exposure as outcome and (ii) the two coefficients associated with the covariates in survival model used for generating times-to-event. For each scenario, we simulated 40 000 samples.

4.2.2. Results. The results of the simulations in terms of wrong H_0 rejection are presented in Table V. When confounding bias was low, we observed that H_0 was rejected in approximately 5% of the samples by using the WMC with robust sandwich covariance or the WWC, and a little less by using the two confounder-adjusted log-rank tests ($\approx$ 4%). In contrast, the percentages of H_0 rejections were larger with SLR, especially for large sample sizes. When the confounding bias in the relationship between the exposure and time-to-event was more significant, the WMC and the two confounder-adjusted log-rank tests remain close to 5%. The WWC rejected more H_0 (between 7% and 10%), in contrast the two SLR methods obtained lower rates but always higher than the ones obtained with the other methods for large sample sizes and 1:1 and 1:3 matching led to identical percentages of matched exposed subjects.

Table VI describes the percentages of non-rejection of H_0 when H_1 is true. When confounding bias was low, we observed very close results between the different approaches regardless of the sample size. One can notice an exception for 1:1 and 1:3 SLR associated with higher type II error rates. When confounding bias was more significant, the WMC obtained the lowest type II error rates. The WWC and the two confounder-adjusted log-rank tests were very close in terms of type II error rates but higher than those obtained with the WMC when the sample sizes were large. The two SLR were always associated with higher type II error rates. Note that the percentages of exposed subjects included in the analysis were high (from 78% to 100%), and the algorithm used to generate pairs led to identical percentages of matched exposed subjects for 1:1 and 1:3 SLR.

We have explored the impact of the hierarchy in the covariates simulations by reversing the order (from Z_4 to Z_1), the results were unchanged (data not shown).

4.3. Application

The WMC and the WWC were estimated despite the violation of the PH assumption given that we wish to compare the performance of the different models including the case when conditions are not respected.

Figure 2 describes the crude and confounder-adjusted survival curves according to KWR level. Confounder-adjusted and crude curves were very close, indicating that the perturbation in the relationship between KWR and graft survival was very low. Four adjustment variables were retained in the final WMC: the recipient age, the donor creatinemia, the retransplantation (yes or no) and the cold ischemia time. The following *p*-values were obtained: 0.0589 for the WMC, 0.0552 for the ALRT, 0.0546 for the

Table VI. Type II error rates (x100) obtained from data simulated according to the application on kidney weight ratio (exposure rate = 21%).

			Censoring rate ≈ 0.75					
Confounding strength	*n*	% matched	WMC	ALRT	ALRB	WWC	SLR 1:1	SLR 1:3
weak	100	94.4	94.2	94.9	94.7	93.2	94.6	94.5
	250	98.5	89.8	90.6	90.6	90.2	92.9	92.1
	500	99.4	81.5	82.8	82.8	83.1	90.9	87.6
	1500	99.9	50.2	52.5	52.5	53.3	82.4	70.5
strong	100	78.4	92.5	96.1	95.6	88.7	95.0	95.5
	250	85.3	88.8	94.1	93.6	84.7	93.9	93.9
	500	88.5	81.5	91.8	91.5	82.5	92.3	91.3
	1500	91.0	52.8	78.6	78.9	72.8	86.2	82.0

The first four lines correspond to the confounding strength as observed in the application, the second four lines to results obtained with an assumed increased confounding strength.
Six approaches were compared: the multivariable Cox model (WMC), two confounder-adjusted log-rank tests proposed by Xie and Liu (ALRT) and by Sugihara (ALRB), the weighted Cox model (WWC) proposed by Cole and Hernán and two caliper matching-based stratified log rank tests (SLR 1:1 and 1:3).
% matched indicates the percentage of exposed subjects included in SLR 1:1 and SLR 1:3 analyses.
The area under the ROC curve from the logistic model between the exposure and the confounders equaled 0.57 and 0.78 when confounding strength was assumed as 'weak' and 'strong', respectively.

ALRB and 0.0543 for the WWC. For the 1:1 SLR, we have re-performed the generation of pairs four times. The following *p*-values were obtained: 0.5961, 0.8348, 0.1037 and 0.2695. We also repeated the 1:3 SLR four times and obtained the following *p*-values 0.3328, 0.1420, 0.0391 and 0.1227. Except for the SLR, all these *p*-values obtained were very close, all at the significance limit. As in the first illustration, other variable selection strategies were performed (LASSO penalization and consideration of all the covariates associated with the outcome). These two strategies allowed different covariates to be considered (7 and 8 respectively). The results obtained for each of these strategies are presented in the Supporting Information Table S4. The *p*-values obtained were all relatively close, except for the SLR test, for which the *p*-values partly differed substantially between the different strategies and also between the repetitions of the matching procedure.

5. Discussion

In this paper, we compared the performance in terms of type I and II error rates of different tests: the WMC, two IPW adaptations of the log-rank test [10, 11], the WWC proposed by Cole and Hernán [13] and the SLR as proposed by Austin and Schuster [5]. In accordance with the results of Austin and Schuster, we did not include stratification on the propensity score, a method which resulted in the greatest bias as in the results obtained by Lunceford [30]. Moreover, as we only studied time-invariant exposure, we did not include the confounder-adjusted log-rank test proposed by Xu *et al.* [31], which is an extension for time-varying treatments. We also limited our study to the nearest neighbor caliper matching which showed better results compared with the other matching-based methods [5]. Finally, we did not include the test based on the average covariate method (i.e. estimating confounder-adjusted survival curves by applying the parameter estimates obtained from multivariable regression analysis to the average value of the covariates in the entire sample) because this method has been shown to be both mathematically and conceptually problematic and could lead to misleading survival curves [32]. Indeed, the survival curve estimated for an average individual may be different compared with the average survival curve estimated from a heterogeneous group of individuals [33, 34].

When the PH assumption holds, the results of our simulations indicated that the WMC had the best performance. The WWC should be used with caution with small sample sizes due to the violated type I error rates. Permutation procedures have been shown to preserve type I error rates in general settings [35, 36], and if applied to the WWC may avoid the violation of the type I error rate observed for this test. Among the two confounder-adjusted log-rank tests, it appeared the ALRT had slightly better type I error rate control. Nevertheless, in comparison with the WMC, the ALRT had (i) higher type I error rates

especially when there were few events observed or when the percentages of exposed and non-exposed subjects were imbalanced, (ii) higher type II error rates when the sample sizes were large. Finally, SLR seemed to respect the type I error rates, while type II error rates were higher than the ones obtained from the WMC. It also seems that 1:X incomplete matching should be preferred, especially when the percentages of exposed and non-exposed subjects were imbalanced. Compared with both confounder-adjusted log-rank tests (ALRT and ALRB), the type II error rates obtained with 1:3 SLR were higher for small sample sizes and lower for large sample sizes.

The superiority of the WMC in our results may be because of the simulated data respected the PH assumption. We therefore extended these comparisons to non-PH situations. Our previous conclusions in favour of the WMC remained valid. Under non-PH, we observed an important decrease in the statistical power for the SLR methods. Of course, in practice, Cox model has to be extended by time-varying coefficients when the PH assumption is in doubt.

In our applications, the compared tests led to different conclusions about rejecting or not rejecting the null hypothesis. This underlines the importance of our simulation study to guide the practitioner in choosing an appropriate method for confounder-adjusted comparisons of survival. In these two applications, we also compared different variable selection strategies. Globally, we obtained close results regardless the variable selection strategy used, except for SLR which was sensitive to the variable selection method and to the random number seed. Note that in the first application, we probably underestimated the outcome difference between ECD and SCD. Indeed, this difference was estimated using only ECD that were not discarded due to their poor quality [37].

Altogether, these simulation and application results tended to favour two complementary approaches: the WMC and the confounder-adjusted survival curves with the corresponding ALRT, each presenting advantages and disadvantages. The WMC is the most efficient but requires verification of assumptions before being carried out and summarizes the results in a single coefficient [38–40]. Confounder-adjusted survival curves make it possible to more precisely illustrate the differences in survivals between differently exposed groups, but the corresponding ALRT is not as efficient as the WMC.

In contrast, Austin and Schuster concluded that the best performance were obtained by the SLR and the WWC [5]. These apparent discordances can be explained. Firstly, regarding the caliper matching method, our results were in agreement with the results of Austin and Schuster: the SLR obtains the best type I error rates. However, we also highlighted that this method leads to a loss of statistical power for small sample sizes and when the PH assumption does not hold. These points were previously unexplored by Austin and Schuster. Secondly, the simulations performed by Austin and Schuster showed actual type I error rates of 5%, most likely because they only considered very large samples with 10 000 subjects.

In conclusion, the IPW confounder-adjusted survival curves and the corresponding confounder-adjusted log-rank test proposed by Xie and Liu [10] seemed to be useful as a first descriptive approach. The use of a multivariable Cox model should nevertheless be performed further to this in order to validate the results, especially if the ALRT is non significant. To facilitate its use in practice, we proposed an R package, entitled IPWsurvival, to plot confounder-adjusted survival curves and to calculate the ALRT (www.divat.fr, 2015 May 01).

Acknowledgements

We thank S. Le Floch for the collection of the data in the DIVAT cohort. We also thank the Roche Laboratory for its support to DIVAT cohort. Florent Le Borgne obtained a grant from IDBC for this work.

References

1. Cox DR. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society* 1972; **B**(24): 187–220.
2. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 2011; **46**(3):399–424.
3. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**(1):41–55.
4. Austin PC. A tutorial and case study in propensity score analysis: an application to estimating the effect of in-hospital smoking cessation counseling on mortality. *Multivariate Behavioral Research* 2011; **46**(1):119–151.
5. Austin PC, Schuster T. The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: a simulation study. *Statistical Methods in Medical Research* 2014; [Epub ahead of print].
6. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2013.

7. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**(5):550–560.
8. Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, Smith D. Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value in Health* 2010; **13**(2):273–277.
9. Liu C, Xie J, Zhang Y. Weighted nonparametric maximum likelihood estimate of a mixing distribution in nonrandomized clinical trials. *Statistics in Medicine* 2007; **26**(29):5303–5319.
10. Xie J, Liu C. Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine* 2005-10; **24**(20):3089–3110.
11. Sugihara M. Survival analysis using inverse probability of treatment weighted methods based on the generalized propensity score. *Pharmaceutical Statistics* 2010; **9**(1):21–34.
12. Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* 1989; **84**(408):1074–1078.
13. Cole SR, Hernán MA. Adjusted survival curves with inverse probability weights. *Computer Methods and Programs in Biomedicine* 2004; **75**(1):45–49.
14. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 1985; **39**(1):33.
15. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics* 2011; **10**(2):150–161.
16. Wang Y, Cai H, Li C, Jiang Z, Wang L, Song J, Xia J. Optimal caliper width for propensity score matching of three treatment groups: A Monte Carlo study. *PLoS ONE* 2013; **8**(12):e81045.
17. Ho DE, Imai K, King G, Stuart EA. MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software* 2011; **42**(8):1–28.
18. Port FK, Bragg-Gresham JL, Metzger RA, Dykstra DM, Gillespie BW, Young EW, Delmonico FL, Wynn JJ, Merion RM, Wolfe RA, Held PJ. Donor characteristics associated with reduced graft survival: an approach to expanding the pool of kidney donors. *Transplantation* 2002; **74**(9):1281–1286.
19. Praehauser C, Hirt-Minkowski P, Saydam Bakar K, Amico P, Vogler E, Schaub S, Mayr M. Risk factors and outcome of expanded-criteria donor kidney transplants in patients with low immunological risk. *Swiss Medical Weekly Wkly* 2013; **143**:w13883.
20. Singh SK, Kim SJ. Does expanded criteria donor status modify the outcomes of kidney transplantation from donors after cardiac death? *American Journal of Transplantation* 2013; **13**(2):329–336.
21. Sung RS, Guidinger MK, Leichtman AB, Lake C, Metzger RA, Port FK, Merion RM. Impact of the expanded criteria donor allocation system on candidates for and recipients of expanded criteria donor kidneys. *Transplantation* 2007; **84**(9): 1138–1144.
22. Stratta RJ, Rohr MS, Sundberg AK, Armstrong G, Hairston G, Hartmann E, Farney AC, Roskopf J, Iskandar SS, Adams PL. Increased kidney transplantation utilizing expanded criteria deceased organ donors with results comparable to standard criteria donor transplant. *Annals of Surgery* 2004; **239**(5):688–697.
23. Roels L, Rahmel A. The European experience. *Transplant International* 2011; **24**(4):350–367.
24. De Fijter JW. An old virtue to improve senior programs. *Transplant International* 2009; **22**(3):259–268.
25. Tibshirani R. The lasso method for variable selection in the Cox model. *Statistics in Medicine* 1997; **16**(4):385–395.
26. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable delection for propensity score models. *American Journal of Epidemiology* 2006; **163**(12):1149–1156.
27. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine* 2007; **26**(4): 734–753.
28. Giral M, Foucher Y, Karam G, Labrune Y, Kessler M, Hurault de Ligny B, Büchler M, Bayle F, Meyer C, Trehet N, Daguin P, Renaudin K, Moreau A, Soulillou JP. Kidney and recipient weight incompatibility reduces long-term graft survival. *Journal of the American Society of Nephrology* 2010; **21**(6):1022–1029.
29. Andersen PK, Perme MP. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research* 2010; **19**(1):71–99.
30. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 2004; **23**(19):2937–2960.
31. Xu S, Shetterly S, Powers D, Raebel MA, Thomas Tsai MD, Ho PM, Magid D. Extension of Kaplan-Meier methods in observational studies with time-varying treatment. *Value in Health* 2012; **15**(1):167–174.
32. Ghali WA, Quan H, Brant R, van Melle G, Norris CM, Faris PD, Galbraith PD, Knudtson ML, APPROACH (Alberta Provincial Project for Outcome Assessment in Coronary Heart Disease) Investigators. Comparison of 2 methods for calculating adjusted survival curves from proportional hazards models. *JAMA* 2001; **286**(12):1494–1497.
33. Nieto FJ, Coresh J. Adjusting survival curves for confounders: a review and a new 'method. *American Journal of Epidemiology* 1996; **143**(10):1059–1068.
34. Thomsen BL, Keiding N, Altman DG. A note on the calculation of expected survival, illustrated by the survival of liver transplant patients. *Statistics in Medicine* 1991; **10**(5):733–738.
35. Simon R, Simon NR. Using randomization tests to preserve type I error with response-adaptive and covariate-adaptive randomization. *Statistics & probability letters* 2011; **81**(7):767–772.
36. Berger VW. Pros and cons of permutation tests in clinical trials. *Statistics in Medicine* 2000; **19**(10):1319–1328.
37. Pan Q, Schaubel DE. Proportional hazards models based on biased samples and estimated selection probabilities. *Canadian Journal of Statistics* 2008; **36**(1):111–127.
38. Spruance SL, Reid JE, Grace M, Samore M. Hazard ratio in clinical trials. *Antimicrobial Agents and Chemotherapy* 2004; **48**(8):2787–2792.

39. Combescure C, Perneger TV, Weber DC, Daurès JP, Foucher Y. Prognostic ROC curves: a method for representing the overall discriminative capacity of binary markers with right-censored time-to-event endpoints. *Epidemiology* 2014; **25**(1): 103–109.
40. Hernán MA. The Hazards of Hazard Ratios. *Epidemiology* 2010; **21**(1):13–15.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.

3.2 Annexe publiée sur le WEB uniquement

Research Article

Statistics in Medicine

Received XXXX

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

Web-based Supplementary Materials for: "Comparisons of the performance of different statistical tests for time-to-event analysis with confounding factors: Practical illustrations in kidney transplantation."

Florent Le Borgne[1,2,3], Bruno Giraudeau[4,5,6,7], Anne Hélène Querard[1,3,8], Magali Giral[3] and Yohann Foucher[1,3]*

This Web appendix contains supplementary simulation results according to the first illustration related to the ECD and the complete R code used to generate and analyse the simulated data and to analyse the real data of the two illustrations.

[1] *SPHERE (EA 4275): Biostatistics, Clinical Research and Subjective Measures in Health Sciences, University of Nantes, France*
[2] *IDBC/A2com, Espace Antrium Parc de la Teillais 35740 PACE, France*
[3] *Transplantation, Urology and Nephrology Institute (ITUN), Nantes Hospital and University, INSERM U1064, Nantes, France*
[4] *Centre de recherche Epidémiologie et Biostatistique, INSERM U1153, Paris, France*
[5] *Centre d'Investigation clinique (CIC), INSERM 1415, Tours, France*
[6] *Université François-Rabelais de Tours, PRES Centre-Val de Loire Université, Tours, France*
[7] *CHRU de Tours, Tours, France*
[8] *Médecine néphrologie - Hémodialyse, Centre Hospitalier Départemental Vendée Site de La Roche sur Yon, France*
* *Correspondence to: E-mail: Yohann.Foucher@univ-nantes.fr*

Prepared using simauth.cls *[Version: 2010/03/10 v3.00]*

Statistics
in Medicine

A. Supplementary simulation results related to the first illustration

% of exposed subjects	β_X	n	% matched	Censoring rate ≈ 0.68						Censoring rate ≈ 0.25					
				WMC	ALRT	ALRB	WWC	SLR 1:1	SLR 1:3	WMC	ALRT	ALRB	WWC	SLR 1:1	SLR 1:3
	(a)	250	61	6.3	9.5	12.7	32.5	3.6	4.4	6.3	9.5	12.7	32.5	3.6	4.4
	0.000	500	71	5.3	8.5	11.1	27.2	4.4	5.2	5.3	8.5	11.1	27.2	4.4	5.2
		1500	80	5.1	7.6	9.6	19.5	5.2	4.9	5.1	7.6	9.6	19.5	5.2	4.9
	(b)	250	61	88.3	84.5	81.3	64.8	95.6	93.5	88.3	84.5	81.3	64.8	95.6	93.5
	0.250	500	71	86.7	85.2	82.1	69.5	94.3	91.2	86.7	85.2	82.1	69.5	94.3	91.2
5%		1500	80	74.4	84.2	82.0	74.8	88.9	84.1	74.4	84.2	82.0	74.8	88.9	84.1
	(b)	250	61	83.7	81.3	78.1	62.4	94.8	91.8	83.7	81.3	78.1	62.4	94.8	91.8
	0.365	500	71	77.9	80.5	77.7	65.9	92.4	87.4	77.9	80.5	77.7	65.9	92.4	87.4
		1500	80	52.2	78.7	76.4	69.5	81.9	72.2	52.2	78.7	76.4	69.5	81.9	72.2
	(b)	250	61	76.1	76.6	73.0	58.3	93.4	89.0	76.1	76.6	73.0	58.3	93.4	89.0
	0.500	500	71	64.3	75.4	72.4	60.9	89.8	82.0	64.3	75.4	72.4	60.9	89.8	82.0
		1500	80	24.7	70.7	68.3	60.7	69.6	52.6	24.7	70.7	68.3	60.7	69.6	52.6

Table S1. Error rates (x100) obtained from data simulated according to the application related to the ECD for 5% of exposed subjects. Six approaches were compared: the multivariable Cox model (WMC), two confounder-adjusted log-rank tests proposed by Xie and Liu (ALRT) and by Sugihara (ALRB), the weighted Cox model (WWC) proposed by Cole and Hernán and two caliper matching-based stratified log rank tests (SLR 1:1 and 1:3). β_X is the regression coefficient associated with the variable of interest. (a) Type I error rates. (b) Type II error rates. % matched indicates the percentage of exposed subjects included in SLR 1:1 and SLR 1:3 analyses. The area under the ROC curve from the logistic model between the exposure and the covariates equaled 0.92.

% of exposed subjects	β_X	n	% matched	Censoring rate ≈ 0.68						Censoring rate ≈ 0.25					
				WMC	ALRT	ALRB	WWC	SLR 1:1	SLR 1:3	WMC	ALRT	ALRB	WWC	SLR 1:1	SLR 1:3
		100	46	6.0	5.8	6.5	10.0	4.5	4.8	7.0	6.1	6.8	9.7	4.7	5.1
	(a)	250	50	5.3	5.7	6.1	8.2	5.2	5.3	5.5	5.8	6.1	7.6	4.9	4.8
	0.000	500	52	5.3	5.4	5.7	6.8	5.0	5.0	5.6	5.8	6.0	6.8	5.1	5.0
		1500	54	5.1	5.7	5.8	6.0	5.1	5.2	5.2	5.5	5.6	5.8	5.3	5.5
		100	46	90.0	88.9	88.3	84.2	94.4	92.6	84.5	83.7	83.1	78.2	92.9	91.0
	(b)	250	50	84.9	85.8	85.5	82.2	89.9	88.5	73.7	76.5	76.3	72.3	87.2	83.6
	0.250	500	52	75.3	80.1	80.2	76.7	85.0	81.3	55.3	65.4	65.7	62.8	76.8	71.3
40%		1500	54	41.1	57.5	57.7	54.9	63.0	54.1	11.5	30.6	31.2	31.2	41.9	29.6
		100	46	85.2	84.7	84.0	79.2	92.8	89.9	74.7	74.9	74.7	68.7	89.1	86.0
	(b)	250	50	73.6	77.4	77.4	72.8	84.7	80.9	51.7	61.0	61.1	56.2	78.2	71.4
	0.365	500	52	52.8	64.4	64.7	59.9	73.5	66.5	23.1	41.8	42.5	40.4	58.4	47.2
		1500	54	9.7	27.9	28.2	27.1	34.6	23.2	0.3	7.2	7.5	8.2	12.8	5.5
		100	46	76.8	77.5	77.1	70.6	90.1	85.1	60.2	63.5	63.1	55.8	84.5	79.1
	(b)	250	50	53.7	63.3	63.3	57.8	74.9	68.7	25.4	41.3	41.5	37.7	63.2	52.0
	0.500	500	52	25.2	42.9	43.3	38.8	55.2	44.0	4.0	18.6	19.5	19.2	34.2	21.5
		1500	54	0.4	7.3	7.5	8.4	10.0	4.2	0.0	1.0	1.1	1.6	1.3	0.2
		100	55	6.2	7.4	8.5	16.5	4.5	4.6	7.3	8.2	9.5	17.1	4.3	5.1
	(a)	250	64	5.3	6.3	6.9	12.2	4.3	4.9	5.9	7.6	8.4	13.1	5.1	5.0
	0.000	500	67	5.3	5.7	6.1	9.5	5.0	4.9	5.4	6.8	7.2	10.2	4.7	5.1
		1500	70	5.1	5.4	5.5	6.9	5.2	5.1	5.1	6.8	7.1	8.4	4.9	4.9
		100	55	90.3	86.8	85.6	78.7	94.6	93.0	85.4	81.9	80.8	70.7	94.4	91.7
	(b)	250	64	86.9	85.5	84.8	80.8	92.7	90.4	78.1	78.6	77.6	70.0	89.9	87.0
	0.250	500	67	79.4	82.7	82.2	79.5	88.6	85.1	63.8	72.9	71.9	65.8	84.5	78.5
20%		1500	70	50.9	69.9	69.6	66.9	74.1	65.0	21.6	50.2	50.0	48.0	60.7	45.4
		100	55	86.3	82.8	81.5	74.5	93.6	91.7	77.7	75.7	74.6	63.2	92.8	88.8
	(b)	250	64	77.6	80.1	79.2	74.4	89.4	85.1	60.0	68.9	67.6	58.3	84.2	78.3
	0.365	500	67	61.3	73.8	73.2	69.2	81.2	73.9	34.6	58.4	57.3	50.9	72.3	60.2
		1500	70	18.6	48.5	48.4	44.1	52.3	37.1	2.0	24.7	25.1	25.0	31.5	15.2
		100	55	79.8	77.6	76.5	68.6	92.0	88.3	66.0	67.4	66.1	53.4	90.1	84.0
	(b)	250	64	62.1	71.0	70.2	64.0	83.5	76.1	36.1	55.6	54.1	44.3	74.5	64.5
	0.500	500	67	36.0	60.2	60.0	54.2	68.8	56.5	10.2	40.0	39.1	34.0	53.8	36.3
		1500	70	2.1	23.6	23.7	19.2	25.3	11.6	0.0	6.4	7.3	8.1	8.8	1.9
	(a)	250	61	6.1	9.1	11.3	30.4	3.8	4.4	8.1	10.7	12.8	31.7	4.7	5.1
	0.000	500	71	5.3	8.3	9.9	23.6	4.4	5.0	6.4	9.8	11.3	24.0	4.7	5.3
		1500	80	5.2	6.7	7.5	13.9	5.3	5.0	5.5	8.4	9.4	15.8	4.8	4.8
	(b)	250	61	88.6	84.5	82.6	66.8	95.6	93.2	82.9	80.5	78.7	58.5	94.5	92.0
	0.250	500	71	86.5	85.1	83.3	71.7	94.2	91.3	79.0	79.5	77.7	62.8	92.6	89.4
5%		1500	80	73.7	83.5	82.3	77.5	88.7	83.7	57.0	75.8	73.7	64.5	85.0	77.0
	(b)	250	61	83.9	81.0	79.1	63.7	94.7	91.8	74.6	75.1	73.1	52.5	93.2	89.8
	0.365	500	71	77.8	80.6	78.8	67.6	92.3	87.5	64.8	73.5	71.3	55.1	89.8	84.0
		1500	80	52.6	77.1	75.9	70.7	81.8	72.4	27.1	66.0	63.4	53.5	74.3	59.0
	(b)	250	61	76.0	76.5	74.3	59.9	93.8	89.1	61.8	68.7	66.6	45.2	91.4	85.5
	0.500	500	71	64.0	74.7	72.8	61.5	89.5	81.7	43.6	65.2	62.9	46.5	84.4	74.3
		1500	80	25.1	67.1	65.6	59.5	69.5	53.3	5.9	53.4	50.2	41.0	57.0	34.7

Table S2. Error rates (x100) obtained from data simulated according to the application related to the ECD by correcting the weights as proposed by Liu et al. (2007). Six approaches were compared: the multivariable Cox model (WMC), two confounder-adjusted log-rank tests proposed by Xie and Liu (ALRT) and by Sugihara (ALRB), the weighted Cox model (WWC) proposed by Cole and Hernán and two caliper matching-based stratified log rank tests (SLR 1:1 and 1:3). β_X is the regression coefficient associated with the variable of interest. (a) Type I error rates. (b) Type II error rates. % matched indicates the percentage of exposed subjects included in SLR 1:1 and SLR 1:3 analyses.

Statistics
in Medicine

B. Supplementary application results with different variable selection strategies

	Strategy 1	Strategy 2	Strategy 3
	p-value	*p*-value	*p*-value
WMC	<0.0001	0.0001	0.0001
ALRT	0.0237	0.0253	0.0298
ALRB	0.0638	0.0655	0.0579
WWC	0.0742	0.0749	0.0717
SLR 1:1	0.3086	0.3889	0.0066
SLR 1:1	0.0160	0.0137	0.0032
SLR 1:1	0.0577	0.1573	0.0118
SLR 1:1	0.0097	0.0152	0.0119
SLR 1:3	0.0086	0.0001	0.0012
SLR 1:3	0.0031	0.0019	0.0045
SLR 1:3	0.0306	0.0092	0.0038
SLR 1:3	0.0043	0.0188	0.0074

Table S3. *P*-values obtained for the illustration on ECD from the different tests and the different set of covariates according to the variable selection strategy. Strategy 1: the one initially used in the paper, strategy 2: the LASSO penalized Cox model, strategy 3: the inclusion of the covariates associated with the outcome. ALRB, Adjusted Log-Rank test proposed by Sugihara; ALRT, Adjusted Log-Rank test proposed by Xie and Liu; SLR, Stratified Log-Rank test related to the caliper matching; WMC, Wald test related to the Multivariable Cox model; WWC, Wald test related to the Weighted Cox model.

	Strategy 1	Strategy 2	Strategy 3
	p-value	*p*-value	*p*-value
WMC	0.0589	0.0218	0.0384
ALRT	0.0552	0.0501	0.0337
ALRB	0.0546	0.0514	0.0340
WWC	0.0543	0.0470	0.0354
SLR 1:1	0.5961	0.0527	0.1797
SLR 1:1	0.8348	0.5023	0.5785
SLR 1:1	0.1037	0.3454	0.2328
SLR 1:1	0.2695	0.1851	0.2636
SLR 1:3	0.3388	0.0042	0.0136
SLR 1:3	0.1420	0.0057	0.1187
SLR 1:3	0.0391	0.0239	0.1328
SLR 1:3	0.1227	0.0528	0.0562

Table S4. *P*-values obtained for the illustration on KWR from the different tests and the different set of covariates according to the variable selection strategy. Strategy 1: the one initially used in the paper, strategy 2: the LASSO penalized Cox model, strategy 3: the inclusion of the covariates associated with the outcome. ALRB, Adjusted Log-Rank test proposed by Sugihara; ALRT, Adjusted Log-Rank test proposed by Xie and Liu; SLR, Stratified Log-Rank test related to the caliper matching; WMC, Wald test related to the Multivariable Cox model; WWC, Wald test related to the Weighted Cox model.

C. The R code

```
#----------------------------
# R code used for simulations
#----------------------------

#####################
# The functions used #
#####################

crosssum<-function(seq,point,value)
{
        crosssum<-numeric(length(seq))
        for (i in 1:length(point)) {
                loc<-sum(seq<=point[i])
                crosssum[loc]<-crosssum[loc]+value[i]
        }
        crosssum
}

ALRT<-function(time,delta,treat,wgt,adjust=NULL)
{
        if (!is.null(adjust)) {
                m1<-mean(wgt[treat==1]); m0<-mean(wgt[treat==0])
                mwgt<-ifelse(treat==1,adjust*m1,adjust*m0)
                wgt<-ifelse(wgt>mwgt,mwgt,wgt)
        }
        n.unit<-rep(1,length(time))
        d.time<-survfit(Surv(time,n.unit)~1)$time
        n.risk0<-crosssum(d.time,time[treat==0],n.unit[treat==0])
        n.risk1<-crosssum(d.time,time[treat==1],n.unit[treat==1])
        n.risk1<-rev(cumsum(rev(n.risk1)))
        n.risk<-n.risk0+n.risk1
        n.event<-crosssum(d.time,time[delta==1],delta[delta==1])
        mod.rate<-ifelse(n.risk==1,0,n.event*(n.risk-n.event)/(n.risk*(n.risk-1)))
        mod.rate[is.na(mod.rate)] <- 0

        w.risk0<-crosssum(d.time,time[treat==0],wgt[treat==0])
        w.risk1<-crosssum(d.time,time[treat==1],wgt[treat==1])
        w.risk0<-rev(cumsum(rev(w.risk0)))
        w.risk1<-rev(cumsum(rev(w.risk1)))
        w.risk<-w.risk0+w.risk1
        w.risk0[w.risk0==0]<-0.0001
        w.risk1[w.risk1==0]<-0.0001

        subset0<-(delta==1)*(treat==0)
        w.event0<-crosssum(d.time,time[subset0==1],wgt[subset0==1])
        subset1<-(delta==1)*(treat==1)
        w.event1<-crosssum(d.time,time[subset1==1],wgt[subset1==1])
        w.event0<-w.event0*n.risk0/w.risk0;
        w.event1<-w.event1*n.risk1/w.risk1;
        w.event<-w.event0+w.event1

        ww.risk0<-crosssum(d.time,time[treat==0],(wgt*wgt)[treat==0])
        ww.risk1<-crosssum(d.time,time[treat==1],(wgt*wgt)[treat==1])
        ww.risk0<-rev(cumsum(rev(ww.risk0)))
        ww.risk1<-rev(cumsum(rev(ww.risk1)))
        ww.risk0<-ww.risk0*n.risk0*n.risk0/(w.risk0*w.risk0)
        ww.risk1<-ww.risk1*n.risk1*n.risk1/(w.risk1*w.risk1)

        log.mean<-sum(w.event1-n.risk1*w.event/n.risk)
        log.var<-sum(mod.rate*(n.risk0*n.risk0*ww.risk1+n.risk1*n.risk1*ww.risk0)/(n.risk*n.risk))
        v.akm<-log.mean/sqrt(log.var)
        return(v.akm)
}

ALRB <- function(times,failures,variable,weights=NULL){
        if(is.null(weights)){ .w <- rep(1, length(times)) } else { .w <- weights }
        .data <- data.frame(t=times, f=failures, v=variable, w=.w)
```

```
        .data <- .data[!is.na(.data$v),]
        .tj <- sort(unique(.data$t[.data$f==1]))
        .dj <- sapply(.tj,function(x){sum(.data$t[.data$f==1]==x)})
        .nj <- sapply(.tj,function(x){sum(.data$t>=x)})

        .dj0w <- sapply(.tj, function(x){sum(.data$w[.data$t==x & .data$f==1 & .data$v==0])})
    .dj1w <- sapply(.tj, function(x){sum(.data$w[.data$t==x & .data$f==1 & .data$v==1])})
        .djw <- .dj1w + .dj0w

        .nj0w2 <- sapply(.tj, function(x){sum(.data$w[.data$t>=x & .data$v==0]^2)})
        .nj1w2 <- sapply(.tj, function(x){sum(.data$w[.data$t>=x & .data$v==1]^2)})
        .nj0w <- sapply(.tj, function(x){sum(.data$w[.data$t>=x & .data$v==0])})
        .nj1w <- sapply(.tj, function(x){sum(.data$w[.data$t>=x & .data$v==1])})
        .njw <- .nj0w + .nj1w

        .ej1w <- .nj1w*.djw/.njw

        .vjw <- ((.dj*(.nj-.dj))/(.nj*(.nj-1))) *
        (((.nj0w/.njw)^2)*.nj1w2   + ((.nj1w/.njw)^2)*.nj0w2)

        .stat_test <- (sum(.dj1w)-sum(.ej1w))/sqrt(sum(.vjw))
        .pvalue <- 2*(1-pnorm(abs(.stat_test)))

        return(data.frame(stat_test=.stat_test, pvalue=.pvalue))
}

KM <- function(times,failures,variable,weights=NULL)
{
        if(is.null(weights)){ .w <- rep(1, length(times)) }     else { .w <- weights }
        .data <- data.frame(t=times, f=failures, v=variable, w=.w)
        .data <- .data[!is.na(.data$v),]
        .table <- data.frame(times=NULL, n.risk=NULL, n.event=NULL, survival=NULL, variable=NULL)
        for(i in  unique(variable)){
        .d <- .data[.data$v==i,]
        .tj <- c(0,sort(unique(.d$t[.d$f==1])))
        .dj <- sapply(.tj, function(x){sum(.d$w[.d$t==x & .d$f==1])})
        .nj <- sapply(.tj, function(x){sum(.d$w[.d$t>=x])})
        .st <- cumprod((.nj-.dj)/.nj)
        .table <- rbind(.table,data.frame(times=.tj, n.risk=.nj, n.event=.dj, survival=.st, variable=i))
        }
        return(.table)
}



fct_simul_ECD <- function(b.ecd, cens_min, cens_max){
        N <- j # sample size
        ## age
        ageR1 <- rtnorm(N, mean=param[param$Label=="ageR1.mu",2], sd=param[param$Label=="ageR1.sigma",2], lower
            =18^2/100, upper=Inf)
        ## retransplantation
        proba.z1 <- exp(param[param$Label=="retransplant.beta0",2] + param[param$Label=="retransplant.beta1",2]
             * ageR1)/(1+exp(param[param$Label=="retransplant.beta0",2] + param[param$Label=="retransplant.
            beta1",2] * ageR1))
        retransplant <- sapply(proba.z1, FUN=function(x) { rbinom(1, 1, prob=x) })
        ## PRA
        proba.z1 <- exp(param[param$Label=="PRA.beta0",2] + param[param$Label=="PRA.beta1",2] * ageR1 + param[
            param$Label=="PRA.beta2",2] * retransplant)/(1+exp(param[param$Label=="PRA.beta0",2] + param[param$
            Label=="PRA.beta1",2] * ageR1 + param[param$Label=="PRA.beta2",2] * retransplant))
        PRA <- sapply(proba.z1, FUN=function(x) { rbinom(1, 1, prob=x) })
        ## HLA
        proba.z1 <- exp(param[param$Label=="HLA.beta0",2] + param[param$Label=="HLA.beta1",2] * ageR1 + param[
            param$Label=="HLA.beta2",2] * retransplant + param[param$Label=="HLA.beta3",2] * PRA)/(1+exp(param[
            param$Label=="HLA.beta0",2] + param[param$Label=="HLA.beta1",2] * ageR1 + param[param$Label=="HLA.
            beta2",2] * retransplant + param[param$Label=="HLA.beta3",2] * PRA))
        HLA <- sapply(proba.z1, FUN=function(x) { rbinom(1, 1, prob=x) })
        ## ECD
        proba.z1 <- exp(b.ecd + param[param$Label=="ECD.beta1",2] * ageR1 + param[param$Label=="ECD.beta2",2] *
             retransplant + param[param$Label=="ECD.beta3",2] * PRA + param[param$Label=="ECD.beta4",2] * HLA)/
            (1+exp(b.ecd + param[param$Label=="ECD.beta1",2] * ageR1 + param[param$Label=="ECD.beta2",2] *
            retransplant + param[param$Label=="ECD.beta3",2] * PRA + param[param$Label=="ECD.beta4",2] * HLA))
```



```
        ECD <- sapply(proba.z1, FUN=function(x) { rbinom(1, 1, prob=x) })

        RepW.inverse<-function(t, cova, coef.reg, sigma, nu)
        { ((-1/sigma)*(exp(as.matrix(cova) %*% as.vector(-1*coef.reg)))*log(1-t))^(1/nu) }
        u <- runif(N, min=0, max=1)
        tps <- RepW.inverse(u, cbind(ECD, ageR1, HLA, PRA, retransplant), c(l, param[4:7,2]), param[1,2], param
             [2,2])
        cens <- runif(N, min=cens_min, max=cens_max)
        evt <- 1*(tps<cens)
        tps <- pmin(tps, cens)
        tps <- round(tps*100,2)

        D <- data.frame(tps=tps, ECD=ECD, ageR1=ageR1, retransplant=retransplant, PRA=PRA, HLA=HLA, evt=evt)
        D[D$tps==max(D$tps),7]<-0


        Pr0 <- glm(ECD~1, family = binomial(link=logit), data=D)$fitted.values[1]
        Pr1 <- glm(ECD~ageR1 + retransplant + PRA + HLA, family = binomial(link=logit), data=D)$fitted.values
        #IPW
        W <- (D$ECD==1) * Pr0/Pr1 + (D$ECD==0) * (1-Pr0)/(1-Pr1)

        fct_matchit <- function(.ratio){
        .matchit <- matchit(ECD~ageR1 + retransplant + PRA + HLA, data = D, method = "nearest", replace=FALSE,
             ratio=.ratio, caliper=0.2)
        D.app <- D
        D.app$app <- NULL
        D.app[row.names(.matchit$match.matrix)[!is.na(.matchit$match.matrix[,1])],"app"] <- row.names(.matchit$
             match.matrix)[!is.na(.matchit$match.matrix[,1])]
        sapply(c(1:dim(.matchit$match.matrix)[1]), FUN=function(x){
        D.app[.matchit$match.matrix[x,][!is.na(.matchit$match.matrix[x,])],"app"] <<- row.names(.matchit$match.
             matrix)[x]
        })
        D.app <- D.app[!is.na(D.app$app),]
        if(.ratio == 1){res <- c(NA,NA)
        }else{res <- c(NA,NA,NA)}
        try(res[1] <- 1 - pchisq(survdiff(Surv(D.app$tps, D.app$evt)~D.app$ECD + strata(D.app$app))$chisq, 1))
        res[2] <-  sum(D.app$ECD) / sum(D$ECD) * 100
        if(.ratio > 1){res[3] <- dim(D.app[D.app$ECD==0,])[1] / sum(D.app$ECD) }
        return(res)
        }

        res <- rep(NA,9)
        if(sum(ECD)>0){
                res[1:4] <- c(
                summary(coxph(Surv(tps,evt)~ECD + ageR1 + retransplant + PRA + HLA, robust=TRUE, data=D))$
                     coefficients[1,6],
                2*(1-pnorm(abs(ALRT(D$tps, D$evt, D$ECD, W)))),
                as.numeric(ALRB(D$tps, D$evt, D$ECD, W)[2]),
                summary(coxph(Surv(tps,evt)~ECD, data=D, robust=TRUE, weights=W))$coefficients[1,6]
                )
                ### matching ###
                try(res[5:6] <- fct_matchit(1))
                try(res[7:9] <- fct_matchit(3))
        }
        return(res)
}



fct_simul_KwRw <- function(forcees=FALSE){
        N <- j # sample size
        ## age
        .age <- rtnorm(N, mean=param[param$Label=="age.mu",2], sd=param[param$Label=="age.sigma",2], lower=18,
             upper=Inf)
        ## ischemia
        proba.z1 <- exp(param[param$Label=="isc2cl.beta0",2] + param[param$Label=="isc2cl.beta_age",2] * .age)/
             (1+exp(param[param$Label=="isc2cl.beta0",2] + param[param$Label=="isc2cl.beta_age",2] * .age))
        .isc2cl <- sapply(proba.z1, FUN=function(x) { rbinom(1, 1, prob=x) })
        ## creatininemia
```

```
proba.z1 <- exp(param[param$Label=="creat.beta0",2] + param[param$Label=="creat.beta_age",2] * .age +
    param[param$Label=="creat.beta_isc2cl",2] * .isc2cl)/(1+exp(param[param$Label=="creat.beta0",2] +
    param[param$Label=="creat.beta_age",2] * .age + param[param$Label=="creat.beta_isc2cl",2] * .isc2cl
    ))
.creat2cl <- sapply(proba.z1, FUN=function(x) { rbinom(1, 1, prob=x) })
## retransplantation
proba.z1 <- exp(param[param$Label=="retransplant.beta0",2] + param[param$Label=="retransplant.beta_age"
    ,2] * .age + param[param$Label=="retransplant.beta_isc2cl",2] * .isc2cl + param[param$Label=="
    retransplant.beta_creat2cl",2] * .creat2cl)/(1+exp(param[param$Label=="retransplant.beta0",2] +
    param[param$Label=="retransplant.beta_age",2] * .age + param[param$Label=="retransplant.beta_isc2cl
    ",2] * .isc2cl + param[param$Label=="retransplant.beta_creat2cl",2] * .creat2cl))
.retransplant <- sapply(proba.z1, FUN=function(x) { rbinom(1, 1, prob=x) })
## KwRw
if(forcees==TRUE){m <- 5; .b0 <- 2}else{m <- 1; .b0 <- 1}
proba.z1 <- exp(param[param$Label=="ratioP.beta0",2] * .b0 + param[param$Label=="ratioP.beta_age",2] *
    m * .age + param[param$Label=="ratioP.beta_isc2cl",2] * m * .isc2cl + param[param$Label=="ratioP.
    beta_creat2cl",2] * m * .creat2cl + param[param$Label=="ratioP.beta_retransplant",2] * m * .
    retransplant)/(1+exp(param[param$Label=="ratioP.beta0",2] * .b0 + param[param$Label=="ratioP.beta_
    age",2] * m * .age + param[param$Label=="ratioP.beta_isc2cl",2] * m * .isc2cl + param[param$Label==
    "ratioP.beta_creat2cl",2] * m * .creat2cl + param[param$Label=="ratioP.beta_retransplant",2] * m *
    .retransplant))
.ratioP <- sapply(proba.z1, FUN=function(x) { rbinom(1, 1, prob=x) })
.ratioP <- 1*(.ratioP==0)

RepW.inverse<-function(t, cova, coef.reg, sigma, nu)
{ ((-1/sigma)*(exp(as.matrix(cova) %*% as.vector(-1*coef.reg)))*log(1-t))^(1/nu) }
u0 <- runif(length(.ratioP[.ratioP==0]), min=0, max=1)
u1 <- runif(length(.ratioP[.ratioP==1]), min=0, max=1)
if(forcees==TRUE){m <- 0.35; .b0 <- 2}else{m <- 1; .b0 <- 1}
tps0 <- RepW.inverse(u0, cbind(.age[.ratioP==0], .isc2cl[.ratioP==0], .creat2cl[.ratioP==0], .
    retransplant[.ratioP==0]),  c(param[param$Label=="beta_age",2],param[param$Label=="beta_isc2cl",2],
    param[param$Label=="beta_creat2cl",2],param[param$Label=="beta_retransplant",2]) * .b0, param[1,2]*
    m, param[2,2])
tps1 <- RepW.inverse(u1, cbind(.age[.ratioP==1], .isc2cl[.ratioP==1], .creat2cl[.ratioP==1], .
    retransplant[.ratioP==1]),  c(param[param$Label=="beta_age",2],param[param$Label=="beta_isc2cl",2],
    param[param$Label=="beta_creat2cl",2],param[param$Label=="beta_retransplant",2]) * .b0, param[3,2]*
    m, param[4,2])

D <- data.frame(cbind(.age,.isc2cl,.creat2cl,.retransplant, .ratioP))
D <- D[order(D[,5],decreasing=FALSE),]
D$tps <- c(tps0, tps1)
D$cens  <- runif(N, min=param[param$Label=="censure.min",2], max=param[param$Label=="censure.max",2]*
    1.2)
D$evt <- 1*(D$tps<D$cens)
D$Temps <- pmin(D$tps, D$cens)
D$Temps <- round(D$Temps,2)
D[D$Temps==max(D$Temps),"evt"] <- 0

Pr0 <- glm(.ratioP~1, family = binomial(link=logit), data=D)$fitted.values[1]
Pr1 <- glm(.ratioP~.age + .isc2cl + .creat2cl + .retransplant , family = binomial(link=logit), data=D)$
    fitted.values
W <- (D$.ratioP==1) * Pr0/Pr1 + (D$.ratioP==0) * (1-Pr0)/(1-Pr1)

fct_matchit <- function(.ratio){
.matchit <- matchit(.ratioP~.age + .isc2cl + .creat2cl + .retransplant, data = D, method = "nearest",
    replace=FALSE, ratio=.ratio, caliper=0.2)
D.app <- D
D.app$app <- NULL
D.app[row.names(.matchit$match.matrix)[!is.na(.matchit$match.matrix[,1])],"app"] <- row.names(.matchit$
    match.matrix)[!is.na(.matchit$match.matrix[,1])]
sapply(c(1:dim(.matchit$match.matrix)[1]), FUN=function(x){
D.app[.matchit$match.matrix[x,][!is.na(.matchit$match.matrix[x,])],"app"] <<- row.names(.matchit$match.
    matrix)[x]
})
D.app <- D.app[!is.na(D.app$app),]
if(.ratio == 1){res <- c(NA,NA)
}else{res <- c(NA,NA,NA)}
try(res[1] <- 1 - pchisq(survdiff(Surv(D.app$Temps, D.app$evt)~D.app$.ratioP + strata(D.app$app))$chisq
    , 1))
res[2] <-  sum(D.app$.ratioP) / sum(D$.ratioP) * 100
if(.ratio > 1){res[3] <- dim(D.app[D.app$.ratioP==0,])[1] / sum(D.app$.ratioP) }
```



```
        return(res)
        }

        res <- rep(NA,9)
        res[1:4] <- c(
        summary(coxph(Surv(Temps,evt)~.ratioP + .age + .isc2cl + .creat2cl + .retransplant, robust=TRUE, data=D
            ))$coefficients[1,6],
        2*(1-pnorm(abs(ALRT(D$Temps, D$evt, D$.ratioP, W)))),
        as.numeric(ALRB(D$Temps, D$evt, D$.ratioP, W)[2]),
        summary(coxph(Surv(Temps,evt)~.ratioP, data=D, robust=TRUE, weights=W))$coefficients[1,6])
        ### matching ###
        try(res[5:6] <- fct_matchit(1))
        try(res[7:9] <- fct_matchit(3))

        return(res)
}


fct_simul_KwRw_hyp_nulle <- function(forcees=FALSE){
        N <- j # sample size
        ## age
        .age <- rtnorm(N, mean=param[param$Label=="age.mu",2], sd=param[param$Label=="age.sigma",2], lower=18,
            upper=Inf)
        ## ischemia
        proba.z1 <- exp(param[param$Label=="isc2cl.beta0",2] + param[param$Label=="isc2cl.beta_age",2] * .age)/
            (1+exp(param[param$Label=="isc2cl.beta0",2] + param[param$Label=="isc2cl.beta_age",2] * .age))
        .isc2cl <- sapply(proba.z1, FUN=function(x) { rbinom(1, 1, prob=x) })
        ## creatininemia
        proba.z1 <- exp(param[param$Label=="creat.beta0",2] + param[param$Label=="creat.beta_age",2] * .age +
            param[param$Label=="creat.beta_isc2cl",2] * .isc2cl)/(1+exp(param[param$Label=="creat.beta0",2] +
            param[param$Label=="creat.beta_age",2] * .age + param[param$Label=="creat.beta_isc2cl",2] * .isc2cl
            ))
        .creat2cl <- sapply(proba.z1, FUN=function(x) { rbinom(1, 1, prob=x) })
        ## retransplantation
        proba.z1 <- exp(param[param$Label=="retransplant.beta0",2] + param[param$Label=="retransplant.beta_age"
            ,2] * .age + param[param$Label=="retransplant.beta_isc2cl",2] * .isc2cl + param[param$Label=="
            retransplant.beta_creat2cl",2] * .creat2cl)/(1+exp(param[param$Label=="retransplant.beta0",2] +
            param[param$Label=="retransplant.beta_age",2] * .age + param[param$Label=="retransplant.beta_isc2cl
            ",2] * .isc2cl + param[param$Label=="retransplant.beta_creat2cl",2] * .creat2cl))
        .retransplant <- sapply(proba.z1, FUN=function(x) { rbinom(1, 1, prob=x) })
        ## KwRw
        if(forcees==TRUE){m <- 5; .b0 <- 2}else{m <- 1; .b0 <- 1}
        proba.z1 <- exp(param[param$Label=="ratioP.beta0",2] * .b0 + param[param$Label=="ratioP.beta_age",2] *
            m * .age + param[param$Label=="ratioP.beta_isc2cl",2] * m * .isc2cl + param[param$Label=="ratioP.
            beta_creat2cl",2] * m * .creat2cl + param[param$Label=="ratioP.beta_retransplant",2] * m * .
            retransplant)/(1+exp(param[param$Label=="ratioP.beta0",2] * .b0 + param[param$Label=="ratioP.beta_
            age",2] * m * .age + param[param$Label=="ratioP.beta_isc2cl",2] * m * .isc2cl + param[param$Label==
            "ratioP.beta_creat2cl",2] * m * .creat2cl + param[param$Label=="ratioP.beta_retransplant",2] * m *
            .retransplant))
        .ratioP <- sapply(proba.z1, FUN=function(x) { rbinom(1, 1, prob=x) })
        .ratioP <- 1*(.ratioP==0)

        RepW.inverse<-function(t, cova, coef.reg, sigma, nu)
        { ((-1/sigma)*(exp(as.matrix(cova) %*% as.vector(-1*coef.reg)))*log(1-t))^(1/nu) }
        u0 <- runif(length(.ratioP[.ratioP==0]), min=0, max=1)
        u1 <- runif(length(.ratioP[.ratioP==1]), min=0, max=1)
        if(forcees==TRUE){m <- 0.35; .b0 <- 2}else{m <- 1; .b0 <- 1}
        tps0 <- RepW.inverse(u0, cbind(.age[.ratioP==0], .isc2cl[.ratioP==0], .creat2cl[.ratioP==0], .
            retransplant[.ratioP==0]),  c(param[param$Label=="beta_age",2],param[param$Label=="beta_isc2cl",2],
            param[param$Label=="beta_creat2cl",2],param[param$Label=="beta_retransplant",2]) * .b0, 0.021 * m,
            0.32)
        tps1 <- RepW.inverse(u1, cbind(.age[.ratioP==1], .isc2cl[.ratioP==1], .creat2cl[.ratioP==1], .
            retransplant[.ratioP==1]),  c(param[param$Label=="beta_age",2],param[param$Label=="beta_isc2cl",2],
            param[param$Label=="beta_creat2cl",2],param[param$Label=="beta_retransplant",2]) * .b0, param[3,2]
            * m, param[4,2])

        D <- data.frame(cbind(.age,.isc2cl,.creat2cl,.retransplant, .ratioP))
        D <- D[order(D[,5],decreasing=FALSE),]
        D$tps <- c(tps0, tps1)
        D$cens  <- runif(N, min=param[param$Label=="censure.min",2], max=param[param$Label=="censure.max",2]*
            1.2)
```

```
        D$evt <- 1*(D$tps<D$cens)
        D$Temps <- pmin(D$tps, D$cens)
        D$Temps <- round(D$Temps,2)
        D[D$Temps==max(D$Temps),"evt"] <- 0

        Pr0 <- glm(.ratioP~1, family = binomial(link=logit), data=D)$fitted.values[1]
        Pr1 <- glm(.ratioP~.age + .isc2cl + .creat2cl + .retransplant , family = binomial(link=logit), data=D)$
             fitted.values
        W <- (D$.ratioP==1) * Pr0/Pr1 + (D$.ratioP==0) * (1-Pr0)/(1-Pr1)

        fct_matchit <- function(.ratio){
        .matchit <- matchit(.ratioP~.age + .isc2cl + .creat2cl + .retransplant, data = D, method = "nearest",
             replace=FALSE, ratio=.ratio, caliper=0.2)
        D.app <- D
        D.app$app <- NULL
        D.app[row.names(.matchit$match.matrix)[!is.na(.matchit$match.matrix[,1])],"app"] <- row.names(.matchit$
             match.matrix)[!is.na(.matchit$match.matrix[,1])]
        sapply(c(1:dim(.matchit$match.matrix)[1]), FUN=function(x){
        D.app[.matchit$match.matrix[x,][!is.na(.matchit$match.matrix[x,])],"app"] <<- row.names(.matchit$match.
             matrix)[x]
        })
        D.app <- D.app[!is.na(D.app$app),]
        if(.ratio == 1){res <- c(NA,NA)
        }else{res <- c(NA,NA,NA)}
        try(res[1] <- 1 - pchisq(survdiff(Surv(D.app$Temps, D.app$evt)~D.app$.ratioP + strata(D.app$app))$chisq
             , 1))
        res[2] <-  sum(D.app$.ratioP) / sum(D$.ratioP) * 100
        if(.ratio > 1){res[3] <- dim(D.app[D.app$.ratioP==0,])[1] / sum(D.app$.ratioP) }
        return(res)
        }

        res <- rep(NA,9)
        res[1:4] <- c(
        summary(coxph(Surv(Temps,evt)~.ratioP + .age + .isc2cl + .creat2cl + .retransplant, robust=TRUE, data=D
             ))$coefficients[1,6],
        2*(1-pnorm(abs(ALRT(D$Temps, D$evt, D$.ratioP, W)))),
        as.numeric(ALRB(D$Temps, D$evt, D$.ratioP, W)[2]),
        summary(coxph(Surv(Temps,evt)~.ratioP, data=D, robust=TRUE, weights=W))$coefficients[1,6])
        ### matching ###
        try(res[5:6] <- fct_matchit(1))
        try(res[7:9] <- fct_matchit(3))
        return(res)
}

################################################
# Simulations related to the first illustration #
################################################

# import parameters estimated from the real data set
param <- read.csv("param.csv",fill=TRUE,header=TRUE, sep = ";",dec = ".")

B <- 40000 # number of simulated samples
for(.b.ecd in c(-8.8, -6.4, -4.7)){ # percentage of exposed subjects
        for (l in c(0, 0.25, 0.365, 0.5)){ # beta_x
                for (j in c(250, 500, 1500)) { # sample sizes
                        inter <- matrix(NA, ncol=9, nrow=B)
                        inter <- foreach(i=1:B, .combine=rbind, .packages=c('msm','MatchIt','survival')) %dopar
                             % {
                        fct_simul_ECD(b.ecd=.b.ecd, cens_min=param[param$Label=="censure.min",2], cens_max=99);
                        }
                        assign(paste("res_",l,"_",j,sep=""),inter)
                        f = get(paste("res_",l,"_",j,sep=""))
                        assign(paste("res_",l,"_",j,sep=""), f)
                        write.table(get(paste("res_",l,"_",j,sep="")), file = paste("res_",l,"_",j,".csv",sep="
                             "), append = FALSE, quote = TRUE, sep = ";",eol = "\n", na = "NA", dec = ",", row.
                             names = FALSE, col.names = TRUE, qmethod = c("escape", "double"), fileEncoding = ""
                             )
                }
        }
}
```



```
#######################################################
# Simulations related to the second illustration (H1) #
#######################################################

# import parameters estimated from the real data set
param <- read.csv("param.csv",fill=TRUE,header=TRUE, sep = ";",dec = ".")

B <- 40000 # number of simulated samples
for (j in c(100,250,500,1500)) {# sample sizes
        inter <- matrix(NA, ncol=9, nrow=B)
        inter <- foreach(i=1:B, .combine=rbind, .packages=c('msm','MatchIt','survival')) %dopar% {fct_simul_
            KwRw(forcees=FALSE);}#set forcees=TRUE to simulate data with an increase confounding bias
        assign(paste("res_",j,sep=""),inter)
        f = get(paste("res_",j,sep=""))
        assign(paste("res_",j,sep=""), f)
        write.table(get(paste("res_",j,sep="")), file = paste("res_",j,".csv",sep=""), append = FALSE, quote =
            TRUE, sep = ";",eol = "\n", na = "NA", dec = ",", row.names = FALSE,col.names = TRUE, qmethod = c("
            escape", "double"),fileEncoding = "")
}


#######################################################
# Simulations related to the second illustration (H0) #
#######################################################

# import parameters estimated from the real data set
param <- read.csv("param.csv",fill=TRUE,header=TRUE, sep = ";",dec = ".")

B <- 40000 # number of simulated samples
for (j in c(100,250,500,1500)) {# sample sizes
        inter <- matrix(NA, ncol=9, nrow=B)
        inter <- foreach(i=1:B, .combine=rbind, .packages=c('msm','MatchIt','survival')) %dopar% {fct_simul_
            KwRw_hyp_nulle(forcees=FALSE);}#set forcees=TRUE to simulate data with an increase confounding bias
        assign(paste("res_",j,sep=""),inter)
        f = get(paste("res_",j,sep=""))
        assign(paste("res_",j,sep=""), f)
        write.table(get(paste("res_",j,sep="")), file = paste("res_",j,".csv",sep=""), append = FALSE, quote =
            TRUE, sep = ";", eol = "\n", na = "NA", dec = ",", row.names = FALSE, col.names = TRUE, qmethod = c
            ("escape", "double"),fileEncoding = "")
}



#------------------------------
# R code used for illustrations
#------------------------------

####################
## Illustration 1 ##
####################

# Import DATA
base<-read.delim("data.csv",encoding="UTF-8",sep=";", dec=",", header=TRUE, na.string="",fill=TRUE)

# Weights calculation
test <- base[,c("ECD","ageR","cardiovasc","imc3clref2","anti.classI",
"Tps","Evt")]
test <- na.omit(test)
Pr0 <- glm(test$ECD~1, family = binomial(link="logit"), data=test)$fitted.values[1]
Pr1 <- glm(test$ECD~ ageR + cardiovasc + imc3clref2 + anti.classI
, data=test, family=binomial(link = "logit"))$fitted.values
W <- (test$ECD==1) * (Pr0/Pr1) + (test$ECD==0) * (1-Pr0)/(1-Pr1)
adjust <- 3.03
m1<-mean(W[test$ECD==1]); m0<-mean(W[test$ECD==0])
mW<-ifelse(test$ECD==1,adjust*m1,adjust*m0)
W2<-ifelse(W>mW,mW,W)

## Figure 1 ##
pdf("km_ECD.pdf")
```

```
        plot(NULL,xlim=c(0,13),ylim=c(0,1),yaxs="i",xaxs="i",ylab="Graft␣and␣patient␣survival", xlab="Time␣post
            -transplantation␣(years)")
        res.km <- KM(times=test$Tps, failures=test$Evt,  variable=test$ECD, weights=NULL)
        lines(res.km$times[res.km$variable==1],res.km$survival[res.km$variable==1],type="s",col=2,lty=2)
        lines(res.km$times[res.km$variable==0],res.km$survival[res.km$variable==0],type="s",lty=2,lwd=2)
        res.akm <- KM(times=test$Tps, failures=test$Evt,  variable=test$ECD, weights=W3)
        lines(res.akm$times[res.akm$variable==1],res.akm$survival[res.akm$variable==1],type="s",col=2)
        lines(res.akm$times[res.akm$variable==0],res.akm$survival[res.akm$variable==0],type="s",pch=1,lwd=2)
        nb.risk1<-function(x) {sum(test$Tps[test$ECD==0]>x)}
        nb.risk2<-function(x) {sum(test$Tps[test$ECD==1]>x)}
        segments(x0=0.5, y0=0.12, x1=12.5, y1=0.12)
        text(x=6, y=0.14, "number␣of␣at-risk␣patients", cex=0.8)
        tps <- seq(1,12,by=1)
        text(x=tps, y=rep(0.09,length(tps)), as.character(sapply(tps, FUN="nb.risk1")), cex=0.8, col=1)
        text(x=tps, y=rep(0.04,length(tps)), as.character(sapply(tps, FUN="nb.risk2")), cex=0.8, col=2)
        legend("topright", legend=c("Adjusted␣survival␣for␣SCD␣(n=1168)","Crude␣survival␣for␣SCD","Adjusted␣
            survival␣for␣ECD␣(n=721)","Crude␣survival␣for␣ECD"), col=c(1,1,2,2), lty=c(1,2,1,2), lwd=c(2,2,1,1)
            , cex=0.8)
dev.off()

## Tests ##
# ALRT
2*(1-pnorm(abs(ALRT(test$Tps, test$Evt, test$ECD, W)))) # 0.02371923
# ALRB
as.numeric(ALRB(test$Tps, test$Evt, test$ECD, W)[2]) # 0.06385161
# WWC
summary(coxph(Surv(Tps,Evt)~ECD, data=test, robust=TRUE, weights=W))$coefficients[1,6] # 0.07423174
# WMC
summary(coxph(Surv(Tps,Evt)~ECD+ageR + cardiovasc + imc3clref2 + anti.classI, robust=TRUE, data=test)) # 1.21e
    -05
# unadjusted Cox
summary(coxph(Surv(Tps,Evt)~ECD, robust=TRUE, data=test)) #  <2e-16
# SLR 1:1
fct_matchit <- function(.ratio){
        .matchit <- matchit(ECD~ ageR + cardiovasc + imc3clref2 + anti.classI, data = test, method = "nearest",
             distance = "logit", replace=FALSE, ratio=.ratio, caliper=0.2)
        D.app <- test
        D.app$app <- NULL
        D.app[row.names(.matchit$match.matrix)[!is.na(.matchit$match.matrix[,1])],"app"] <- row.names(.matchit$
            match.matrix)[!is.na(.matchit$match.matrix[,1])]
        sapply(c(1:dim(.matchit$match.matrix)[1]), FUN=function(x){
        D.app[.matchit$match.matrix[x,][!is.na(.matchit$match.matrix[x,])],"app"] <<- row.names(.matchit$match.
            matrix)[x]
        })
        D.app <- D.app[!is.na(D.app$app),]
        if(.ratio == 1){tabres <- c(NA,NA)
        }else{tabres <- c(NA,NA,NA)}
        try(tabres[1] <- 1 - pchisq(survdiff(Surv(D.app$Tps, D.app$Evt)~D.app$ECD + strata(D.app$app))$chisq,
            1))
        tabres[2] <-  sum(D.app$ECD) / sum(test$ECD) * 100
        if(.ratio > 1){tabres[3] <- dim(D.app[D.app$ECD==0,])[1] / sum(D.app$ECD) }
        return(tabres)
}
fct_matchit(1) # 0.3086, 0.0160, 0.0577, 0.0097
fct_matchit(3) #  0.00863411, 0.003092814, 0.03062015, 0.004307527


####################
## Illustration 2 ##
####################


# Import DATA
base3 <- read.csv("Data.csv",fill=TRUE,header=TRUE, sep = ";",dec = ",")

# Weights calculation
Pr0 <- glm(ratioP~1, family = binomial(link=logit), data=base3)$fitted.values[1]
Pr1 <- glm(ratioP~age + isc2cl + creat2cl + retransplant , family = binomial(link=logit), data=base3)$fitted.
    values
W <- (base3$ratioP==1) * Pr0/Pr1 + (base3$ratioP==0) * (1-Pr0)/(1-Pr1)
```



```
## Figure 2 ##
pdf("km_KwRw.pdf")
        plot(NULL,xlim=c(0,9.5),ylim=c(0.3,1),yaxs="i",xaxs="i",ylab="Graft␣survival", xlab="Time␣post-
            transplantation␣(years)")
        res.km <- KM(times=base3$Temps/12,failures=base3$status,variable=base3$ratioP,weights=NULL)
        lines(res.km$times[res.km$variable==1],res.km$survival[res.km$variable==1],type="s",lwd=2,lty=2)
        lines(res.km$times[res.km$variable==0],res.km$survival[res.km$variable==0],type="s",lty=2,col=2)
        res.akm <- KM(times=base3$Temps/12,failures=base3$status,variable=base3$ratioP,weights=W)
        lines(res.akm$times[res.akm$variable==1],res.akm$survival[res.akm$variable==1],type="s",lwd=2)
        lines(res.akm$times[res.akm$variable==0],res.akm$survival[res.akm$variable==0],type="s",pch=1,col=2)
        nb.risk1<-function(x) {sum(base3$Temps[base3$ratioP==0]>x)}
        nb.risk2<-function(x) {sum(base3$Temps[base3$ratioP==1]>x)}
        segments(x0=0.5, y0=0.42, x1=9, y1=0.42)
        text(x=4.5, y=0.44, "number␣of␣at-risk␣patients", cex=0.8)
        tps <- seq(1,10,by=1)
        text(x=tps, y=rep(0.39,length(tps)), as.character(sapply(tps, FUN="nb.risk2")), cex=0.8, col=1)
        text(x=tps, y=rep(0.34,length(tps)), as.character(sapply(tps, FUN="nb.risk1")), cex=0.8, col=2)
        legend(x=0.,y=0.65, legend=c(expression("Adjusted␣survival␣for␣for␣KwRw" >="2.3␣g.kg"),expression("
            Crude␣survival␣for␣for␣KwRw␣">="2.3␣g.kg␣"),expression("Adjusted␣survival␣for␣KwRw␣<␣2.3␣g.kg"),"
            Crude␣survival␣for␣for␣KwRw␣<␣2.3␣g.kg"), col=c(1,1,2,2), lty=c(1,2,1,2), lwd=c(2,2,1,1), cex=0.8)
dev.off()

## Tests ##
# ALRT
2*(1-pnorm(abs(ALRT(base3$Temps, base3$status, base3$ratioP, W)))) #  0.05515853
# ALRB
as.numeric(ALRB(base3$Temps, base3$status, base3$ratioP, W)[2]) #  0.05462589
# WWC
summary(coxph(Surv(Temps,status)~ratioP, data=base3, robust=TRUE, weights=W))$coefficients[1,6] #  0.05430477
# WMC
summary(coxph(Surv(Temps,status)~ratioP+age+creat2cl+retransplant+isc2cl,robust=TRUE, data=base3)) # 0.058906
# unadjusted Cox
summary(coxph(Surv(Temps,status)~ratioP,robust=TRUE, data=base3)) # 0.056
# SLR 1:1
fct_matchit(1) # 0.5961, 0.8348, 0.1037, 0.2695
fct_matchit(3) # 0.3388, 0.1420, 0.0391, 0.1227
```

3.3 Errata

Nous avons vu dans la section 2.1.2 que les HR ont la particularité d'être *non-collapsibles* et que dans ce cas, les effets marginaux et conditionnels d'une exposition, lorsqu'ils sont estimés par un HR peuvent être différents. Dans cet article, nous n'évoquons pas la *non-collapsibilité* des HR. Pourtant nous comparons une méthode estimant un effet conditionnel (i.e. le modèle de Cox multiple) à des méthodes estimant un effet marginal (i.e. les méthodes basées sur le score de propension). Nous n'avions pas connaissance lorsque nous avons réalisé ce travail de la différence qu'il peut exister entre un effet marginal et un effet conditionnel.

Notons que la différence entre l'effet marginal et conditionnel de l'exposition est souvent faible voire très faible. Dans une revue systématique de la littérature à partir des publications présentant à la fois l'effet marginal et l'effet conditionnel, Shah et al. (2005) ont montré que l'estimation marginale d'un OR ou d'un HR était en moyenne 6.4% plus proche de la valeur 1 que l'estimation conditionnelle. Dans une autre revue systématique, Stürmer et al. (2006) ont indiqué que dans 60 études parmi 69 analysées, l'effet estimé par une méthode marginale diffère de moins de 20% de celui estimé par une méthode conditionnelle. Cette différence souvent très faible entre l'effet individuel et populationnel explique certainement en partie l'actuelle méconnaissance de ces deux natures d'effet.

Nous avons comparé les différentes méthodes en termes de risques de première et seconde espèces. Il a été montré que les effets marginaux et conditionnels coïncident lorsque l'effet de l'exposition est nul (Gail et al., 1984; Greenland, 1987). Ainsi, en ce qui concerne le risque de première espèce calculé sous l'hypothèse nulle, nos résultats restent corrects. Concernant le risque de seconde espèce, la situation est plus préoccupante. Les données sont simulées à partir d'un modèle conditionnel afin d'avoir un effet conditionnel de l'exposition non nul, dans ce cas l'effet marginal de l'exposition est également non nul. Ainsi, l'ensemble des méthodes appliquées estiment un effet non nul et l'on est donc en droit de calculer un risque de seconde espèce. Cependant, des travaux d'Hernán et al. (2011) et de Pocock et al. (2002) ont montré que les OR et les HR conditionnels étaient souvent plus éloignés de la valeur 1 que leurs équivalents marginaux. On est donc en droit de penser que la supériorité du modèle de Cox multiple en termes de puissance est peut être en partie due à cette différence de taille d'effet.

Pour donner un ordre de grandeur de ce biais et interpréter au mieux les résultats obtenus en termes de risque de seconde espèce, nous avons recalculé les HR marginaux théoriques pour les différents scénarios simulés selon l'approche proposée par Gayat et al. (2012). Cette approche consiste à simuler les covariables pour un large échantillon d'un million d'individus. Ensuite, on considère que les 500 000 premiers sujets sont exposés et que les suivants ne le sont pas. Les temps d'événements et de censure sont ensuite simulés. Les HR marginaux théoriques sont alors obtenus par un modèle de Cox univarié où la seule variable incluse dans le modèle est l'exposition. Dans les simulations correspondantes à la première application sur les donneurs à critères élargis (en anglais, *expanded criteria donor*, ECD), les logarithmes des HR théoriques marginaux obtenus sont les suivants : 0,238, 0,348 et 0,477 correspondants respectivement aux scénarios simulés avec des logarithmes des HR conditionnels égaux à 0,250, 0,365 et 0,500. Dans les simulations correspondantes à la seconde application en lien avec le poids des reins, les logarithmes des HR théoriques marginaux et conditionnels sont respectivement égaux à -0,249 et -0,255 dans la situation où le biais de confusion est faible, et à -0,224 et -0,245 dans la situation où le biais de confusion est fort. Finalement, les différences entre les HR théoriques marginaux et conditionnels sont faibles et il semble peu probable qu'elles remettent en cause nos conclusions en termes de risque de seconde espèce. De plus, les écarts de puissance observés entre les différentes méthodes sont très variables selon la taille de l'échantillon alors que l'écart entre l'effet théorique marginal et conditionnel est indépendant de la taille de l'échantillon.

Une autre distinction omise dans ce travail concerne la nature différente des effets marginaux estimés par l'appariement et par les autres méthodes basées sur le score de propension. En effet, la méthode appariée estime l'effet populationnel moyen chez les exposés (i.e. l'ATT) alors que les trois autres méthodes IPW estiment l'effet populationnel moyen dans l'ensemble de la population (i.e. l'ATE). Là encore, pour évaluer le potentiel impact de cette différence sur les résultats des simulations, nous avons recalculé les HR marginaux théoriques correspondants à l'ATT pour les différents scénarios simulés selon l'approche proposée par Austin (2013). Dans les simulations correspondantes à l'application sur les ECD, les logarithmes des HR théoriques marginaux dans la population des exposés obtenus sont les suivants : 0,239, 0,347 et 0,477 correspondants respectivement aux trois scénarios simulés avec des logarithmes des HR conditionnels égaux à 0,250, 0,365 et 0,500. Dans les simulations correspondantes à l'application en lien avec le poids des reins, les logarithmes des HR théoriques marginaux dans la population des exposés (i.e. ATT) sont égaux à -0,251 dans la situation où le biais de confusion est faible et -0,228 dans la situation où le biais de confusion est fort. On remarque que les effets théoriques ATE et ATT sont extrêmement proches pour tous les scénarios simulés.

Chapitre 4

Courbes ROC standardisées et pondérées dépendantes du temps pour évaluer les capacités pronostiques intrinsèques à un marqueur en tenant compte des facteurs de confusion

Sommaire

4.1 Manuscrit en révision

Résumé de l'article

Une grande partie des recherches biomédicales concerne actuellement la médecine personnalisée. Néanmoins ce concept n'est pas nouveau. Son expansion actuelle est principalement due au développement des nouvelles technologies (omics, bio-imagerie, etc.) permettant d'identifier de nouveaux biomarqueurs pour prédire la progression d'une maladie, le risque de complications, ou la réponse à un traitement. Cependant, très peu de ces nouveaux prédicteurs sont utilisés en clinique (Ioannidis et al., 2014; Macleod et al., 2014). Ce constat est en partie lié à de nombreux problèmes méthodologiques : faibles tailles des échantillons, absence d'échantillon de validation, procédures non-standardisées, etc. Parmi ces écueils, la non-comparabilité entre les cas et les témoins peut aboutir à l'identification de prédicteurs faux-positifs. La courbe ROC (*Receiver Operating Characteristic*) dépendante du temps est un outil couramment utilisé pour évaluer les capacités pronostiques d'un prédicteur. Les courbes ROC ajustées tenant compte des facteurs de confusion ont récemment été proposées mais uniquement en l'absence de données censurées (Pepe et Cai, 2004; Pepe et Longton, 2005). Nous proposons un nouvel estimateur pour obtenir des courbes ROC dépendantes du temps ajustées par une méthode de pondération basée sur le score de propension (IPW). Cet estimateur fournit une mesure de la capacité prédictive du marqueur en prenant en compte les facteurs potentiellement confondants. Les résultats de nos études de simulations permettent de valider cet estimateur, en particulier dans le contexte spécifique d'absence de censure où la comparaison avec la méthode proposée par Pepe et Cai (2004) est possible. Nous illustrons également l'intérêt de notre estimateur par deux applications en transplantation rénale sur les données réelles de la cohorte DIVAT.

Statistical Methods in Medical Research
XX(X):2–??

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Standardized and weighted time-dependent ROC curves to evaluate the intrinsic prognostic capacities of a marker by taking into account confounding factors

Florent Le Borgne[1,2,3], Christophe Combescure[4], Florence Gillaizeau [1,9],Magali Giral[1,3], Marion Chapal [1,10], Bruno Giraudeau[5,6,7,8] and Yohann Foucher[1,11]

Abstract

Time-dependent Receiver Operating Characteristic (ROC) curves allow to evaluate the capacity of a marker to discriminate between subjects who experience the event up to a given prognostic time from those who are free of this event. In this paper, we propose an inverse probability weighting (IPW) estimator of a standardized and weighted time-dependent ROC curve. This estimator provides a measure of the prognostic capacities by taking into account potential confounding factors. We illustrate the robustness of the estimator by a simulation-based study and its usefulness by two applications in kidney transplantation.

Keywords

Prognostic capacities, Confounding factors, Inverse probability weighting, Propensity score, ROC curve, Time-to-event data.

Prepared using sagej.cls *[Version: 2015/06/09 v1.01]*

1 Introduction

Over the last decade the number of publications related to personalized medicine has grown exponentially. This has been primarily due to the emergence and application of *omics*-technologies resulting in the identification of new biomarkers for predicting disease evolution or treatment response. Nevertheless, the transfer of new biomarkers to clinical practice has been relatively underwhelming, and may be partially explained by mistrust of the proposed biomarker performance by healthcare professionals. This may reflect numerous limitations of the publications including small sample sizes, lack of validation sample[6], and patient characteristics and environmental factors are often ignored. This may result in the non-comparability between cases and controls[16], and the identification of a biomarker whose prognostic capacities are in fact due to confounding factors. For instance, when patient age is associated with both the biomarker value and the disease evolution, a part of the prognostic capacities due to the patient age will be attributed to the biomarker. In this case, a substantial risk is to classify as a good predictor a biomarker that in fact, only reflects the expression of clinical or environmental risk factors, which are easier and less expensive to measure. As demonstrated by Janes and Pepe[14], the methods for assessing diagnostic/prognostic capacities should encompass potential confounding factors.

To evaluate the diagnostic or prognostic capacities of any marker defined on a continuous scale, the receiver operating characteristic (ROC) curve is probably the most popular tool. The ROC curve is a graphical plot of the sensitivity versus 1 – specificity across all possible threshold values of the continuous marker. The

[1]EA 4275-SPHERE methodS for Patient-centered outcomes & HEalth REsearch, Nantes, France
[2]IDBC/A2com, Pace, France
[3]ITUN, INSERM U1064, Nantes, France
[4]CRC and Division of Clinical Epidemiology, Department of Health and Community Medicine, University of Geneva, University Hospitals of Geneva, Geneva, Switzerland
[5]Centre de recherche Epidémiologie et Biostatistique, INSERM U1153, Paris, France
[6]Centre d'Investigation clinique (CIC), INSERM 1415, Tours, France
[7]Université François-Rabelais de Tours, PRES Centre-Val de Loire Université, Tours, France
[8]CHRU de Tours, Tours, France
[9]Department of Statistical Science, University College London, London, United Kingdom
[10]Médecine néphrologie - Hémodialyse, Centre Hospitalier Départemental Vendée Site de La Roche sur Yon, France
[11]Nantes University Hospital, Nantes, France

Corresponding author:
Yohann Foucher, Nantes University 1, rue Gaston Veil BP 53508 44035 Nantes Cedex 1 FRANCE
Email: Yohann.Foucher@univ-nantes.fr

Prepared using sagej.cls

sensitivity is the probability that the biomarker be superior to the threshold among the cases and the specificity is the probability that the biomarker be inferior or equal to the threshold among the controls. The area under the ROC curve (AUC) can be interpreted as the probability that among two randomly chosen subjects, one case and one control, the value of the marker be higher for the case than for the control. Therefore, an AUC of 0.5 indicates that the marker is non-informative. An increase in the AUC indicates an improvement in discriminative capacities, with a maximum at 1.0.

A possible solution for taking into account confounder(s) in ROC analysis is to stratify the analysis according to confounder levels and to compute the weighted mean of the AUC. However, only a low number of covariates can be considered, continuous covariates must be categorized, and a global ROC curve cannot be obtained. For diagnostic markers, Janes and Pepe [15] recently proposed regression models to assess covariate-adjusted ROC curves based on the theory of placement values (PV) [20;21]. The principle is to consider the ROC curve as a cumulative distribution function of the marker in cases, the marker being standardized according to the distribution in controls. Over the same period, time-dependent ROC curves have also been developed to accommodate for censored times-to-event data [9;10;17], i.e. for the assessment of long-term prognostic capacities. In this context, two main types of time-dependent ROC curve have been proposed [10]: i) the time-dependent *cumulative* ROC curve where a case is a subject who has the event before the time τ, and ii) the time-dependent *incident* ROC curve where a case is a subject who has the event at the time τ. In this paper, we focus on the *cumulative* definition. Very few methods are available to account for covariates in the presence of time-to-event data, and all were covariate specific, i.e. a specific estimation for each value of the covariates. More specifically, Song and Zhou [27] proposed a semi-parametric estimator, Rodríguez-Álvarez *and others* [23] proposed a nonparametric kernel-based estimator, and Zheng *and others* [32] proposed a predictive values curve.

To our knowledge no estimator has been proposed for covariate-adjusted time-dependent ROC curves, i.e. allowing a global ROC curve to be obtained by taking into account multiple and/or continuous covariates. We propose in this paper such an estimator by standardizing the marker and by inverse probability weighting (IPW). The corresponding propensity score is the probability for a subject to be a case, i.e. the event occurs before the prognostic time given his/her covariates. This paper is arranged as follows. In the next section, we define standardized and weighted time-dependent ROC curve. In the section 3,

a simulation-based study is proposed. In the section 4, we apply the methods to two examples to illustrate the usefulness of the approach. The last section is dedicated to the discussion.

2 Methods

2.1 Observations

Let T the time-to-event, C the time to last follow-up (right censoring) and X the marker of interest. Because of random right censoring, we only observe $Y = \min(T, C)$ and $\delta = \mathbb{1}(T \leq C)$ the indicator of event. Assume that larger values of X are associated with greater hazards; otherwise, X can be recoded if necessary to achieve this. Let Z a p-dimensional vector of baseline covariates. Assume that the data $\{(Y_j, X_j, Z_j, \delta_j) : i = 1, \ldots, n\}$ are independent and identically distributed samples from (Y, X, Z, δ).

2.2 Crude estimators

For a threshold value v, a subject j, is classified in the high-risk group when the test is positive, i.e. $\{X_j > v\}$. Let the sensitivity at time τ, $se_c(\tau, v) = P(X > v|T \leq \tau)$, the proportion of positive tests among patients with the event before τ, according the cumulative definition proposed by Heagerty and Zheng [10]. Let the specificity at time τ, $sp_c(\tau, v) = P(X \leq v|T > \tau)$, the proportion of negative tests among patients without event up to τ. When there is no censoring $(\delta_j = 1, \forall j)$, $se_c(\tau, v)$ can be estimated by the proportion of subjects with $\{X_j > v\}$ among subjects with event before τ and $sp_c(\tau, v)$ can be estimated by the proportion of subjects with $\{X_j \leq v\}$ among subjects free of the event at time τ. In presence of censoring, the sensitivity and specificity can be naïvely assessed by the following proportions:

$$\widehat{se}_n(\tau, v) = \frac{\sum_{j=1}^n \delta_j \mathbb{1}(Y_j \leq \tau, X_j > v)}{\sum_{j=1}^n \delta_j \mathbb{1}(Y_j \leq \tau)} \tag{1}$$

$$\widehat{sp}_n(\tau, v) = \frac{\sum_{j=1}^n \mathbb{1}(Y_j > \tau, X_j \leq v)}{\sum_{j=1}^n \mathbb{1}(Y_j > \tau)} \tag{2}$$

However, in the sensitivity computation removing the subjects censored before the prognostic time τ can induce bias [1]. To account for the information brought by subjects with unknown status at time τ, the inverse probability of censoring weighting (IPCW) estimators have been proposed by Uno *and others* [29] and

Hung and Chiang[11]. These IPCW estimators represent corrections of the "naïve" ones by weighting the uncensored subjects at time τ by their probability of being uncensored:

$$\widehat{se}_c(\tau, v) = \frac{\sum_{j=1}^{n} \delta_j \mathbb{1}(Y_j \leq \tau, X_j > v)/\hat{S}_C(Y_j)}{\sum_{j=1}^{n} \delta_j \mathbb{1}(Y_j \leq \tau)/\hat{S}_C(Y_j)} \tag{3}$$

$$\widehat{sp}_c(\tau, v) = \frac{\sum_{j=1}^{n} \mathbb{1}(Y_j > \tau, X_j \leq v)}{\sum_{j=1}^{n} \mathbb{1}(Y_j > \tau)} \tag{4}$$

where $\hat{S}_C(Y_j)$ is the estimated probability that the censoring time occurs after Y_j, by using the Kaplan-Meier estimator for instance. Note that this specificity estimator is the same as the "naïve" one because all the weights at time τ are equal to $\hat{S}_C(\tau)$, which allows simplification of the formula. Note also that when there is no censoring, both terms δ_j and $S_C(Y_j)$ are equal to 1 for all subjects, they may therefore be omitted. Then, the estimators (3) and (4) are respectively equal to the proportions of positive/negative tests among cases/controls. The prognostic capacities of the marker can be summarized by the ROCτ curve, which represents $\widehat{se}_c(\tau, v)$ plotted against $1 - \widehat{sp}_c(\tau, v)$ for all the threshold values v.

2.3 The standardized and weighted estimators

To account for covariates, we focus on estimating a weighted sensitivity such as $se_w(\tau, v) = \mathbb{E}[se_c(\tau, v|Z)]$ and a weighted specificity such as $sp_w(\tau, v) = \mathbb{E}[sp_c(\tau, v|Z)]$ where the expectation is taken with respect to Z.

To take into account the confounding factors in etiologic studies, Rosenbaum[24] have proposed the use of weights based on the propensity score[25]. The propensity score weighting aim to create a "pseudo-sample" in which the distribution of baseline covariates is similar between exposed and unexposed subjects. The time-dependent sensitivity and specificity are conditional probabilities on the outcomes. Thus, in accordance with the idea proposed by Rosenbaum[24], we adapted the weights in order to create a "pseudo-sample" in which the distribution of baseline covariates is similar between cases and controls. Let $p_j(\tau) = P(T_j \leq \tau|z_j)$ the probability to experience the event before the time τ for the jst subject according to his/her characteristics and $\hat{p}_j(\tau)$ its estimator, for instance derived from a Cox proportional hazard (PH) model. To take into account confounders in the estimator (3). One can propose to correct

the contribution of each subject j by $w_{j1}(\tau) = \hat{p}_j(\tau)$:

$$\widehat{se}_w(\tau, v) = \frac{\sum_{j=1}^n \delta_j \mathbb{1}(Y_j \leq \tau, X_j > v)/\{\hat{S}_C(Y_j) w_{j1}(\tau)\}}{\sum_{j=1}^n \delta_j \mathbb{1}(Y_j \leq \tau)/\{\hat{S}_C(Y_j) w_{j1}(\tau)\}} \tag{5}$$

This correction reduces the contribution of over-represented profiles in cases, and increases the contribution of under-represented ones. Similarly, for the specificity (4). One can propose to correct the contribution of controls by $w_{j0}(\tau) = 1 - \hat{p}_j(\tau)$:

$$\widehat{sp}_w(\tau, v) = \frac{\sum_{j=1}^n \mathbb{1}(Y_j > \tau, X_j \leq v)/w_{j0}(\tau)}{\sum_{j=1}^n \mathbb{1}(Y_j > \tau)/w_{j0}(\tau)} \tag{6}$$

In absence of censoring, one can show that the estimators (5) and (6) correspond to stratified estimators when covariates Z are categorical (see Appendix A.1 of the online supplementary materials for more details). The corresponding time-dependent IPW ROCτ curve is defined as the plot of $\widehat{se}_w(\tau, v)$ versus $1 - \widehat{sp}_w(\tau, v)$ for all the threshold values v. In order to avoid unstable results when $w_{j1}(\tau)$ or $w_{j0}(\tau)$ are close to zero, we propose $w_{j1}(\tau) = \max\{0.1\bar{p}(\tau)^2, w_{j1}(\tau)\}$ and $w_{j0}(\tau) = \max\{0.1(1 - \bar{p}(\tau))^2, w_{j0}(\tau)\}$, where $\bar{p}(\tau) = \hat{P}(T \leq \tau)$. The term $0.1\bar{p}(\tau)^2$ has been chosen from simulations.

To estimate the confounder-adjusted prognostic capacity of a marker, the association between the marker and the covariates has also to be consider. Indeed, as described by Janes and Pepe [14], for a covariate only associated with the marker (not associated with outcome), the same threshold used in the different strata could lead to very different sensitivity and specificity values. In other words, even if the ROC curves in each strata are close, the same threshold v could lead to very different points on the ROC curves per strata. In this situation, calculating the weighted mean of the sensitivities and specificities estimated in each strata result in an underestimation of the AUC. Then, we propose to standardize the marker according to the covariates among the controls before calculating the IPW estimators (5) and (6). This standardization of the marker allow to obtain a ROC curve which is a weighted average of covariate-specific curves as defined by Janes and Pepe [15]. In Appendix A.2 of the online supplementary materials, we additionally show that the proposed standardized and weighted estimators are equivalent to the crude ones when the covariates are independent of both the marker and the time-to-event.

One can note that in the context of uncensored data, the results should be similar to the ones obtained by using placement values[20], $\{T_j \leq \tau\}$ being the cases and $\{T_j > \tau\}$ the controls (see Appendix A.3 of the online supplementary materials for more details on placement values method).

2.4 *Which factors should be considered in IPW, stratified and PV estimators ?*

In etiologic analysis, confounding mainly occurs when a covariate is associated with both the marker and the times-to-event, but perturbation in the measure of the prognostic capacities of a marker could also occurs in other cases. Indeed, as we previously outlined, Janes and Pepe[14] demonstrated the need to take into account a covariate once it is associated with the marker. Similarly, as shown in Figure 1, it is necessary to take into account covariates associated with the times-to-event, regardless of their association with the marker. For this purpose, we simulated a completely arbitrary data set of 10 000 subjects and one binary covariate Z. We simulated Z from a Bernoulli distribution with parameter equal to 0.5, and independently, X from a normal distribution with mean equal to 0 and standard deviation equal to 2. We then simulated the times-to-event from a Weibull PH model such as $F(t|x_j, z_j) = 1 - \exp[0.5 \exp(0.3x_j + 1.1z_j)t^{2.5}]$. We did not consider competing censoring times and fixed the prognostic time τ at 10 years. One can observe (Figure 1) that the crude ROC (estimated for the whole sample) is different from the stratified one estimated in the strata $Z = 1$, and very close to the stratified one estimated in the strata $Z = 0$. The curve estimated with the IPW ROC estimator lies between the two stratified ROC curves.

2.5 *Estimation procedure*

To standardize the marker according to the covariates among the controls, one can fit a multiple linear regression with the marker as the outcome and the other covariates as explicative variables such as: $x_j = \beta_0 + \beta_1 z_{1j} + \cdots + \beta_p z_{pj} + \varepsilon_j$ by considering only subjects with a participating time Y_j superior to τ. Indeed, analogously to the demonstration provided by Blanche *and others*[1] of the consistency of the "naïve" specificity estimator if censoring is independent of X and T, the distribution of the marker among the controls can be estimated by removing all subjects censored before the time τ if censoring is independent of X, Z and T (see Appendix A.4 of the online supplementary materials for

more details). The random error could be assumed to be normally distributed, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Then, we used the standardized marker values obtained as follows: $[x_j - (\hat{\beta}_0 + \hat{\beta}_1 z_{1j} + \cdots + \hat{\beta}_p z_{pj})]/\hat{\sigma}$ instead of x_j in the equations (5) and (6). In case of heteroscedasticity, an alternative is to fit a linear model using generalized least squares (GLS) method by specifying a variance structure function allowing the errors to have unequal variances according covariates and/or marker values. To estimate the conditional probability $p_j(\tau)$, we first fitted a multivariable Cox PH model[5]. From the estimated regression coefficients $\hat{\gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_p)$, we secondly computed the cumulative baseline hazard function $\Lambda_0(\tau)$ using the Breslow estimator[2]. The probability $p_j(\tau)$ could then be estimated as $\hat{p}_j(\tau) = 1 - \exp\{-\hat{\Lambda}_0(\tau)\exp(\hat{\gamma}_1 z_{1j} + \cdots + \hat{\gamma}_p z_{pj})\}$. The AUC was computed by the trapezoidal rule. The 95% confidence interval of the AUC can be obtained from 1000 bootstrap replications, and by using the 2.5th and 97.5th percentiles of the corresponding empirical distribution. All the estimation procedure is summarize in Figure 2.

Note that we have implemented the four estimators (crude, stratified, IPW, and PV) in the R package ROCt.

3 Simulations-based study

3.1 Data-generating process

We simulated data of n subjects ($j = 1, \ldots, n$) for a setting in which there was a single categorical covariate Z with four levels of equal probability. Next, X was generated from a normal distribution with mean equal to $\beta_Z z_j$ and standard deviation equal to 1. The times-to-event were generated from a Weibull PH model such as $F(T_j|x_j, z_j) = 1 - \exp[\lambda \exp(\alpha_X x_j + \alpha_Z z_j) t_j^{\nu}]$. We simulated two scenarios related to the relationship between the covariate and both the marker and the times-to-event. For moderate associations, we aimed to obtain a R^2 at 0.20, and we fixed $\beta_Z = 0.4$ and $\alpha_Z = 0.4$. For strong associations ($R^2 = 0.45$), we fixed $\beta_Z = 0.8$ and $\alpha_Z = 0.8$. The value of the regression coefficient α_X was 0 under the null hypothesis of a non-informative marker (AUC_τ=0.5), and 0.5 under the alternative one ($AUC_\tau > 0.5$). The prognostic time τ was 10 years. The coefficients associated with the baseline Weibull distribution were chosen in order to obtain a 10-year survival equal to 40% regardless the confounding strength and the informative or non-informative status of the marker. The censoring times were independently generated from a Weibull distribution with

parameters defined to obtain three different censoring rates: 0, 0.25 and 0.5. We also evaluated the performance of the estimators for various sample sizes (500, 1000 and 2000). For each scenario, we simulated 10 000 datasets. Note that we conducted a complementary simulations-based study with multiple covariates. The design and the results of these simulations are presented in the Appendix A.5 of the online supplementary materials. In these simulations, we also evaluated the impact of covariates selection on the results.

For the alternative hypothesis, the theoretical values were asymptotically computed by simulating one broad dataset of 10 000 000 subjects, and using the stratified estimator on the four strata of the covariates Z (see section 2.3). Note that we applied the PV method only to validate the proposed IPW estimators in the particular situation where there is no censoring, the PV method being not applicable in the presence of censoring (cases and controls have to be completely observed).

3.2 Results

For scenarios under the null hypothesis, the results are presented in Table 1. Crude AUCs higher than 0.5 illustrated the necessity to take into account the covariates. For uncensored data, both PV and IPW estimators performed very well with the mean bias equal to 0. In terms of standard deviations, the RMSE obtained from the two methods were also close. The IPW estimator performed as well in the presence of censoring, with an expected slight increase in RMSE.

Table 2 describes the results under the alternative hypothesis. In absence of censoring, the IPW and the PV estimators were always associated with a mean bias very close to 0, with a low superiority of the IPW approach. In terms of RMSE, we observed equivalent results with the two estimators. In presence of censoring, we observed a small increase in both bias and RMSE with the IPW estimator. The web supplementary Tables S1 and S2 presented the results of the simulations conducted in presence of four covariates. The results confirmed those obtained with only one covariate. We additionally observed that under the null hypothesis, considering the variables associated with both the marker and the times-to-event results in bias results. Under the alternative hypothesis, considering only the covariates associated with the marker led to under-estimate the AUC and considering only the covariates associated with the outcome led to over-estimate the AUC. The best results were obtained by considering all the covariates (which are at least associated to the marker or the time-to-event).

Globally, these simulation results allow to validate the proposed IPW estimator when there is no censoring by comparing the results with the ones obtained with the PV estimator and in presence of censoring by observing the stability of the results between the scenarios in which the only difference is the introduction of censoring.

4 Applications in kidney transplantation

4.1 Cold ischemia time for the prognosis of delayed graft function

The objective of this application was to validate our approach in the absence of censoring where we would anticipate the results would compare closely to the PV method. Delayed graft function (DGF) is a common complication in kidney transplantation, and reflects the renal injuries linked to ischemia and reperfusion that occur after cold and warm preservation of the graft. DGF is considered as an acute primary renal failure, and is clinically defined by the need for dialysis within the first seven days post-transplantation[31]. DGF is known to be associated with higher rates of acute allograft rejection, poorer 12-month post-transplantation renal function, poorer long-term graft and recipient survival, and higher costs associated with patient management[28;30]. DGF is associated with well-described risk factors[12;22], among which cold ischemia time (CIT) is one of the main predictors[3;19]. In this application we aimed to evaluate for the first time the intrinsic capacities of CIT to discriminate kidney recipients into two groups: those who will not experience DGF after kidney transplantation (controls), and those who will experience DGF (cases). We re-analysed the data from a previous study[3]. Adult recipients of isolated, non-preemptive, non machine-perfused, deceased donor kidneys, who were prospectively collected since January 2007 and computerized in the DIVAT (www.divat.fr, CNIL 891735) databank were included. The outcome was the time between transplantation and the first occurrence of dialysis. The prognostic time τ was fixed to 7 days according to the DGF definition.

Among the 1844 included recipients, 468 experienced DGF. The web supplementary Table S3 outlines the parameters which can be associated with CIT, DGF probability or both. According to the reasons explained in sections 2.3 and 2.4, we selected the covariates significantly associated with either CIT or DGF ($p < 0.05$). Twelve covariates were retained (shown in bold in supplementary web Table S3). The standardized marker values were

computed from a multiple linear model between the CIT and the covariates. The homoscedasticity assumption was graphically verified for each of the covariates, no violation has been detected.

The AUC was estimated at 0.62 (95% CI = [0.59;0.65]) with the crude estimator, 0.62 (95% CI = [0.58;0.65]) with the standardized and weighted estimator, and 0.61 (95% CI = [0.58;0.65]) with the PV estimator. The corresponding ROC curves are presented in Figure 3. The prognostic capacities of the CIT were really attributable to the marker, and were not influenced by confounders. As expected, results for the IPW and PV estimators were similar.

4.2 12-month serum creatinine after kidney transplantation for prognosis of return to dialysis

The objective of this second application is to apply the IPW estimators in the presence of censoring and to illustrate its usefulness in terms of interpretation. Numerous studies have previously demonstrated the crude capacities of the 12-month serum creatinine measurement (SCr) to predict graft lost [7;8;18]. It is also well-established that this marker is associated with other risk factors of graft failure such as delayed graft function (DGF), recipient age and gender [8;30]. Our objective was to study for the first time the intrinsic capacity of the 12-month post-transplantation SCr measurement to discriminate kidney recipients into two groups: those with a functional renal transplant at 10 years post-transplantation (controls) and those who returned to dialysis or required retransplantation (cases). Our analysis was based on recipients prospectively included in the French multicenter DIVAT cohort and transplanted between January 2000 and December 2013. Patients who died or censored before their first graft anniversary were not included. The outcome was the time between the first anniversary of the transplantation and the return to dialysis or retransplantation (death-censored graft survival). Among the 5158 included recipients, 678 returned to dialysis. The web supplementary Figure S1 shows the cumulative incidence function of graft lost. The probability of graft loss was 24.7% at 10 years post-transplantation (95% CI = [22.7%;26.6%]). In contrast with our first example, the times-to-event were censored and the PV estimator proposed by Janes and Pepe [15] can not be applied.

The web supplementary Table S4 describes the parameters potentially associated with the 12-month SCr measurement, or with the risk of graft failure or both. By using the same strategy of variable selection as in the first

application, 20 covariates were retained (shown in bold in supplementary web Table S4). We observed heteroscedasticity in the multiple linear model used to compute the standardized marker values. To deal with this issue, we used a GLS estimator with different variances of residuals for each combinations of donor gender, delayed graft function (DGF) and donor type (expanded criteria donor, ECD). The choice of the covariates included in the variance model were made to obtain for each covariate, a similar distribution of the standardized marker in controls across levels of the covariate. The crude and IPW AUCs for a prognostic time up to 10 years post-transplantation was equal to 0.72 (95% CI = [0.68;0.75]) and 0.66 (95% CI = [0.63;0.70]), respectively. The Figure 4 shows the crude and IPW ROC curves obtained for prognostic times fixed at 3, 7 and 10 years. In addition, a plot of the AUC estimated from the crude and IPW estimators according to prognostic time τ varying from 3 to 10 years is provided. Despite the fact that the prognostic capacity of the SCr value was lower when covariates were considered, this marker remains important to monitor since it offers additional information by itself, and independently of other possible confounding factors.

5 Discussion

In this paper, we proposed an estimator of time-dependent ROC curve to evaluate the prognostic capacity of a marker. With this approach, the advantage is to control for potential confounders and to be applicable when the survival outcome is right censored. In addition, the assessment of covariate-specific time-dependent ROC curves, impracticable when the number of potential confounders is high, is not needed. A package for R, ROCt package, is available to facilitate the use of our approach (http://www.divat.fr/en/softwares).

In absence of censoring, Janes and Pepe[14] showed that the confounding phenomenon is a concern for the assessment of a ROC curve and that confounding factors should not be ignored. The magnitude and the direction of the confounding phenomenon depends on the associations between the confounding factor, the outcome and the marker. Janes *and others*[13] proposed a procedure to assess a ROC curve adjusted for potential confounders using placement values. If the covariate-specific ROC curves are the same for any value of the covariates, the resulted adjusted ROC curve is the common covariate-specific ROC curve and the area under curve is an estimate of the probability for the marker to be higher in a case than in a control with the

same values of covariates. If not, the adjusted ROC is an weighted average of the covariate-specific ROC curves where weights are related to the distribution of the covariates. Equivalently, the principle of this approach is to average the covariate-specific sensitivities for a given specificity (the same specificity for any values of covariates). Janes and Pepe [15] standardized the marker so that the distribution of the standardized marker is similar in controls for any values of covariates. Our approach is similar since the covariate-specific sensitivities corresponding to a common specificity (as in the approach by Janes and Pepe [15]) is similar to averaging the covariate-specific sensitivities at a common threshold. Therefore, the interpretation of the adjusted ROC curve obtained with the two methods is similar. In case of heteroscedasticity, the standardization of the marker for the two approaches should involve a modelling of the variance.

In contrast to the approach based on placement values, the estimator we proposed is applicable to assess the discriminatory capacities of the marker when the survival outcome is right censored. For this purpose, we proposed IPW estimators of the time-dependent sensitivities and specificities. These estimators are based on the IPCW approach [1] which have been already applied to assess time-dependent ROC curve, and the IPW ones aiming to create a "pseudo-sample" in which the distribution of baseline covariates is similar between cases and controls. The IPW estimators are unbiased if the censoring is independent of the time-to-event, the marker and the covariates.

The simulation study confirmed that the estimators we proposed perform well when the censoring is independent of the time-to-event, the marker and the covariates : the mean bias was close to 0 in presence of a confounding factor moderately associated with the time-to-event and the marker and was small in presence of a confounding factor strongly associated with the time-to-event and the marker. In absence of censoring, our approach and the approach proposed by Janes and Pepe [15] performed similarly with low bias and similar RMSE. We investigated the impact of covariates selection in presence of multiple covariates. The results of the simulation study were in agreement with the recommendation made by Janes and Pepe [14]: the ROC curve should be adjusted for all covariates associated with either the marker or the time-to-event.

In addition, we illustrated the usefulness of the method with two applications in kidney transplantation. In the first application, we evaluated and confirmed the capacity of CIT to discriminate between patients who will experience DGF or not, with adjustment on potential confounders. In this application, the survival outcome was not censored and the estimates obtained with the IPW

estimator and the placement values estimator were very close. Moreover, the area under the adjusted and non-adjusted ROC curves was similar, showing a lack of confounding phenomenon attributable to the accounted covariates. In the second application, the survival outcome was right-censored and the placement values estimator was not applicable. The results showed that the area under the adjusted ROC curve of the serum creatinine was lower than the area under the non-adjusted ROC curve, especially for long prognostic times. Ignoring the confounders, the apparent ability of serum creatinine to discriminate patients who will experience a loss of graft and patients without a loss of graft is greater than the intrinsic ability of serum creatinine.

Note that, in this paper, we focused on the cumulative definition of the ROC curve provided by Heagerty and Zheng [10]. The confounding phenomenon is also a concern for other methods or statistics aiming to capture the prognostic ability of a marker, for instance the incident time-dependent ROC curve [10] or the prognostic ROC curve [4]. Further developments are needed to make possible the adjustment on potential confounders in these methods.

6 Supplementary material

Supplementary material is available at http://smm.sagepub.com

7 Acknowledgements

This work was supported by a grant of the French Agency of Research (ANR-11-JSV1-0008-01). We also wish to thank the clinical research assistant team (S. Le Floch, C. Scellier, P. Przednowed, V. Godel, K. Zurbonsen, C. Dagot, C. Coelle and J. Posson).

References

1. Blanche, Paul, Dartigues, Jean-François and Jacqmin-Gadda, Hélène. (2013, September). Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal. Biometrische Zeitschrift* **55**(5), 687–704.
2. Breslow, NE. (1972). Discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society* **B**(34), 216–217.
3. Chapal, Marion, Le Borgne, Florent, Legendre, Christophe, Kreis, Henri, Mourad, Georges, Garrigue, Valérie, Morelon, Emmanuel,

Buron, Fanny, Rostaing, Lionel, Kamar, Nassim, Kessler, Michèle, Ladrière, Marc, Soulillou, Jean-Paul, Launay, Katy, Daguin, Pascal, Offredo, Lucile, Giral, Magali *and others*. (2014). A useful scoring system for the prediction and management of delayed graft function following kidney transplantation from cadaveric donors. *Kidney International* **86**(6), 1130–1139.

4. Combescure, Christophe, Perneger, Thomas V., Weber, Damien C., Daurès, Jean-Pierre and Foucher, Yohann. (2014, January). Prognostic ROC curves: a method for representing the overall discriminative capacity of binary markers with right-censored time-to-event endpoints. *Epidemiology (Cambridge, Mass.)* **25**(1), 103–109.
5. Cox, DR. (1972). Regression Models and Life-Tables (with discussion). *Journal of the Royal Statistical Society* **B**(24), 187–220.
6. Drucker, Elisabeth and Krapfenbauer, Kurt. (2013, February). Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine. *EPMA Journal* **4**(1), 7.
7. Foucher, Yohann, Daguin, Pascal, Akl, Ahmed, Kessler, Michèle, Ladrière, Marc, Legendre, Christophe, Kreis, Henri, Rostaing, Lionel, Kamar, Nassim, Mourad, Georges, Garrigue, Valérie, Bayle, François, H de Ligny, Bruno, Büchler, Mathias, Meier, Carole, Daurès, Jean P., Soulillou, Jean-Paul *and others*. (2010, December). A clinical scoring system highly predictive of long-term kidney graft survival. *Kidney International* **78**(12), 1288–1294.
8. Hariharan, Sundaram, McBride, Maureen A., Cherikh, Wida S., Tolleris, Christine B., Bresnahan, Barbara A. and Johnson, Christopher P. (2002, July). Post-transplant renal function in the first year predicts long-term kidney transplant survival. *Kidney International* **62**(1), 311–318.
9. Heagerty, P. J., Lumley, T. and Pepe, M. S. (2000, June). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**(2), 337–344.
10. Heagerty, Patrick J. and Zheng, Yingye. (2005, March). Survival model predictive accuracy and ROC curves. *Biometrics* **61**(1), 92–105.
11. Hung, Hung and Chiang, Chin-Tsang. (2010, March). Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics* **38**(1), 8–26.
12. Irish, W D, Ilsley, J N, Schnitzler, M A, Feng, S and Brennan, D C. (2010, October). A risk prediction model for delayed graft function in the current era of deceased donor renal transplantation. *American journal of transplantation:*

official journal of the American Society of Transplantation and the American Society of Transplant Surgeons **10**(10), 2279–2286.

13. Janes, Holly, Longton, Gary and Pepe, Margaret. (2009, January). Accommodating Covariates in ROC Analysis. *The Stata Journal* **9**(1), 17–39.
14. Janes, Holly and Pepe, Margaret S. (2008, July). Adjusting for Covariates in Studies of Diagnostic, Screening, or Prognostic Markers: An Old Concept in a New Setting. *American Journal of Epidemiology* **168**(1), 89–97.
15. Janes, Holly and Pepe, Margaret S. (2009, June). Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika* **96**(2), 371–382.
16. Joyner, Michael J. and Paneth, Nigel. (2015, September). Seven Questions for Personalized Medicine. *JAMA* **314**(10), 999–1000.
17. Moskowitz, Chaya S. and Pepe, Margaret S. (2004, May). Quantifying and comparing the accuracy of binary biomarkers when predicting a failure time outcome. *Statistics in Medicine* **23**(10), 1555–1570.
18. Nicol, D., MacDonald, A. S., Lawen, J. and Belitsky, P. (1993, May). Early prediction of renal allograft loss beyond one year. *Transplant International: Official Journal of the European Society for Organ Transplantation* **6**(3), 153–157.
19. Opelz, Gerhard and Döhler, Bernd. (2007). Multicenter analysis of kidney preservation. *Transplantation* **83**(3), 247–253.
20. Pepe, Margaret Sullivan and Cai, Tianxi. (2004, June). The analysis of placement values for evaluating discriminatory measures. *Biometrics* **60**(2), 528–535.
21. Pepe, Margaret Sullivan and Longton, Gary. (2005, September). Standardizing diagnostic markers to evaluate and compare their performance. *Epidemiology (Cambridge, Mass.)* **16**(5), 598–603.
22. Perico, Norberto, Cattaneo, Dario, Sayegh, Mohamed H and Remuzzi, Giuseppe. (2004, November). Delayed graft function in kidney transplantation. *Lancet* **364**(9447), 1814–1827.
23. Rodríguez-Álvarez, María Xosé, Meira-Machado, Luís, Abu-Assi, Emad and Raposeiras-Roubín, Sergio. (2015, October). Nonparametric estimation of time-dependent ROC curves conditional on a continuous covariate. *Statistics in Medicine*.
24. Rosenbaum, Paul R. (1987, June). Model-Based Direct Adjustment. *Journal of the American Statistical Association* **82**(398), 387.
25. Rosenbaum, Paul R. and Rubin, Donald B. (1983, April). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**(1),

41–55.

26. Schemper, M. and Smith, T. L. (1996, August). A note on quantifying follow-up in studies of failure time. *Controlled Clinical Trials* **17**(4), 343–346.
27. Song, Xiao and Zhou, Xiao-Hua. (2008). A semiparametric approach for the covariate specific roc curve with survival outcome. *Statistica Sinica* **18**(947-965), 84.
28. Tapiawala, S. N., Tinckam, K. J., Cardella, C. J., Schiff, J., Cattran, D. C., Cole, E. H. and Kim, S. J. (2010). Delayed graft function and the risk for death with a functioning graft. *J Am Soc Nephrol* **21**(1), 153–61.
29. Uno, Hajime, Cai, Tianxi, Tian, Lu and Wei, L. J. (2007). Evaluating Prediction Rules for t-Year Survivors With Censored Regression Models. *Journal of the American Statistical Association* **102**, 527–537.
30. Yarlagadda, Sri G., Coca, Steven G., Formica, Richard N., Poggio, Emilio D. and Parikh, Chirag R. (2009, March). Association between delayed graft function and allograft and patient survival: a systematic review and meta-analysis. *Nephrology Dialysis Transplantation* **24**(3), 1039–1047.
31. Yarlagadda, S. G., Coca, S. G., Garg, A. X., Doshi, M., Poggio, E., Marcus, R. J. and Parikh, C. R. (2008). Marked variation in the definition and diagnosis of delayed graft function: a systematic review. *Nephrol Dial Transplant* **23**(9), 2995–3003.
32. Zheng, Yingye, Cai, Tianxi, Stanford, Janet L. and Feng, Ziding. (2010, March). Semiparametric models of time-dependent predictive values of prognostic biomarkers. *Biometrics* **66**(1), 50–60.

confounding strength	censoring rate	n	mean AUC			mean bias		relative bias (%)		RMSE	
			crude	IPW	PV	IPW	PV	IPW	PV	IPW	PV
moderate	0.00	500	0.576	0.500	0.500	-0.000	0.000	-0.008	0.001	0.029	0.030
		1000	0.576	0.500	0.500	0.000	0.000	0.021	0.026	0.020	0.021
		2000	0.576	0.500	0.500	-0.000	-0.000	-0.014	-0.023	0.014	0.015
	0.25	500	0.577	0.500		0.000		0.093		0.033	
		1000	0.576	0.500		-0.000		-0.065		0.023	
		2000	0.576	0.500		-0.000		-0.009		0.016	
	0.50	500	0.576	0.500		0.000		0.062		0.041	
		1000	0.576	0.500		-0.000		-0.039		0.029	
		2000	0.576	0.500		-0.000		-0.017		0.020	
strong	0.00	500	0.723	0.500	0.500	0.000	0.000	0.058	0.028	0.042	0.041
		1000	0.722	0.500	0.500	0.000	0.000	0.053	0.006	0.030	0.029
		2000	0.722	0.500	0.500	0.000	0.000	0.042	0.004	0.021	0.021
	0.25	500	0.722	0.501		0.001		0.110		0.049	
		1000	0.722	0.500		-0.000		-0.001		0.035	
		2000	0.722	0.500		0.000		0.048		0.025	
	0.50	500	0.722	0.500		0.000		0.053		0.062	
		1000	0.722	0.500		-0.000		-0.088		0.044	
		2000	0.722	0.500		0.000		0.026		0.031	

Table 1. Results of simulations for a non-informative marker (AUC $= 0.5$) and a single confounder. Three approaches were compared: the crude ROC, the standardized and weighted ROC (IPW) and the ROC curve based on placement values (PV). RMSE: root mean square error.

confounding strength	censoring rate	n	mean AUC			mean bias		relative bias (%)		RMSE	
			crude	IPW	PV	IPW	PV	IPW	PV	IPW	PV
moderate	0.00	500	0.757	0.713	0.718	0.000	0.004	0.020	0.604	0.026	0.028
		1000	0.757	0.714	0.718	0.000	0.005	0.056	0.693	0.019	0.020
		2000	0.757	0.714	0.718	0.000	0.005	0.032	0.680	0.013	0.014
	0.25	500	0.756	0.713		0.000		0.012		0.030	
		1000	0.756	0.714		0.000		0.057		0.021	
		2000	0.756	0.714		0.000		0.068		0.015	
	0.50	500	0.756	0.713		0.000		0.003		0.038	
		1000	0.756	0.714		0.001		0.075		0.027	
		2000	0.756	0.714		0.001		0.078		0.019	
strong	0.00	500	0.849	0.732	0.729	-0.013	-0.017	-1.757	-2.216	0.044	0.042
		1000	0.848	0.734	0.730	-0.011	-0.016	-1.528	-2.102	0.032	0.031
		2000	0.848	0.734	0.730	-0.011	-0.015	-1.457	-2.078	0.024	0.025
	0.25	500	0.847	0.729		-0.017		-2.221		0.051	
		1000	0.848	0.732		-0.013		-1.778		0.037	
		2000	0.847	0.733		-0.012		-1.591		0.028	
	0.50	500	0.848	0.725		-0.020		-2.719		0.062	
		1000	0.847	0.729		-0.016		-2.181		0.046	
		2000	0.847	0.732		-0.013		-1.803		0.034	

Table 2. Results of simulations for a marker associated with the times-to-event and with a single confounder. Three approaches were compared: the crude ROC, the standardized and weighted ROC (IPW) and the ROC curve based on placement values (PV). RMSE: root mean square error. The theoretical values asymptotically computed were 0.7133 and 0.7453 for moderate and strong confounding strength, respectively.

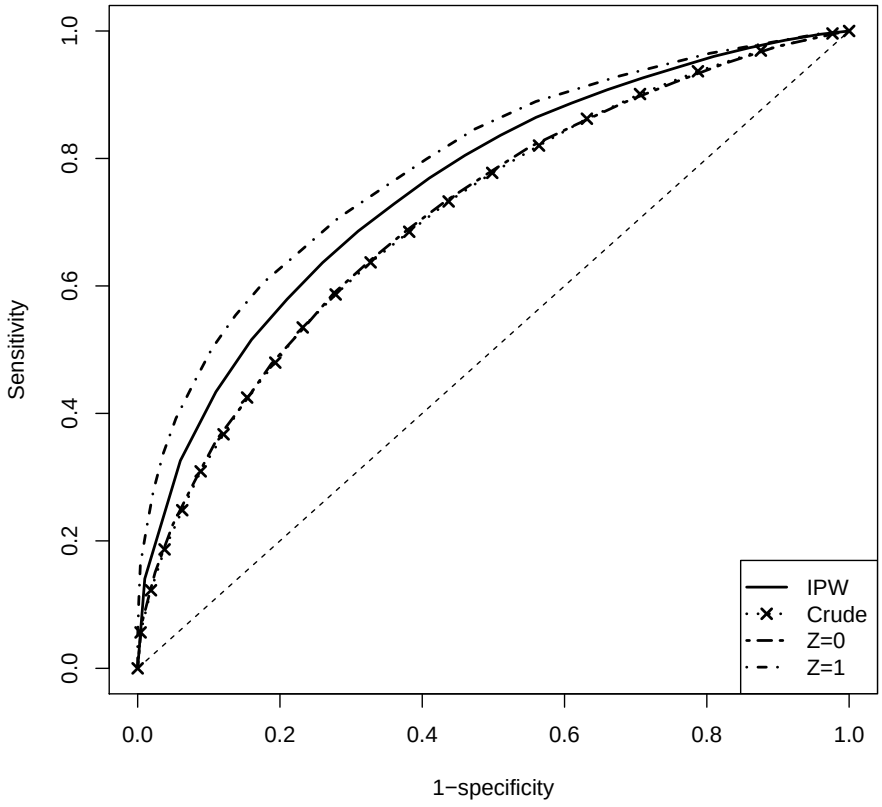


Figure 1. The marker X and the covariate Z are independent, they are both associated with the times-to-event. Four ROCτ curves are computed: the standardized and weighted ROC (IPW) curve, the crude ROC curve and the stratified ROC curve obtained by considering either the subjects in strata $Z = 0$ or the subjects in strata $Z = 1$.

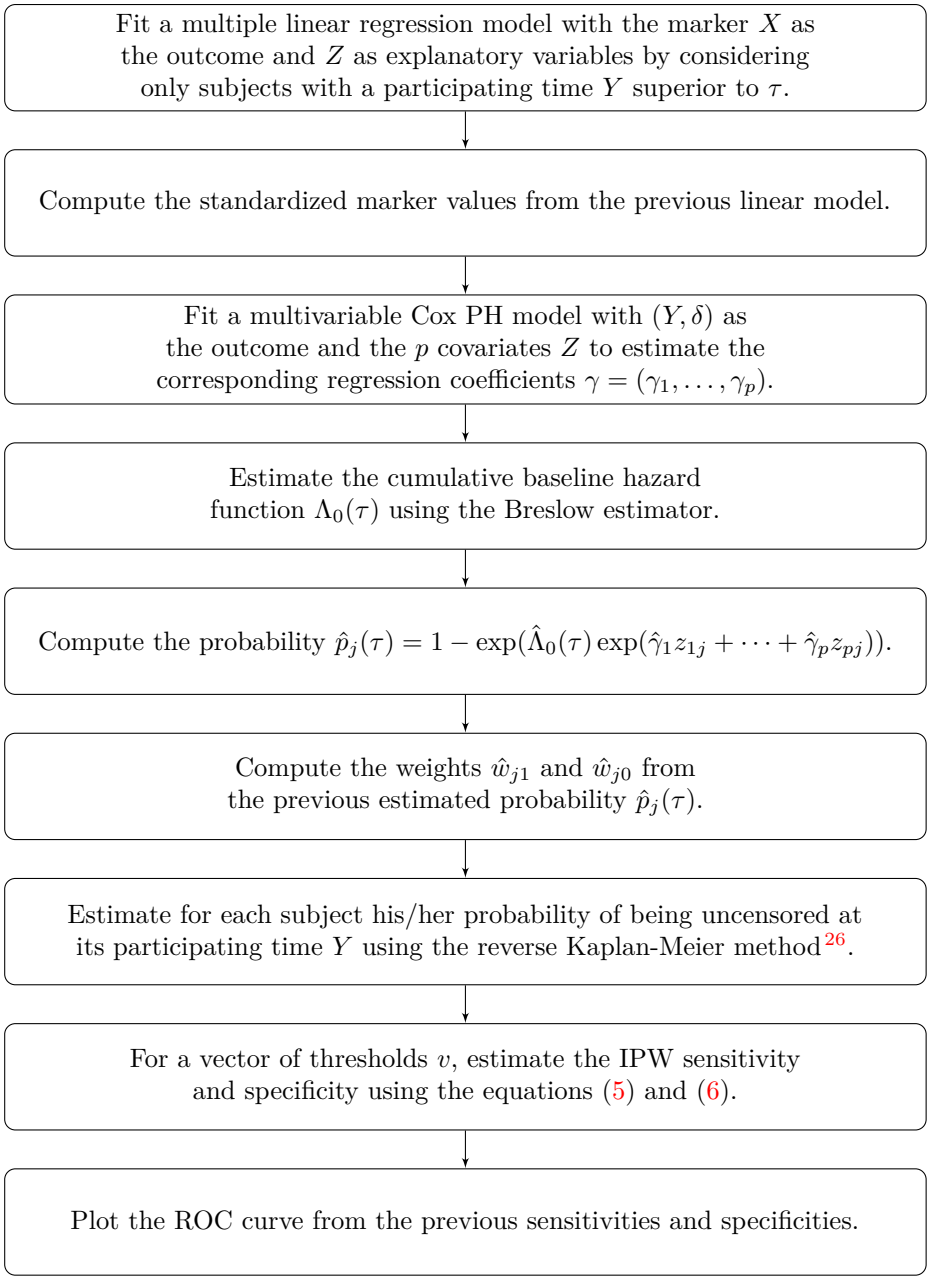


Figure 2. Estimation procedure of the standardized and weighted ROC estimator.

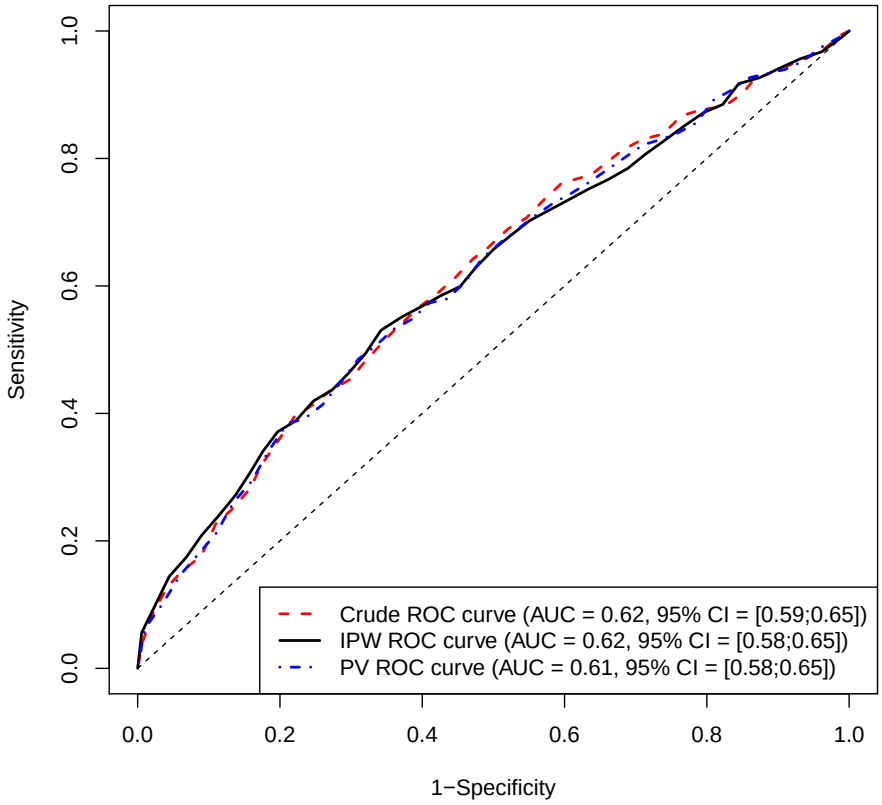


Figure 3. ROC curves estimated by the crude estimator, the standardized and weighted (IPW) estimator and the placement values (PV) estimator for evaluating the capacity of the cold ischemia time (CIT) to discriminate kidney recipients into two groups: those who will not experience delayed graft function (DGF) after kidney transplantation, and those who will experience DGF.

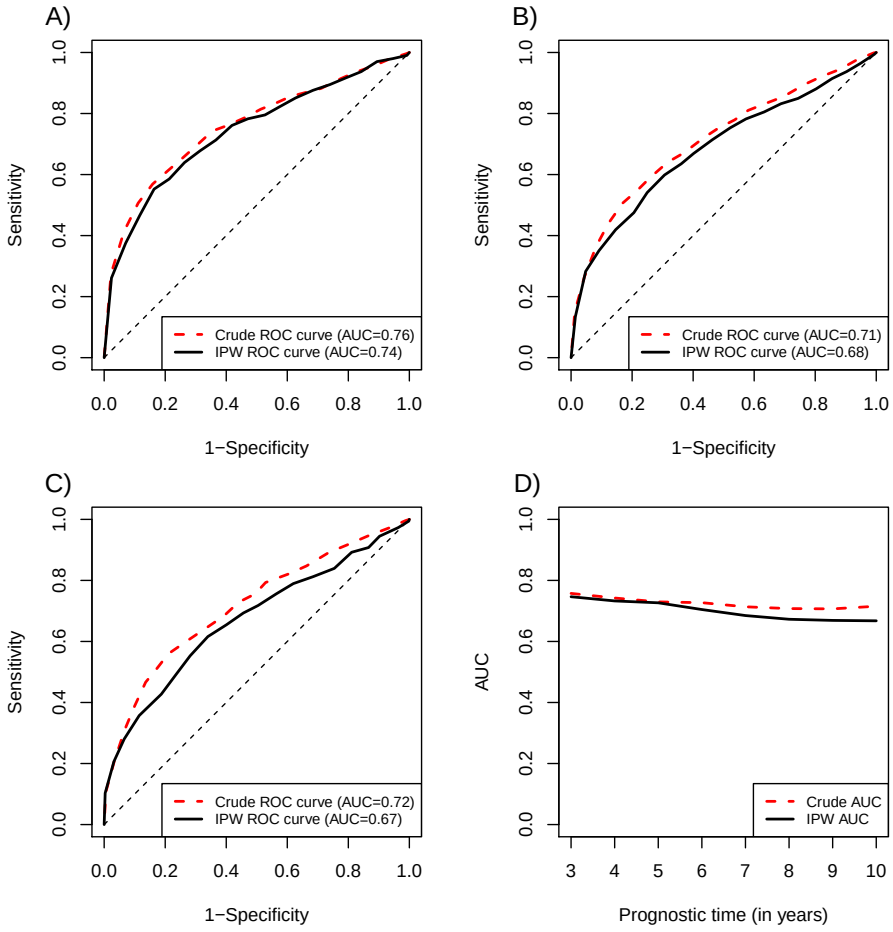


Figure 4. ROC curves estimated by the crude estimator and the standardized and weighted (IPW) estimator for evaluating the capacities of the 12-month serum creatinine measurement (SCr) to predict graft lost of the transplanted recipient up to 3 years (A), 7 years (B) and 10 years (C). The Figure D displays the crude and IPW AUC obtained from different prognostic times varying from 3 to 10 years.

4.2 Annexes du manuscrit

Appendix of the manuscript by Le Borgne et al. entitled "Standardized and weighted time-dependent ROC curves to evaluate the intrinsic prognostic capacities of a marker by taking into account confounding factors"

Statistical Methods in Medical Research
XX(X):1–??

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

A.1 Estimators (5) and (6) when covariates are categorical

In absence of censoring and regardless the proposed correction when weights are close to zero, the IPW estimator of the sensitivity (equation 5) equals to:

$$\widehat{se}_w(\tau, v) = \frac{\sum_{j=1}^{n} \mathbb{1}(T_j \leq \tau, X_j > v)/\hat{P}(T_j \leq \tau|z_j)}{\sum_{j=1}^{n} \mathbb{1}(T_j \leq \tau)/\hat{P}(T_j \leq \tau|z_j)}$$

By only considering categorical confounding factors in Z, let S the covariate defined by the combinations of Z and S_j the observed strata for the subject j $(j = 1, \ldots, n)$:

$$\begin{aligned}
\widehat{se}_w(\tau, v) &= \frac{\sum_s \{\hat{P}(T \leq \tau|S = s)^{-1} \sum_{j|s_j=s} \mathbb{1}(T_j \leq \tau, X_j > v)\}}{\sum_s \{\hat{P}(T \leq \tau|S = s)^{-1} \sum_{j|s_j=s} \mathbb{1}(T_j \leq \tau)\}} \\
&= \frac{\sum_s \{\hat{P}(T \leq \tau|S = s)^{-1} n\hat{P}(S = s)\hat{P}(T \leq \tau, X > v|S = s)\}}{\sum_s \{\hat{P}(T \leq \tau|S = s)^{-1} n\hat{P}(S = s)\hat{P}(T \leq \tau|S = s)\}} \\
&= \sum_s \hat{P}(S = s)\hat{P}(T \leq \tau, X > v|S = s)/\hat{P}(T \leq \tau|S = s)
\end{aligned}$$

Note that $\hat{P}(X > v|T \leq \tau, S = s)$ is an estimator of the crude sensitivity for a threshold v at time τ among subjects in the strata s, noted $\widehat{se}_s(\tau, v)$, therefore:

$$\widehat{se}_w(\tau, v) = \sum_s \hat{P}(S = s)\widehat{se}_s(\tau, v)$$

When Z is composed by categorical covariates, the proposed estimator $\widehat{se}_w(\tau, v)$ (equation 5) is equivalent to the mean of the crude sensitivities per strata weighted by the proportion of patients in the strata.

Similarly in absence of censoring, the IPW estimator of the specificity (equation 6) equals to:

$$\begin{aligned}
\widehat{sp}_w(\tau, v) &= \frac{\sum_{j=1}^{n} \mathbb{1}(T_j > \tau, X_j \leq v)/\hat{P}(T_j > \tau|z_j)}{\sum_{j=1}^{n} \mathbb{1}(T_j > \tau)/\hat{P}(T_j > \tau|z_j)} \\
&= \frac{\sum_s \{\hat{P}(T > \tau|S = s)^{-1} \sum_{j|s_j=s} \mathbb{1}(T_j > \tau, X_j \leq v)\}}{\sum_s \{\hat{P}(T > \tau|S = s)^{-1} \sum_{j|s_j=s} \mathbb{1}(T_j > \tau)\}} \\
&= \frac{\sum_s \{\hat{P}(T > \tau|S = s)^{-1} n\hat{P}(S = s)\hat{P}(T > \tau, X \leq v|S = s)\}}{\sum_s \{\hat{P}(T > \tau|S = s)^{-1} n\hat{P}(S = s)\hat{P}(T > \tau|S = s)\}} \\
&= \sum_s \hat{P}(S = s)\hat{P}(T > \tau, X \leq v|S = s)/\hat{P}(T > \tau|S = s)
\end{aligned}$$

Note that $\hat{P}(X \leq v|T > \tau, S = s)$ is an estimator of the crude specificity for a threshold v at time τ among subjects in the strata s, noted $\widehat{sp}_s(\tau, v)$, therefore:

$$\widehat{sp}_w(\tau, v) = \sum_s \hat{P}(S = s)\widehat{sp}_s(\tau, v)$$

When Z is composed by categorical covariates, the proposed estimator $\widehat{se}_w(\tau, v)$ (equation 6) is equivalent to the mean of the crude sensitivities per strata weighted by the proportion of patients in the strata.

A.2 Conditions under which the estimators (5) and (6) correspond to the crude estimators

In absence of censoring, the crude estimators of the sensitivity and the specificity are:

$$\widehat{se}_c(\tau, v) = \hat{P}(X > v|T \leq \tau)$$
$$\widehat{sp}_c(\tau, v) = \hat{P}(X \leq v|T > \tau)$$

Let X' the marker X standardized according to the covariates among the controls. We have shown in the previous section that when Z is composed by categorical covariates, the IPW estimators of sensitivity and specificity are:

$$\widehat{se}_w(\tau, v) = \sum_s \hat{P}(S = s)\hat{P}(X' > v|T \leq \tau, S = s)$$
$$\widehat{sp}_w(\tau, v) = \sum_s \hat{P}(S = s)\hat{P}(X' \leq v|T > \tau, S = s)$$

where S is the covariate define by the unique combinations of Z. In order to illustrate the proposed estimators, we provide two conditions under which $\widehat{se}_c(\tau, v) = \widehat{se}_w(\tau, v)$. The first condition is $X' = X$, i.e. the marker is independent of the covariates. With this first condition, we have:

$$\widehat{se}_w(\tau, v) = \sum_s \hat{P}(S = s)\hat{P}(X > v|T \leq \tau, S = s)$$

and

$$\widehat{sp}_w(\tau, v) = \sum_s \hat{P}(S = s)\hat{P}(X \leq v|T > \tau, S = s)$$

By using the total probability theorem, the crude estimators of the sensitivity and specificity can be developed as:

$$\widehat{se}_c(\tau, v) = \sum_s \hat{P}(S = s|T \leq \tau)\hat{P}(X > v|T \leq \tau, S = s)$$
$$\widehat{sp}_c(\tau, v) = \sum_s \hat{P}(S = s|T > \tau)\hat{P}(X \leq v|T > \tau, S = s)$$

Therefore, to obtain $\widehat{se}_c(\tau, v) = \widehat{se}_w(\tau, v)$, the second condition is $\hat{P}(S = s|T \leq \tau) = \hat{P}(S = s), \forall s$, i.e. the time-to-event is independent from the covariates. Two conditions under which the estimators (5) and (6) provide an estimation of the crude sensitivity and specificity are therefore sufficient: i) the marker is independent of the covariates and ii) the time-to-event is independent of the covariates.

A.3 Placement values estimator

The placement values (PV) estimator has been proposed by Pepe and Cai[1] in order to take into account covariates in ROC curves when there is no censoring. Let $F_{T_j>\tau}(x) = P(X \leq x|T_j > \tau)$ the cumulative distribution function of X in the controls, i.e. the false positive rate at x. Let $u = 1 - F_{T_j>\tau}(v)$ the placement values of X defined as the proportions of the controls with marker values superior than v, v being all the possible threshold values of X. The ROC curve represents sensitivities plotted against 1 - specificities, i.e. the vector u. Let $F_{T_j\leq\tau}(x) = P(X \leq x|T_j \leq \tau)$ the cumulative distribution function of X in the cases. Then, we have $\text{ROC}(u) = 1 - F_{T_j\leq\tau}(v)$. Because $v = F^{-1}_{T_j>\tau}(1 - u)$ we obtained the following relationship:

$$\text{ROC}(u) = 1 - F_{T_j\leq\tau}(F^{-1}_{T_j>\tau}(1 - u)) \quad (1)$$

According to equation (1), one can take into account the covariates Z by following these steps:

i) Estimate a linear regression with X as the outcome and Z as explanatory variables among controls.

[1]Pepe, Margaret Sullivan and Cai, Tianxi. (2004, June). The analysis of placement values for evaluating discriminatory measures. *Biometrics* **60**(2), 528–535.

ii) Estimate the corresponding empirical distribution function of the standardized residuals.
iii) Compute the standardized residuals of the observations X among the cases by applying the previous linear regression.
iv) Compute the placement values by applying the previous cumulative distribution to the standardized residuals obtained in the cases.
v) Estimate the empirical distribution function of the placement values.
vi) For a vector U from 0 to 1, compute the expected cumulative probabilities of $1-U$ according to the empirical distribution function of the placement values.
vii) The ROC curves represent a plot of u against the previous expected cumulative probabilities of $1-U$.

A.4 Estimation of the distribution of the marker among the controls

The cumulative distribution function of X among the controls conditionally to the covariates, noted $F(x|\tau,z)$, can be *naively* estimated by removing all subjects censored before the time τ.

$$\hat{F}(x|\tau,z)=\frac{n^{-1}\sum_{i=1}^{n}\mathbb{1}(X_i\leq x,Y_i>\tau,Z_i=z)}{\sum_{i=1}^{n}\mathbb{1}(Y_i>\tau,Z_i=z)} \tag{2}$$

Due to the law of large numbers, as $n\to+\infty$, one can develop the "naive" estimator as follows:

$$\begin{aligned}\hat{F}(x|\tau,z)&\to\frac{P(X\leq x,T>\tau,C>\tau,Z=z)}{P(T>\tau,C>\tau,Z=z)} &(3)\\ &=\frac{P(X\leq x,T>\tau,C>\tau,Z=z)}{P(X\leq x,T>\tau,Z=z)}\times\frac{P(X\leq x,T>\tau,Z=z)}{P(T>\tau,Z=z)}\\ &\times\frac{P(T>\tau,Z=z)}{P(T>\tau,C>\tau,Z=z)} &(4)\\ &=P(X\leq x|T>\tau,Z=z)\frac{P(C>\tau|X\leq x,T>\tau,Z=z)}{P(C>\tau|T>\tau,Z=z)} &(5)\end{aligned}$$

This result shows that the cumulative distribution function of the marker among the controls conditionally to the covariates can be estimated by removing all subjects censored before the time τ if censoring is independent of X, Z and T.

A.5 Supplementary simulations-based study with multiple covariates

We simulated data of $n = 1000$ subjects ($j = 1, \ldots, n$) for a setting in which there were 4 covariates Z. First, we simulated Z_1 from Bernoulli distribution with parameter equal to 0.5. Next, Z_2 and Z_3 were successively simulated according to Bernoulli distributions with the probabilities respectively equal to the following linear predictors of logistic regressions: -0.5 + $0.5z_{1j}$ (to obtain an AUC at 0.56) and $0.3z_{1j} + 0.5z_{2j}$ (to obtain an AUC at 0.58). Z_4 was simulated from a normal distribution with mean equal to $0.3z_{1j} + 0.2z_{2j} + 0.3z_{3j}$, and standard deviation equal to 1 (to obtain an R^2 at 0.10). In order to apply the stratified estimator, we discretized Z_4 from 1 to 10 by using deciles. X was generated from a normal distribution with mean equal to $0z_{1j} + \beta_2 z_{2j} + \beta_3 z_{3j} + \beta_4 z_{4j}$ and standard deviation equal to 1. We simulated two scenarios related to the relationship between the covariates and both the marker and the time-to-event. For moderate associations, we aimed to obtain a R^2 at 0.30, and we fixed $\beta_2 = 0.6$, $\beta_3 = 0.5$, $\beta_4 = 0.15$, $\alpha_1 = 0.5$, $\alpha_3 = 0.4$ and $\alpha_4 = 0.02$. For strong associations ($R^2 = 0.60$), we fixed $\beta_2 = 1.2$, $\beta_3 = 1.0$, $\beta_4 = 0.25$, $\alpha_1 = 1.0$, $\alpha_3 = 0.8$ and $\alpha_4 = 0.1$. The times-to-event were generated from a Weibull PH model such as $F(t_j|x_j, z_j) = 1 - \exp[\lambda \exp(\alpha_X x_j + \alpha_1 z_{1j} + 0z_{2j} + \alpha_3 z_{3j} + \alpha_4 z_{4j}) t_j^{\nu}]$. The coefficients associated with the baseline Weibull distribution were chosen in order to obtain a 10-year survival equal to 40% regardless the confounding strength and the informative or non-informative status of the marker. The value of the regression coefficient associated with the marker of interest (α_X) was 0 under the null hypothesis of a non-informative marker (AUC_τ=0.5), and 0.5 under the alternative one ($AUC_\tau > 0.5$). The prognostic time τ was 10 years. The censoring times were independently generated from a Weibull distribution with parameters defined to obtain three different censoring rates: 0, 0.25, 0.5. We also evaluated the impact of the set of covariates included in the model. We considered four set of covariates: i) all the four covariates, ii) the three covariates that are associated with the times-to-event, iii) the three covariates that are associated with the marker, and iv) the two covariates that are associated with both the marker and the times-to-event. For each scenario, we simulated 10 000 datasets.

For the alternative hypothesis, the theoretical values were asymptotically computed by simulating one broad dataset of 10 000 000 subjects, and using the stratified estimator on the 80 strata defined by the unique combinations of the four covariates. Note that we applied the PV method only to validate the

confounding	censoring		mean AUC			mean bias		relative bias (%)		RMSE	
strength	rate	covariates	crude	IPW	PV	IPW	PV	IPW	PV	IPW	PV
		all	0.547	0.500	0.500	0.000	0.000	0.063	0.061	0.020	0.020
	0.00	outcome	0.547	0.500	0.500	0.000	0.000	0.005	0.042	0.019	0.020
		marker	0.547	0.500	0.500	-0.000	-0.000	-0.040	-0.039	0.019	0.020
		both	0.547	0.503	0.503	0.003	0.003	0.634	0.664	0.019	0.020
		all	0.547	0.500		0.000		0.041		0.022	
moderate	0.25	outcome	0.546	0.500		0.000		0.015		0.022	
		marker	0.547	0.500		0.000		0.023		0.022	
		both	0.547	0.503		0.003		0.646		0.022	
		all	0.547	0.500		-0.000		-0.015		0.028	
	0.50	outcome	0.547	0.500		-0.000		-0.004		0.028	
		marker	0.547	0.500		-0.000		-0.013		0.027	
		both	0.546	0.503		0.003		0.558		0.027	
		all	0.653	0.500	0.500	-0.000	-0.000	-0.084	-0.051	0.028	0.027
	0.00	outcome	0.653	0.500	0.502	-0.000	0.002	-0.078	0.341	0.028	0.026
		marker	0.653	0.500	0.500	0.000	0.000	0.059	0.069	0.021	0.023
		both	0.653	0.511	0.513	0.011	0.013	2.276	2.575	0.025	0.026
		all	0.653	0.500		0.000		0.079		0.032	
strong	0.25	outcome	0.653	0.500		-0.000		-0.045		0.032	
		marker	0.653	0.500		0.000		0.017		0.025	
		both	0.653	0.512		0.012		2.397		0.027	
		all	0.653	0.499		-0.001		-0.123		0.040	
	0.50	outcome	0.653	0.500		0.000		0.051		0.041	
		marker	0.653	0.500		0.000		0.056		0.032	
		both	0.653	0.512		0.012		2.407		0.033	

Table S1. Results of simulations for a non-informative marker (AUC = 0.5) and with four confounders. Three approaches were compared: the crude ROC, the standardized and weighted ROC (IPW) and the ROC curve based on placement values (PV). Four set of covariates were considered: all the four covariates (all), the three covariates that are associated with the time-to-event (outcome), the three covariates that are associated with the marker (marker), and the two covariates that are associated with both the marker and the time-to-event (both). For each scenario, 10 000 datasets of 1 000 subjects were simulated. RMSE: root mean square error.

proposed IPW estimators in the particular situation where there is no censoring, the PV method being not applicable in the presence of censoring (cases and controls have to be completely observed).

The results of these simulations are presented in the Tables S1 and S2 for respectively, scenarios under the null hypothesis and scenarios under the alternative hypothesis.

confounding	censoring		mean AUC			mean bias		relative bias (%)		RMSE	
strength	rate	covariates	crude	IPW	PV	IPW	PV	IPW	PV	IPW	PV
moderate	0.00	all	0.751	0.706	0.710	-0.001	0.003	-0.121	0.465	0.019	0.018
		outcome	0.751	0.712	0.716	0.006	0.010	0.848	1.396	0.019	0.021
		marker	0.751	0.699	0.702	-0.008	-0.004	-1.064	-0.559	0.020	0.019
		both	0.751	0.707	0.711	0.001	0.004	0.152	0.634	0.017	0.018
	0.25	all	0.750	0.705		-0.001		-0.149		0.022	
		outcome	0.751	0.713		0.006		0.875		0.022	
		marker	0.750	0.699		-0.008		-1.090		0.022	
		both	0.750	0.707		0.001		0.128		0.020	
	0.50	all	0.750	0.704		-0.003		-0.409		0.027	
		outcome	0.750	0.711		0.005		0.717		0.026	
		marker	0.750	0.698		-0.008		-1.176		0.027	
		both	0.750	0.707		0.001		0.104		0.025	
strong	0.00	all	0.831	0.737	0.722	-0.010	-0.025	-1.391	-3.330	0.029	0.035
		outcome	0.831	0.764	0.751	0.017	0.004	2.284	0.521	0.030	0.023
		marker	0.831	0.709	0.700	-0.038	-0.047	-5.146	-6.322	0.047	0.053
		both	0.831	0.741	0.734	-0.006	-0.013	-0.759	-1.743	0.025	0.025
	0.25	all	0.831	0.736		-0.012		-1.542		0.033	
		outcome	0.831	0.763		0.016		2.079		0.032	
		marker	0.830	0.708		-0.039		-5.231		0.050	
		both	0.831	0.740		-0.007		-0.972		0.029	
	0.50	all	0.831	0.734		-0.014		-1.817		0.042	
		outcome	0.831	0.761		0.014		1.901		0.038	
		marker	0.831	0.707		-0.040		-5.417		0.057	
		both	0.831	0.739		-0.008		-1.027		0.036	

Table S2. Results of simulations for a marker associated with the times-to-event and with four confounders. Three approaches were compared: the crude ROC, the standardized and weighted ROC (IPW) and the ROC curve based on placement values (PV). Four set of covariates were considered: all the four covariates (all), the three covariates that are associated with the time-to-event (outcome), the three covariates that are associated with the marker (marker), and the two covariates that are associated with both the marker and the time-to-event (both). For each scenario, 10 000 datasets of 1 000 subjects were simulated. The theoretical values asymptotically computed were 0.7064 and 0.7471 for moderate and strong confounding strength, respectively. RMSE: root mean square error.

A.6 Supplementary results

	CIT (mean hours ± SD (n))			Percentage of DGF		
	yes	no	p-value	yes	no	p-value
Male recipient	18.9 ± 7.0 (1132)	19.5 ± 6.9 (712)	0.0723	26.0	24.4	0.4954
Recipient age > 55 years	19.1 ± 6.8 (818)	19.2 ± 7.1 (1026)	0.8044	28.1	23.2	0.0184
Recipient BMI > 30 kg.m^{-2}	19.1 ± 6.9 (190)	19.1 ± 7.0 (1644)	0.9313	32.6	24.6	0.0210
Retransplantation	20.7 ± 7.0 (453)	18.6 ± 6.9 (1391)	<0.0001	25.8	25.2	0.8491
History of cardiovascular disease	19.6 ± 7.2 (680)	18.8 ± 6.8 (1164)	0.0185	29.1	23.2	0.0057
History of diabetes	19.1 ± 6.6 (239)	19.1 ± 7.0 (1605)	0.9580	32.6	24.3	0.0073
History of hypertension	19.3 ± 7.0 (1483)	18.6 ± 6.7 (361)	0.1041	25.5	24.9	0.8799
History of dyslipemia	18.9 ± 6.6 (566)	19.2 ± 7.1 (1278)	0.3137	27.9	24.3	0.1080
History of malignancy	19.4 ± 6.9 (197)	19.1 ± 7.0 (1647)	0.5113	28.9	25.0	0.2600
History of hepatitis B or C	20.4 ± 7.0 (115)	19.1 ± 6.9 (1729)	0.0419	33.9	24.8	0.0393
Dialysis before surgery >1.5 years	20.1 ± 7.2 (452)	18.8 ± 6.8 (1386)	0.0011	29.0	24.2	0.0514
Recurrent causal nephropathy	19.2 ± 7.0 (564)	19.1 ± 6.9 (1277)	0.9574	25.0	25.5	0.8834
Male donor	19.0 ± 7.0 (1101)	19.3 ± 6.9 (734)	0.4739	26.7	23.2	0.0978
Donor age > 55 years	19.2 ± 6.8 (838)	19.1 ± 7.1 (999)	0.8556	29.1	22.3	0.0010
Donor serum creatinine >108 μmol/l	19.1 ± 7.0 (386)	19.1 ± 6.9 (1449)	0.9879	35.2	22.8	<0.0001
CV cause of donor death	19.2 ± 6.9 (1058)	19.0 ± 7.0 (783)	0.4862	26.1	24.4	0.4404
Donor Dobutamine treatment	18.5 ± 6.4 (151)	19.3 ± 7.0 (1572)	0.1432	25.2	25.9	0.9226
Donor Epinephrine treatment	18.9 ± 6.8 (377)	19.3 ± 7.0 (1350)	0.2850	29.2	24.9	0.1062
HLA-A-B-DR incompatibilities > 4	18.7 ± 7.3 (191)	19.2 ± 6.9 (1653)	0.3477	30.9	24.7	0.0783
Historical peak of anti-class I PRA > 0%	20.5 ± 6.9 (693)	18.3 ± 6.8 (1151)	<0.0001	27.8	23.9	0.0663
Historical peak of anti-class II PRA > 0%	20.4 ± 6.8 (669)	18.4 ± 6.9 (1175)	<0.0001	26.3	24.9	0.5251
Depleting induction	19.6 ± 7.1 (1019)	18.6 ± 6.7 (825)	0.0024	21.1	30.7	<0.0001

Table S3. Results of univariate analyses between the characteristics of kidney transplant recipients and i) cold ischemia time, and ii) delayed graft function. p-values were obtained by two-tailed t-tests for means comparisons and Chi-squared tests for proportions comparisons. BMI, body mass index; CIT, cold ischemia time; CV, cerebro-vascular; DGF, delayed graft function; HLA, Human Leukocyte Antigen; PRA, panel reactive antibody; SD, standard deviation. Variables in bold characters were retained as covariates for the analysis.

	1-year SCr (mean ± SD (n))			Cox model		
	yes	no	p-value	HR	95% CI	p-value
Male recipient	151.1 ± 56.6 (3631)	123.0 ± 50.3 (2254)	<0.0001	1.01	[0.90,1.13]	0.8570
Recipient age > 55 years	146.4 ± 54.8 (2104)	137.0 ± 56.3 (3781)	<0.0001	1.72	[1.54,1.93]	<0.0001
Pre-transplant BMI > 30 kg.m^{-2}	150.8 ± 56.5 (544)	139.3 ± 55.8 (5305)	<0.0001	1.43	[1.20,1.70]	0.0001
12-month BMI > 30 kg.m^{-2}	146.5 ± 51.4 (662)	139.4 ± 56.3 (4974)	0.0011	1.24	[1.05,1.47]	0.0133
Retransplantation	139.4 ± 58.2 (1177)	140.6 ± 55.4 (4708)	0.5140	1.52	[1.34,1.72]	<0.0001
History of cardiovascular disease	144.5 ± 56.8 (2071)	138.1 ± 55.4 (3814)	<0.0001	1.67	[1.50,1.86]	<0.0001
History of diabetes	144.5 ± 54.0 (715)	139.8 ± 56.2 (5170)	0.0312	1.96	[1.69,2.27]	<0.0001
History of hypertension	140.5 ± 54.6 (4701)	139.8 ± 61.0 (1184)	0.7327	1.05	[0.92,1.21]	0.4719
History of dyslipemia	143.6 ± 54.5 (1691)	139.1 ± 56.5 (4194)	0.0043	1.32	[1.17,1.48]	<0.0001
History of malignancy	143.6 ± 57.2 (498)	140.1 ± 55.8 (5387)	0.1884	1.17	[0.96,1.42]	0.1297
History of hepatitis B or C	141.2 ± 63.0 (418)	140.3 ± 55.4 (5467)	0.7851	1.43	[1.19,1.72]	0.0002
Positive recipient CMV serology	140.1 ± 56.0 (3599)	140.8 ± 55.9 (2242)	0.6423	1.22	[1.09,1.36]	0.0007
Positive recipient EBV serology	140.2 ± 55.8 (5602)	143.6 ± 57.0 (185)	0.4152	0.98	[0.73,1.33]	0.9066
Recurrent causal nephropathy	141.7 ± 56.4 (1889)	139.7 ± 55.7 (3981)	0.2008	0.92	[0.82,1.03]	0.1569
12-month acute rejection	155.6 ± 62.9 (1812)	133.6 ± 51.1 (4073)	<0.0001	1.72	[1.54,1.92]	<0.0001
Male donor	138.1 ± 55.9 (3551)	143.9 ± 55.5 (2314)	0.0001	0.81	[0.73,0.90]	0.0002
Donor age > 55 years	155.9 ± 58.1 (2201)	131.0 ± 52.4 (3670)	<0.0001	1.92	[1.72,2.14]	<0.0001
CV cause of donor death	148.7 ± 59.1 (3139)	130.9 ± 50.6 (2729)	<0.0001	1.42	[1.27,1.59]	<0.0001
ECD	158.0 ± 59.8 (1983)	130.0 ± 51.0 (3652)	<0.0001	2.03	[1.81,2.27]	<0.0001
Positive donor CMV serology	144.1 ± 58.2 (2990)	136.4 ± 53.3 (2883)	<0.0001	1.31	[1.17,1.46]	<0.0001
Positive donor EBV serology	140.4 ± 55.5 (5097)	139.4 ± 60.5 (359)	0.7482	1.12	[0.91,1.38]	0.2947
HLA-A-B-DR incompatibilities > 4	144.5 ± 56.6 (714)	139.7 ± 55.7 (5051)	0.0319	1.11	[0.94,1.30]	0.2163
Preemptive transplantation	137.0 ± 53.8 (433)	140.6 ± 56.1 (5449)	0.1752	0.68	[0.52,0.89]	0.0041
Cold ischemia time > 16 hours	142.8 ± 58.3 (3920)	135.4 ± 50.6 (1959)	<0.0001	1.28	[1.13,1.46]	0.0001
DGF	157.3 ± 64.4 (1885)	132.0 ± 49.3 (3951)	<0.0001	1.67	[1.49,1.86]	<0.0001
Depleting induction	141.2 ± 56.9 (2881)	139.5 ± 55.0 (3004)	0.2566	1.10	[0.99,1.23]	0.0797

Table S4. Results of univariate analyses between the characteristics of kidney transplant recipients and i) the 12-month post-transplant serum creatinine value, and ii) graft survival (death-censored). p-values were obtained by two-tailed t-tests for means comparisons and Wald tests related to the regression coefficient of the Cox models. BMI, body mass index; CI, confidence interval; CMV, cytomegalovirus; CV, cerebro-vascular; DGF, delayed graft function; EBV, Epstein-Barr virus; ECD, expanded criteria donor; HLA, Human Leukocyte Antigen; HR, hazard ratio (yes versus no); SCr, serum creatinine; SD, standard deviation. Variables in bold characters were retained as covariates for the analysis.

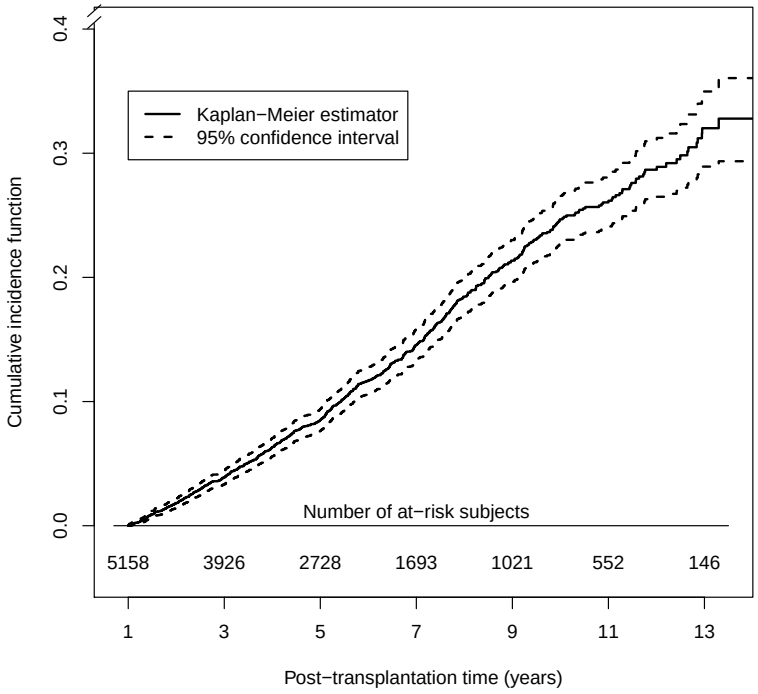


Figure S1. Cumulative incidence function of graft lost related to the 5158 kidney transplant recipients of the DIVAT cohort. The death were right censored.

4.3 Discussions complémentaires

4.3.1 Interprétation de l'AUC standardisée et pondérée

Dans les chapitres précédents nous étions dans une approche étiologique avec l'objectif d'évaluer au mieux la causalité d'une relation entre un facteur de risque supposé et la survenue d'un événement d'intérêt. Dans ce chapitre nous sommes dans un contexte pronostique avec l'objectif d'évaluer les capacités discriminantes d'un biomarqueur (ou d'un marqueur de façon plus générale). L'approche courante dans ce contexte consiste à étudier les capacités discriminantes sans se préoccuper de la causalité et donc à ne pas tenir compte des facteurs de confusion potentiels ou alors, à les combiner avec le marqueur étudié pour obtenir un score avec de meilleures capacités discriminantes (Tripepi et al., 2008). Au contraire, nous proposons un estimateur ajusté des courbes ROC dépendantes du temps mêlant les approches étiologique et pronostique. L'intérêt d'un tel estimateur a été montré par Janes et Pepe (2008) pour prendre en compte un effet centre ou un biais de mesure ou de technique, par exemple.
La mesure de la capacité discriminante du marqueur fournie par notre estimateur est quelque peu théorique et elle peut être difficile à interpréter. En présence de covariables uniquement qualitatives, on peut la formaliser comme la moyenne des AUC obtenues dans chacune des strates formées par les covariables pondérée par la probabilité d'appartenir à la strate. Cette AUC standardisée et pondérée est donc un indicateur de la capacité discriminante du marqueur indépendant des covariables. On parle aussi de capacité discriminante intrinsèque au marqueur. Une lacune de l'article est certainement de ne pas avoir trouvé dans les données dont nous disposions des illustrations plus parlantes. Une application proche de la situation simulée dans la Figure 1 de l'article de Janes et Pepe (2008), avec un effet centre important conduisant, par exemple, à l'obtention d'une courbe ROC non ajustée supérieure aux courbes ROC obtenues pour chaque centre (ou inversement) aurait été plus informative.

4.3.2 Standardisation du marqueur

Les estimateurs pondérés par IPW des sensibilités et spécificités dépendantes du temps proposent d'estimer ces deux quantités dans des pseudo-échantillons de cas et de témoins dans lesquels la distribution des covariables observées est similaire. Nous avons illustré dans les annexes du manuscrit qu'en présence de covariables uniquement qualitatives, ces estimateurs IPW sont équivalents à calculer la moyenne pondérée des estimateurs bruts appliqués chez les individus de chaque strate, les strates étant définies par les combinaisons uniques des différentes modalités des covariables. Cependant, comme nous l'expliquons brièvement dans le manuscrit, lorsqu'une covariable est associée au marqueur, la même valeur seuil peut conduire à des couples sensibilités/spécificités très différents selon les strates. La Figure 4.1 illustre cette situation avec une covariable Z à deux modalités ($Z = 0$ et $Z = 1$, tel que $P(Z = 0) = P(Z = 1) = 0.5$) indépendante de l'événement de sorte que la probabilité d'appartenir à la strate $Z = 0$ (et $Z = 1$) est identique chez les cas et chez les témoins. La partie A) de la Figure présente les deux courbes ROC stratifiées obtenues avec le marqueur dans son échelle naturelle et le point de la courbe correspondant à une valeur seuil de 2,80. On remarque que les courbes obtenues dans chaque strate sont équivalentes, par contre le point de la courbe associé à la valeur seuil de 2,80 est très différent entre les deux strates. Autrement dit, les deux courbes étant égales les mêmes couples sensibilité/spécificité sont obtenus dans chaque strate mais pour des valeurs seuils différentes. Dans ce cas, de par la concavité des courbes ROC, la courbe obtenue à partir des moyennes des sensibilités et spécificités stratifiées pour chaque valeur seuil est inférieure aux courbes stratifiées. On l'observe sur la Figure 4.1 (partie A) où la courbe ROC pondérée estimée par les estimateurs IPW des sensibilités et spécificités, sous-estime les deux courbes obtenues dans chaque strate en passant notamment par l'isobarycentre du segment reliant les points

correspondants à une valeur seuil égale à 2,80 dans les deux strates. L'AUC de la courbe ROC estimée par l'estimateur IPW sera donc inférieure à la moyenne des AUC obtenues dans chaque strate. L'écart sera d'autant plus important que les courbes se rapprochent du coin supérieur gauche.

C'est pour éviter ce biais dû au fait que l'on moyenne des points non alignés ni verticalement ni horizontalement que nous avons proposé de standardiser le marqueur par rapport à sa distribution chez les témoins conditionnellement aux covariables avant de calculer l'estimateur pondéré. En effet, cette standardisation du marqueur permet d'obtenir une distribution du marqueur dans chaque strate similaire et donc des spécificités pour une même valeur seuil très proches dans les différentes strates. La partie B de la Figure 4.1 montre les courbes ROC stratifiées estimées après avoir standardisé le marqueur dans chaque strate. Après standardisation, la valeur seuil de 2,80 devient -0,46 dans une strate et 0,52 dans l'autre strate. Les points de la courbe {1-spécificité, sensibilité} pour ces deux valeurs seuils dans chaque strate sont aussi présentés sur la Figure. On remarque que les points sont alignés verticalement, en conséquence la courbe ROC pondérée obtenue en moyennant les sensibilités et spécificités obtenues pour chaque valeur seuil est aussi très proche des courbes stratifiées.

Standardiser le marqueur par rapport à sa distribution chez les témoins conditionnellement aux covariables avant de calculer les estimateurs des sensibilités et spécificités IPW nous permet donc d'estimer une courbe ROC s'approchant d'une moyenne pondérée par la distribution de Z de courbes ROC stratifiées, identiquement à l'estimateur des *placement values* proposé par Janes et Pepe (2009) dans un contexte diagnostique. Notons cependant que cette standardisation ne garantit pas que l'entière distribution du marqueur chez les témoins soit la même dans toutes les strates. Les spécificités obtenues dans chaque strate pour une même valeur seuil peuvent donc être légèrement différentes et conduire à une faible sous-estimation de la courbe ROC. Cela explique probablement le très faible biais négatif montré par nos simulations pour les scénarios où la confusion est forte.

En présence de données censurées, une difficulté dans l'estimation de la distribution du marqueur chez les témoins réside dans le fait que l'on ne connait pas le statut des individus censurés avant le temps de pronostique τ. Cependant, Blanche et al. (2013) ont montré que l'estimateur "naïf" de la spécificité calculé en omettant les individus censurés avant le temps de pronostique τ est consistent si la censure est indépendante du marqueur X et du temps de survie $\tilde{T}$. De façon analogue, on peut montrer qu'estimer la fonction de répartition du marqueur chez les témoins conditionnellement aux covariables, notée $F(x|\tilde{T} > \tau,z)$, en omettant les individus censurés avant τ est consistent si la censure est indépendante du marqueur X, des covariables Z et du temps de survie $\tilde{T}$.

$$\hat{F}(x|\tilde{T} > \tau,z) = \frac{n^{-1}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x,T_i > \tau,Z_i = z)}{\sum_{i=1}^{n} \mathbb{1}(T_i > \tau,Z_i = z)} \tag{4.1}$$

Selon la loi des grands nombres, l'estimateur de la fonction de répartition du marqueur chez les témoins conditionnellement aux covariables en omettant les individus censurés avant τ tend asymptotiquement vers :

$$\hat{F}(x|\tilde{T} > \tau,z) \rightarrow \frac{P(X \leq x,\tilde{T} > \tau,C > \tau,Z = z)}{P(\tilde{T} > \tau,C > \tau,Z = z)} \tag{4.2}$$

$$= \frac{P(X \leq x,\tilde{T} > \tau,C > \tau,Z = z)}{P(X \leq x,\tilde{T} > \tau,Z = z)} \frac{P(X \leq x,\tilde{T} > \tau,Z = z)}{P(\tilde{T} > \tau,Z = z)} \frac{P(\tilde{T} > \tau,Z = z)}{P(\tilde{T} > \tau,C > \tau,Z = z)} \tag{4.3}$$

$$= P(X \leq x|\tilde{T} > \tau,Z = z) \frac{P(C > \tau|X \leq x,\tilde{T} > \tau,Z = z)}{P(C > \tau|\tilde{T} > \tau,Z = z)} \tag{4.4}$$

Si $C \perp\!\!\!\perp X$, $\tilde{T}$ et Z alors $P(C > \tau|X \leq x,\tilde{T} > \tau,Z = z) = P(C > \tau|\tilde{T} > \tau,Z = z) = P(C > \tau)$ et le second terme de l'équation (4.4) vaut 1. Ce résultat montre que l'estimateur de la fonction de

répartition du marqueur chez les témoins conditionnellement aux covariables en omettant les individus censurés avant τ est un estimateur consistent de la fonction de répartition du marqueur chez les témoins conditionnellement aux covariables si la censure est indépendante du marqueur X, des covariables Z et du temps de survie $\tilde{T}$.

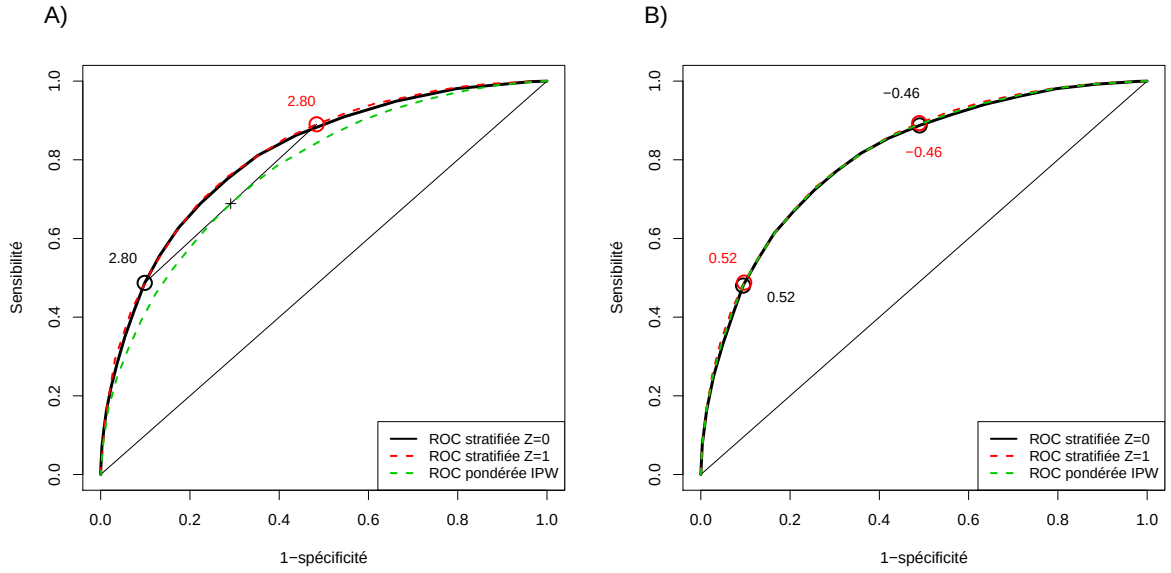


FIGURE 4.1 – Exemple illustratif en présence d'une seule covariable Z à deux modalités. Le marqueur X est associé avec la covariable et avec l'événement. La covariable Z est indépendante de l'événement tel que : $P(T < \tau|Z = 0) = P(T < \tau|Z = 1) = 0.50$. A) Courbes ROC obtenues dans chaque strate sans standardisation du marqueur, le point correspondant à une valeur seuil égale à 2,80 est indiqué pour chaque strate. La courbe ROC pondérée globale obtenue avec les estimateurs IPW des sensibilités et spécificités est également tracée. B) Courbes ROC obtenues dans chaque strate après standardisation du marqueur dans chaque strate par rapport à sa distribution chez les témoins dans la strate, les points correspondants à des valeurs seuils égales à -0,46 et 0,52 sont indiqués pour chaque strate.

4.4 Perspectives

Nous avons évoqué la pénurie de greffons dont pâtissent les patients atteints d'IRCT dans la section 1.2.2. Il est crucial dans ce contexte d'optimiser l'allocation des greffons pour diminuer le nombre de sujets réinscrits sur liste d'attente et limiter la morbidité induite par les retransplantations. Dans cette optique Rao et al. (2009) ont proposé un score donneur nommé *kidney donor risk index* (KDRI) indiquant le profil du greffon rénal et fournissant une estimation du risque d'échec de greffe. Pour évaluer ce score, les auteurs présentent les courbes de survie "ajustées" pour un receveur de 50 ans non diabétique par quintile de KDRI. En effet l'ajustement sur le receveur est primordial lors de l'évaluation d'un tel score donneur car l'allocation des greffons est déjà réalisée à partir d'un score d'attribution qui prend en compte des critères d'équité pour la répartition des greffons et des critères éthiques (comme le différentiel d'âge entre le donneur et le receveur). Évaluer le score donneur sans tenir compte des caractéristiques des receveurs serait donc biaisé car les taux de survie plus faibles chez les receveurs d'un greffon avec un KDRI élevé sont en partie expliqués par les caractéristiques des receveurs. Un travail en cours au sein de notre équipe a pour objectif de valider le KDRI à partir des données Françaises recueillies dans la base DIVAT. En cas de non-validation du score KDRI, il sera envisagé de construire un nouveau score. Dans ce travail l'estimateur des courbes ROC dépendantes du temps standardisées et pondérées sera

intéressant pour évaluer les capacités discriminantes indépendamment des caractéristiques des receveurs du KDRI et éventuellement du nouveau score proposé. De manière plus générale, beaucoup de scores sont créés à partir du prédicteur linéaire d'un modèle de Cox stratifié sur le centre. Pourtant les capacités discriminantes de ces scores sont ensuite évaluées par une courbe ROC ne prenant pas en compte l'effet centre.

Un autre domaine dans lequel nous envisageons d'appliquer l'estimateur des courbes ROC dépendantes du temps standardisées et pondérées est celui de la génomique et du séquençage à haut débit. Ces gènes peuvent avoir des capacités pour pronostiquer la survenue d'un événement de santé mais leurs coûts et leur complexité ne les rendent intéressants que s'ils apportent une information supplémentaire à celle apportée par les facteurs de risque cliniques et environnementaux plus facilement mesurables. L'objectif est donc d'identifier et classer les gènes susceptibles de venir composer, en plus des facteurs de risque cliniques et environnementaux, un score pronostique. Dans leurs travaux, Yu et al. (2012) s'intéressent à cette problématique et proposent trois méthodes pour obtenir des AUC ajustées, c'est à dire pour classer les gènes selon leurs capacités diagnostiques ou pronostiques ajustées sur les facteurs de risque cliniques et environnementaux. Brièvement, les trois méthodes proposées sont les suivantes :

1. Pour chaque gène, estimer un score à partir d'un modèle logistique incluant les facteurs de risque cliniques et environnementaux et le gène d'intérêt. Calcul et classement de l'AUC de chaque score ainsi obtenu.
2. Estimer un score à partir d'un modèle logistique incluant uniquement les facteurs de risque cliniques et environnementaux. Puis, pour chaque gène estimer un nouveau score à partir d'un modèle logistique incluant les facteurs de risque cliniques et environnementaux et le gène d'intérêt mais sans ré-estimer les coefficients de régression associés aux facteurs de risque cliniques et environnementaux. Calcul et classement de l'AUC de chaque score ainsi obtenu.
3. Identique à la méthode précédente hormis la première étape où l'ensemble des gènes disponibles sont inclus dans le modèle logistique utilisé pour calculer un score à partir des facteurs de risque cliniques et environnementaux.

Les méthodes proposées sont discutables, la première pour sa complexité de mise en œuvre avec la nécessité de construire autant de scores que de gènes disponibles, les deux autres pour l'estimation des coefficients associés aux facteurs de risque cliniques et environnementaux à partir d'un modèle différent de celui utilisé pour l'estimation du coefficient associé au gène d'intérêt.

Compte tenu des limites apparentes des méthodes proposées par Yu et al. (2012), nous envisageons de proposer l'utilisation des courbes ROC dépendantes du temps standardisées et pondérées pour classer les gènes selon leur capacité diagnostique/pronostique intrinsèque (i.e. en prenant en compte les facteurs de risque cliniques et environnementaux). Les gènes présentant une bonne AUC ajustée peuvent être intéressant dans la compréhension de la mécanistique de la maladie.

Chapitre 5

Présentation de l'application Plug-Stat®

Sommaire

5.1 Manuscrit en préparation

Résumé de l'article

Motivé par le souhait de fournir de meilleurs outils aux biostatisticiens, épidémiologistes et cliniciens-chercheurs ayant une formation en Biostatistique afin de leur permettre de réaliser des analyses statistiques prenant en compte les facteurs de confusion sans extraire les données (i.e. en lien direct avec la base de données), nous avons développé dans le cadre de la cohorte DIVAT une version personnalisée du logiciel Plug-Stat®. Cette nouvelle version de Plug-Stat® inclut des modèles statistiques adaptés au contexte de la transplantation rénale permettant de prendre en compte les facteurs de confusion. *Via* l'utilisation de ces modèles, l'application permet à l'utilisateur d'étudier l'effet causal d'une exposition qu'il a lui-même définie à partir des variables pré-greffes sur un critère de jugement qu'il choisit parmi ceux proposés. Les variables d'ajustement sont également choisies par l'utilisateur parmi une liste qui regroupe les principaux facteurs confondants de la pathologie. L'objectif est de proposer un outil intuitif et ergonomique qui, une fois qu'il a été adapté à la cohorte, facilite la réalisation d'études épidémiologiques en évitant notamment les étapes longues, coûteuses et sources d'erreurs associées aux phases d'extraction des données, de data-management et de programmation. Dans Plug-Stat®, l'ensemble des analyses statistiques sont réalisées avec le logiciel R (R Development Core Team, 2010), utilisé de façon complètement transparente pour l'utilisateur. Dans ce second article, nous présentons l'outil Plug-Stat®, les différentes méthodes statistiques implémentées, les fonctionnalités offertes par l'application, et nous illustrons son fonctionnement et son utilité à travers une application sur la cohorte DIVAT.

Ci-après est présenté le manuscrit en préparation. Des développements dans l'application Plug-Stat® sont en cours (voir sous-section 5.2.1). Cela explique que certains résultats et certaines figures sont référés dans le texte sous la lettre "X" : ils seront complétés lorsque les développements dans Plug-Stat® seront achevés.

Plug-Stat®: a user friendly web application to query, extract and analyze observational data in clinical epidemiology

Authors: Florent Le Borgne[123], Cyril Loncle[2], Anne Hélène Querard[134] and Yohann Foucher[15*]

Abstract:

Software solutions exist for the collection and the management of data from observational cohorts and others ones for the statistical analysis of these data. Nevertheless, to the best of our knowledge, no software allows to perform statistical analysis directly linked to the database. This lack is detrimental for the valorization of observational cohort: data have to be extracted and managed before statistical analysis, and statistical software are generics, i.e. not specifically adapted for clinical epidemiology and even less for a specific pathology. Both the complexity of the data extraction procedure and the need of advanced statistical skills for the analysis explains partially the current under-exploitation of cohort databases. Here, we present Plug-Stat®, an easy to use web-based application, personalized for each pathology, which allows to query and extract the data but also to perform statistical analyses. To reach this goal, for each new cohort, the first step consists in adapting Plug-Stat® to obtain a personalized application providing methods well-adapted for the main outcomes. Once personalized, Plug-Stat® allows to reduce the time and the cost necessary to the realization of each study and to improve the quality of the results. The French DIVAT cohort of kidney transplant recipients has been chosen to develop and illustrate the use and usefulness of Plug-Stat®.

Keywords: cohort analysis**;** personalized software; web application; clinical epidemiology; biostatistics.

Abbreviations: ANOVA, Analysis of variance; CI, confidence interval; DIVAT, computerized and validated data in transplantation; ECD, expanded criteria donor; HR, hazard ratio; IPW, inverse probability weighting; OR, odds ratio; PS, propensity score; SCD, standard criteria donor; STROBE, STrengthening the Reporting of OBservational studies in Epidemiology

[1] *SPHERE (EA 4275): bioStatistics, Pharmacoepidemiology & Human sciEnces REsearch, University of Nantes, France*
[2] *IDBC/A2com, Espace Antrium Parc de la Teillais 35740 PACE, France*
[3] *Transplantation, Urology and Nephrology Institute (ITUN), Nantes Hospital and University, INSERM U1064, Nantes, France*
[4] *Médecine néphrologie - Hémodialyse, Centre Hospitalier Départemental Vendée Site de La Roche sur Yon, France*
[5] *Nantes Hospital and University, Nantes, France*
[*] *Correspondence to: E-mail: Yohann.Foucher@univ-nantes.fr*

1. Introduction

Over the last decades, the computerization of medical files has provided opportunities to collect numerous clinical and biological parameters into databanks for epidemiologic research. In view of the increasing amount of data available, it appears essential to be able to easily query and exploit these observational data. In this context, we have developed an easy-to-use web application so-called **Plug-Stat®**, allowing to extract data of any computerized data banking systems and to propose appropriate statistical analyses with a direct link with the data bank. The main objective of this innovative web application is to provide a simple and expert tool for studies in clinical epidemiology.

Whatever the pathology, it is noticeable that only few endpoints are studied to evaluate the effect of an exposure. When a new cohort choose Plug-Stat® as software solution, the first step consists in adapting Plug-Stat® to obtain a personalized application providing methods well-adapted for the main outcomes. It is possible through a preliminary collaborative work between the investigators and the biostatisticians in charge of the Plug-Stat® development. For instance in kidney transplantation, we identified that 6 important outcomes allow to respond to the majority of clinical hypotheses: the graft and/or patient survival, the delayed graft function (DGF), the estimated glomerular filtration rate and the acute rejection episode. Using these outcomes, we have been able to implement and adapted biostatistics models into Plug-Stat® to the largest computerized cohort of French transplanted patients (www.divat.fr).

Plug-Stat® is a full-web application running on the main web browsers. It allows to select a sample of patients according to inclusion criteria defined by the user and then to perform various queries such as counting the number of patients, extracting the data, drawing a flowchart, comparing the distribution of a variable between two exposure groups (i.e. two sub-groups of patients among the patients respecting the inclusion criteria), computing and plotting various statistical indicators (means, proportions and survival curves). Beside these usual descriptive methods, Plug-stat® brings novelty for a software solution directly connected with the data bank since it proposes to take into consideration the main difficulties of long-term observational studies: confounding factors, competing events, longitudinal markers, etc. We focused this paper on how account for the confounding factors when estimating the effect of an exposure of interest, which is the main methodological issue in etiologic studies. To deal with confounders, the majority of observational studies remain based on the multivariable models (mainly linear, logistic, or Cox models) allowing an estimation of the conditional effect (i.e. the average effect of treatment or exposure on the individual). However, an increasingly popular solution consists in propensity score (PS) based methods (1–5), allowing a direct estimation of marginal effect (i.e. the average effect of treatment or exposure on the population), as the one tentatively estimated in a randomized clinical trial. The conditional and marginal effects are both meaningful. In consequence, we have chosen to implement the possibility to estimate both effects. Among the different methods based on PS, we have chosen to implement in Plug-stat® the Inverse Probability Weighting (IPW) method that presents the best performances in terms of both bias and statistical power (6–8).

The next Section describes the preliminary step of personalization of Plug-Stat® to the cohort and the pathology. Section 3 presents the main interface and Section 4 gives an overview of the functionalities offered by the application, in particular the methods to obtain confounder-adjusted results. Section 5 proposes an illustrative example in renal transplantation, aiming to demonstrate the usefulness of Plug-Stat® in practice.

2. The preliminary phase for personalized Plug-Stat® to the pathology

When a new cohort chooses Plug-Stat® as software solution, a first collaborative work during four to six months is initiated to personalize Plug-Stat® to the pathology. This work consists in two principal tasks : i) the automatic importation of the data in Plug-Stat® to link the software and the data; and ii) the setting of the main outcomes and the corresponding appropriate statistical models.

To achieve the first task, a module has been developed. It consists in the importation in an Oracle database of the patient table (with one row for each patient) and an unlimited number of files (for instance for treatments, visits, longitudinal biological data, clinical events, etc.). Then it allows to establish the links between the tables, to define for each variable its type (date, numeric or character), label and definition, and to organize the tree view of variables such as it will appear to users in Plug-Stat®. In this module it is also possible to proceed at few data management operations (recode variables for instance) and to parameter checking of consistency in order to prevent the importation of errors. This task is done only once for each cohort, the following data importation will be done automatically as often as required.

The second task consists in studying the most commonly used outcomes and confounding factors in the pathology through a review of the literature and multidisciplinary consultation meetings (with clinicians, researchers, epidemiologists, biostatisticians and data-managers, for instance). Then the corresponding adequate statistical models will be setting.

3. The presentation of the main interface

Plug-Stat® is a full-web application running on the main web browsers (Google Chrome, Internet Explorer, Safari and Firefox). The application is written in Javascript (Ext JS library) for the client part and in VB.net for the server part. The application opens on the main page, as described in Figure 1**,** after authentication with login and password. This main window is composed by six blocks:

A- Tree view of variables. The variables are listed into folders and subfolders. Initially defined during the setting up of the application, this tree view can be modified by each user. It is for instance possible to modify variables or create new variables.

B- List of available operators. After the selection of one variable in the block A, different operators are available. For example, the operator “superior to” is only available for numeric variables.

C- List of available values or modalities. This area allows to define the value(s) or the modalities related to the operator selected in the block B and the variable selected variable in the block A.

D- Definition of the inclusion criteria. When the blocks A, B and C are completed, a condition is created, the results can be TRUE or FALSE. If TRUE the corresponding individuals will be included. By repeating the steps A to C, this area allows combining different conditions with conjunctions AND, OR, or WITHOUT. The combination of these conditions defines the inclusion criteria in order to select the sample of patients to be studied.

E- Definition of the exposure groups. This area allows to define the groups of patients that we want to compare. The creation of these groups is performed by following the steps A to C as for the creation of the inclusion criteria.

F- List of available actions. The first button counts the number of patients according to the previous inclusion criteria and exposure groups. The sixth and second buttons allows respectively, to generate a univariate description between one variable and the exposure groups and to plot survival curves for each exposure groups. The fourth button provides a descriptive table between the exposure groups and different variables selected by the user. The third button gives access to adjusted comparison of means, proportions or survival taking into consideration possible confounding factors. Note that adjusted comparisons are accessible only when two exposure groups are defined. The third button also allows to generate matching pairs of subjects according to selected variables. The fifth button gives access to the specific interface for data extraction.

4. Overview of the functionalities and statistical methods

For all statistical analyses, the software R (10) is used in an invisible manner.

4.1. Description of categorical or quantitative baseline variable

Accessible by the sixth button of the block F (Figure 1) this functionality first requires to select one variable (block A). Then graphical representation is proposed automatically according to the type of the selected variable: boxplot and histogram for quantitative variable and bar graph for qualitative variable. In addition, a descriptive table provides, for the whole sample and for each exposure groups, the numbers and the percentages of each modality of the variable for qualitative variable or the effective, the mean, the standard error, the minimum, the median, the maximum and the first and third quartiles for quantitative variable. Furthermore, if several exposure groups are compared, the adequate statistical test is automatically proposed: a Chi-squared test for qualitative variables (or a Fisher's exact test if at least one of the theoretical frequencies is inferior to 5) and a one-way analysis of variance (ANOVA) for quantitative variables (or a Kruskal–Wallis test if at least one of the groups has less than 30 subjects). The

Figure 2 presents an example of results obtained by the execution of this functionality on a quantitative variable (i.e. the recipient age).

4.2. Description of survival curves

By using the second button of the block F (Figure 1), a dialog window proposes to select the event of interest, the possibility to obtain 95% confidence interval (CI) and the desired annotation (units, labels). The resulting graph combines the survival curves based on the Kaplan-Meier estimator, the number of at-risk subjects and the corresponding log-rank test if several exposure groups are compared. An additional table presents the survival rate at different times.

4.3. Descriptive table

Another functionality available through the fourth button of the block F (Figure 1) is to perform a descriptive table of the characteristics of the patients according the exposure groups. The users are asked to select the variables they want to display in the descriptive table. The comparisons between exposures groups are automatic respecting the same methodology as the one detailed in the section 4.1.

4.4. Confounder-adjusted analysis

In observational studies, subject characteristics may be unbalanced between the exposure groups of interest. To deal with the possible confounding factors, two main estimations can be distinguished: marginal and conditional effects.

For the first one, a solution is the use of weighting methods such as IPW. This method based on PS makes possible the reduction of confounding bias between groups by correcting the contribution of each individual i by a weight equal to:

$$sw_i = \frac{X_i P(X_i = 1)}{P(X_i = 1|Z_i)} + \frac{(1 - X_i)P(X_i = 0)}{P(X_i = 0|Z_i)}$$

Where X_i is the exposure variable of interest ($X_i = 1$ if the individual i is exposed and $X_i = 0$ if unexposed) and Z_i is the vector of covariates. Using these weights allows to estimate the average treatment or exposure effect (ATE), i.e. the average effect of moving an entire population from unexposed to exposed. Another measure of effect which could be of interest is the average treatment or exposure effect on the treated (ATT), i.e. the average effect of treatment or exposure on those subjects who ultimately received the treatment or exposure (9), which could be estimated by using weights equal to:

$$X_i + \frac{(1 - X_i)P(X_i = 1|Z_i)}{P(X_i = 0|Z_i)}$$

With ATT weights the exposed sample is being used as the reference population while with ATE weights the overall sample is being used as the reference population. The weights are easily estimated by logistic regression where quantitative covariates are modeled using natural cubic splines functions (three knots) to allow non-linear effect.

For the second one, multivariable models (linear, logistic, or Cox models, for instance) can be used. Methods based on multivariable modeling differ to the ones based on IPW by the nature of estimated effect. A conditional effect is the average effect, at the individual level, of changing a subject's exposure status from unexposed to exposed.

The three confounder-adjusted comparisons (IPW ATE, IPW ATT, multivariable models) describe below are available through the third button of the block F (Figure 1). When the users choose one of these functionalities, the first step consists to choose between conditional and marginal method, except for adjusted means comparison for which both methods estimate the same effect. If the users choose marginal method, they are then invited to choose between the ATE and the ATT. At each of these steps, explanations on methods are provided to help the users in their choices.

The definition of the covariates to take in consideration is complex: some authors preferentially proposed to consider the variables associated with the outcome (10–12), or with the exposure under interest (13), or with both the outcomes and the exposure under interest (14). This choice also depends if the aim is to estimate the marginal or conditional effect of the exposure. But to be pragmatic, the large majority of data analysts are not aware of such developments. We therefore choose in Plug-Stat® to advise the consideration of the variables associated with the outcome or the exposure, in order to ensure the consideration of all the confounding factors even if the price is probably a small overestimation of the variance effect. The first step consists in a dialog window (Figure 3A) which: i) reminds the definition of a confounding factor with a particular warning for over-adjustment due to covariate in the pathway between the exposure and the outcome, ii) presents a table with crude associations between the outcome and the possible confounding covariates, and iii) allows the selection of covariates in this table by check boxes. The second step consists in a similar dialog window (Figure 3B) but to identify covariates associated with the exposure.

After the choice of the covariates and according the model used, a step displays various indicators in order to verify the assumptions of the model. After validation, the results are provided to the users. As example, for multivariable linear model, graphical representations of the residuals are provided (normal quantile plot and plot of observed versus predicted values) and for method based on IPW a descriptive table of the covariates according the exposure groups in the weighted sample is provided to assess whether the distribution of baseline covariates is similar in the weighted sample between the two exposure groups.

4.5. Presentation of the results interface

The results interface is an html editor in which the results are sequentially inserted in image format. The interface works like a word processor allowing to add text, tables, images, page breaks, etc. It is also possible to change the typography (bold, italic, color, background-color), insert bullets, justify, etc. The users can generate a report that will then be able to save, print or convert to portable document format (PDF). The entire contents of the html editor can be saved in order to be re-opened. In addition, this interface also enables to modify the tables and graphs: titles, axis names, colors, etc. For survival curves it is also possible to move legend and/or log-rank test result, to display or not the number of at risk subject, to choose the minimum and maximum of axis, etc. Finally, one can refresh it with updated data.

4.6. Analyses saving

On the left of the application, a button accessible from all the pages gives access to saving options. The users are invited to select an existing study or create a new one. For each study, three subfolders are automatically created: inclusion criteria, files and reports. The users can save i) the list of both inclusion criteria and exposure groups, ii) the list of variables of the file to extract and iii) the report constitutes by the results (subsection 4.5).

5. An illustrative application

In the context of organ shortage, the United Network for Organ Sharing proposed to increase the pool of donor kidneys by including Expanded Criteria Donor (ECD). A meta-analysis conducted by Querard et al. (15) showed that the ECD classification has been defined for kidney transplant recipients from the United States of America (USA) and despite its use in clinical practice all over the world, it had never been studied on a European cohort. Therefore, we aim to evaluate the marginal effect of the donor type (ECD vs. Standard Criteria Donor (SCD)) on the patient and graft survival. The ATE is meaningless since there is no reason to only performed transplantations from ECD kidneys. We therefore focuses on the ATT. The outcome was the time between transplantation and graft failure, which was the first event occurring between the return to dialysis and patient death.

5.1. Definition and description of the population

The DIVAT cohort contains data of kidney transplant recipients We considered the following inclusion criteria: patients aged from 18 years or older, receiving a first or second transplant and receiving a kidney alone from a heart beating deceased donor. For this study we only included transplantations performed in Nantes hospital. These inclusion criteria are presented in the Figure 1 (block D). Then, we created the two exposure groups under interest. ECD are defined by donors older than 60 years of age or 50-59 years of age with at least two of the following characteristics: history of hypertension, cerebrovascular accident as the cause of death or terminal serum creatinine >1.5 mg/dL (16). These groups are presented in the Figure 1 (block E).

5.2. Comparison of the long-term survival

It is essential to compare survival between ECD and SCD recipients by taking into account the confounding factors. The main one is due to the old-to-old and young-to-young allocation, creating a difference in term of recipient age between ECD and SCD. Thanks to the two steps for covariates selection (subsection 4.4, Figure 3), seven variables were selected: the recipient age, the body mass index, the history of diabetes, the history of malignancy, the history of cardiovascular disease, the retransplantation and the anti HLA class I immunization.

5.3. Results

The results are automatically provided by Plug-Stat®.

The flowchart (Figure X) details the number of patients not included in the analysis. None of the non-included patients has missing data on the variables used for the inclusion criteria. In the case where there are patients not included due to missing data, a table comparing the characteristics between the patients included in the analysis and those not included for missing data is automatically proposed by Plug-Stat®. 948 grafts from SCD kidneys and 752 grafts from ECD kidneys were included in the analysis. A descriptive table (Table 1) presents their baseline characteristics in the whole sample and in the two exposure groups. We observed important imbalance between the two exposure groups: recurrent nephropathy, past history of diabetes, hypertension, cardiovascular disease, hepatitis, dyslipemia and malignancy, male donor, positive donor and recipient CMV serology, retransplantation and depleting induction were more frequent among recipients of ECD kidneys. Compared to recipients of SCD kidneys, recipients of ECD were also older. Then, a summary of follow-up time indicates: the median survival time (13.9 and 8.1 years for SCD and ECD, respectively), the number of observed events (366 and 275 for SCD and ECD, respectively) and the crude survival curves. Finally, the Figure X presents the adjusted survival curves, the *p*-value obtained with the adjusted Log-rank test proposed by Xie and Liu (17), the restricted mean survival time (RMST), the risk difference (RD) and the number needed to treat (NNT). In the context of the ATT, the difference in RMST indicates that the mean individual increase of graft life time was 6.7 (95% CI: -4.31, 18.11) months over the first 12 years post-transplantation if the recipients of ECD kidneys would have received SCD kidneys. The marginal probability of graft failure within 12 years post-transplantation in the population of recipients who actually received a kidney from ECD donor was 0.679 if all subjects had received a kidney from ECD, and 0.549 if all subjects had received a kidney from SCD. Therefore, the absolute RD in the risk of graft failure within 12 years post-transplantation in the population of recipients of ECD kidney was 0.13 (13%), with a 95% CI: 0.033, 0.220. Thus, the NNT is 7.7 (95% CI: 4.53, 29.59). Therefore, one would need to transplant 8 (95%CI: 5, 30) patients in this population with a kidney from SCD donor to avoid one graft failure within 12 years post-transplantation. We observed that after adjustment on the confounding factors by taking the population of patients receiving a graft from ECD as reference (i.e. by using ATT weights), the survival in the group of patients receiving a graft from ECD remains slightly lower than those receiving a graft from SCD, however the gap between confounder-adjusted curves was strongly reduced than between unadjusted curves. The marginal HR was equal to 1.15 (95% CI: [0.84, 1.57], *p*-value = 0.3971), confirms that the difference in survival between the two groups is no longer significant.

6. Discussion

Exploitation and valorization of observational data requires knowledge in data management and biostatistics in addition to the availability of appropriate software. Generally, the phase of data management related to the extraction and the recoding of the data is repeated for each study from the same cohort. The data base is sent to the biostatistician who imports the data into a generalist statistical software and performs the analyses. Very often the statistician is not a specialist of the methods/models specially used in a given pathology. To avoid these long, expensive, and source of error stages, we have developed Plug-Stat®, an easy-to-use web

application. The concept of this application is to provide a tool directly connected with the data base which allows both to extract data and to perform statistical analyses. To reach this goal Plug-Stat® needs to be adapted to each cohort. Once personalized, Plug-Stat® allows to reduce the time and the cost necessary to a study and providing automatically most of the indicators asked in the STROBE guidelines (18). Furthermore, by implementing for each cohort and for each outcome the apropriate statistical model, we hope to reduce the current use of the same statistical models for all the pathologies, i.e. regardless their specificities.

We implemented multivariable modelling and IPW method based on PS. We chose to implement these two methods because they allow to estimate complementary effects. For IPW analysis we proposed to estimate the ATE or the ATT since in observational study there is no reason to expect the ATE and the ATT to coincide (9). The choice of the method should principally depend of the study question but also to data availability and the underlying assumptions of the models (9,19).

In this paper, we focused on the main methodological issue in observational studies, i.e. taking into consideration confounding factors. However, in Plug-Stat others methods have been implemented to deal with the other difficulties frequently encounter in observational cohort studies with a long-term follow-up. For example, we implemented a competing risk model for patients in medical intensive care, especially to correctly estimate the time-to-unassisted breathing and the corresponding risk factors by taking into account the important mortality before extubating. Many studies prefer the use of a composite indicator entitled ventilator-free days, which merges both extubating and death. Nevertheless, this indicator is difficult to interpret and its application is very heterogeneous (20). We also implemented an illness-death model for a cohort of patients with a bioprosthetic aortic valve, the aim being the correct estimation of the cumulative incidence of structural valve deterioration, which represents a severe interval-censored complication. The current literature does not deal with this issue (21), resulting in a dramatic under-estimation of this complication incidence (22). We finally proposed additive and multiplicative risk models for relative survival, particularly meaningful in the context of kidney transplantation (23,24).

The development of Plug-Stat® is an ongoing process to bring new functionalities and take into account user feedback in order to constantly improve the application and simplify its use. Future work will for example bring new statistical methods for prognostic analysis such as time-dependent receiver operating characteristic (ROC) curves or joint models for longitudinal and survival processes.

7. Mode of availability of the application

Plug-Stat® is a property of IDBC/A2com. For additional information, demonstration, or the prospect of acquisition for your cohort data base, please visit the IDBC/A2com website (http://www.idbc.fr/) or contact contact@idbc.fr.

8. Acknowledgments

We wish to thank the professor M. Giral and P. Daguin for their. This work was partially supported by a public grant overseen by the French National Research Agency (ANR) as part of the progam “Laboratoire Commun” (reference: ANR-16-LCV1-0003-01).

9. References

1. Cole SR, Hernán MA. Adjusted survival curves with inverse probability weights. Comput Methods Programs Biomed. 2004 Jul;75(1):45–9.

2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983 Apr;70(1):41–55.

3. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiol Camb Mass. 2000 Sep;11(5):550–60.

4. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol. 2006 May;59(5):437–47.

5. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med. 2015 décembre;34(28):3661–79.

6. Austin PC, Schuster T. The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: A simulation study. Stat Methods Med Res. 2014 Feb 3;

7. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. Stat Med. 2013 Jul 20;32(16):2837–49.

8. Le Borgne F, Giraudeau B, Querard AH, Giral M, Foucher Y. Comparisons of the performance of different statistical tests for time-to-event analysis with confounding factors: practical illustrations in kidney transplantation. Stat Med. 2015 Oct 29;

9. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. Stat Med. 2014 Mar 30;33(7):1242–58.

10. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable Selection for Propensity Score Models. Am J Epidemiol. 2006 Jun 15;163(12):1149–56.

11. Wyss R, Girman CJ, LoCasale RJ, Brookhart AM, Stürmer T. Variable selection for propensity score models when estimating treatment effects on multiple outcomes: a simulation study. Pharmacoepidemiol Drug Saf. 2013 Jan;22(1):77–85.

12. Caruana E, Chevret S, Resche-Rigon M, Pirracchio R. A new weighted balance measure helped to select the variables to be included in a propensity score model. J Clin Epidemiol. 2015 Dec;68(12):1415–22.e2.

13. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. Stat Med. 2007 Jul 20;26(16):3078–94.

14. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Stat Med. 2007 Feb 20;26(4):734–53.

15. Querard A-H, Foucher Y, Combescure C, Dantan E, Larmet D, Lorent M, et al. Comparison of survival outcomes between Expanded Criteria Donor and Standard Criteria Donor kidney transplant recipients: a systematic review and meta-analysis. Transpl Int. 2016 avril;29(4):403–15.

16. Port FK, Bragg-Gresham JL, Metzger RA, Dykstra DM, Gillespie BW, Young EW, et al. Donor characteristics associated with reduced graft survival: an approach to expanding the pool of kidney donors. Transplantation. 2002 Nov 15;74(9):1281–6.

17. Xie J, Liu C. Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. Stat Med. 2005 Oct 30;24(20):3089–110.

18. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. Int J Surg Lond Engl. 2014 Dec;12(12):1495–9.

19. Pirracchio R, Carone M, Rigon MR, Caruana E, Mebazaa A, Chevret S. Propensity score estimators for the average treatment effect and the average treatment effect on the treated may yield very different estimates. Stat Methods Med Res. 2013 Nov 6;

20. Contentin L, Ehrmann S, Giraudeau B. Heterogeneity in the definition of mechanical ventilation duration and ventilator-free days. Am J Respir Crit Care Med. 2014 Apr 15;189(8):998–1002.

21. Sénage T, Le Tourneau T, Foucher Y, Pattier S, Cueff C, Michel M, et al. Early structural valve deterioration of Mitroflow aortic bioprosthesis: mode, incidence, and impact on outcome in a large cohort of patients. Circulation. 2014 Dec 2;130(23):2012–20.

22. Gillaizeau F. Développement et validation de modèles multi-états semi-Markoviens pour le pronostic de patients atteints de maladies chroniques [Internet]. Nantes; 2015 [cited 2016 Aug 3]. Available from: http://www.theses.fr/2015NANT15VS

23. Foucher Y, Akl A, Rousseau V, Trébern-Launay K, Lorent M, Kessler M, et al. An alternative approach to estimate age-related mortality of kidney transplant recipients compared to the general population: results in favor of old-to-old transplantations. Transpl Int Off J Eur Soc Organ Transplant. 2014 Feb;27(2):219–25.

24. Trébern-Launay K, Giral M, Dantal J, Foucher Y. Comparison of the risk factors effects between two populations: two alternative approaches illustrated by the analysis of first and second kidney transplant recipients. BMC Med Res Methodol. 2013;13:102.

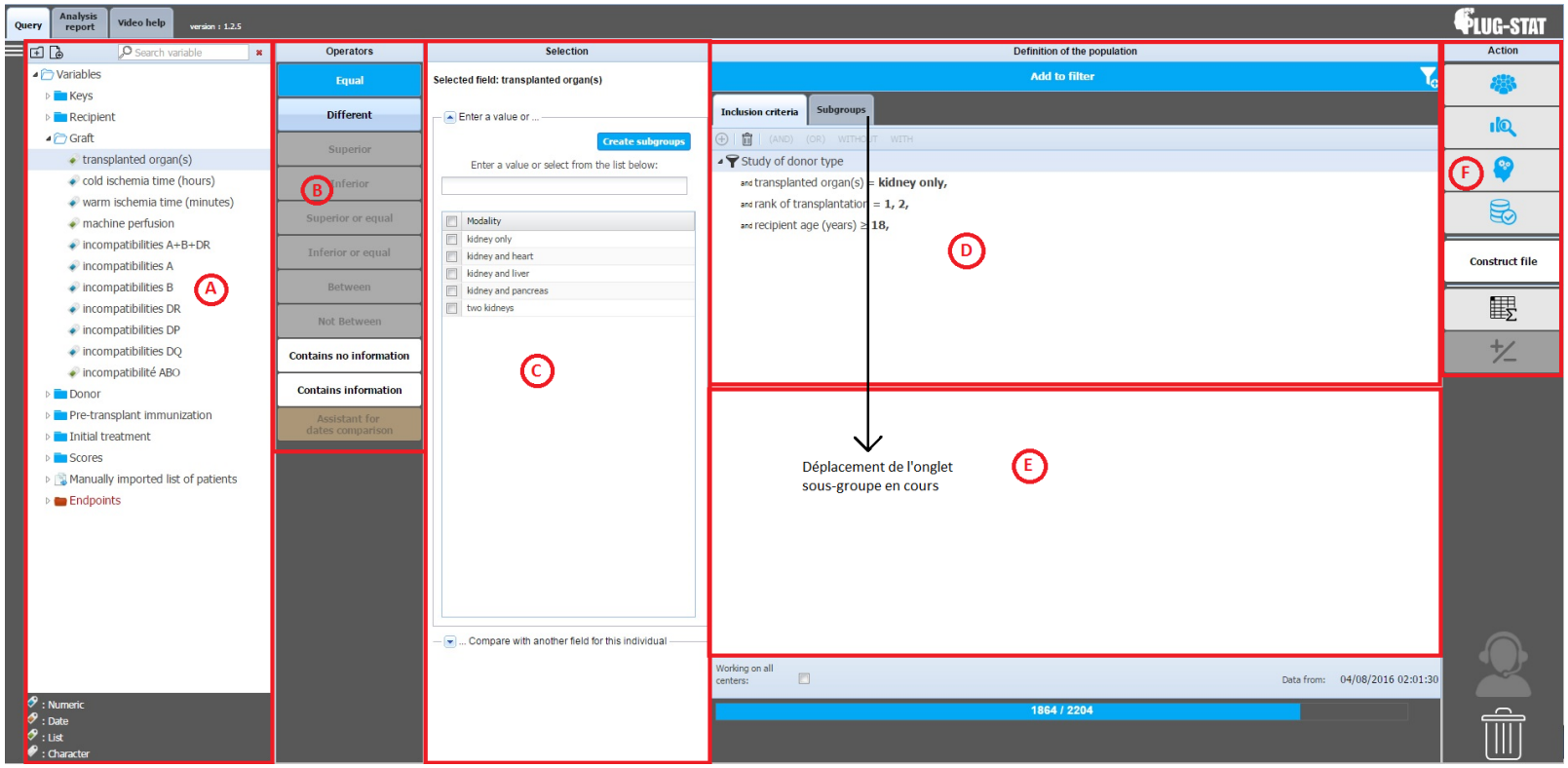
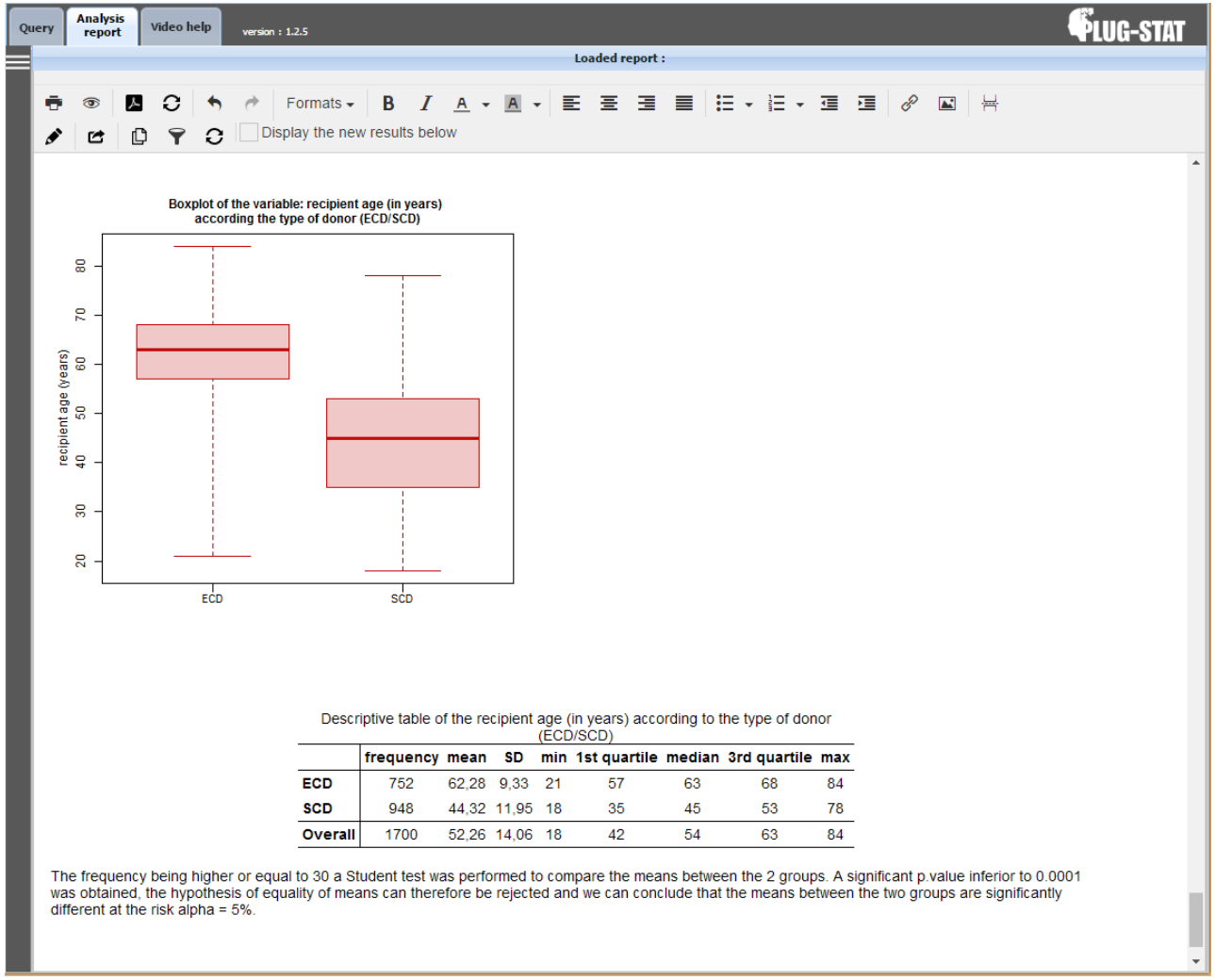


Figure 1: Home interface of the Integr@lis application

A- Tree view of variables
B- List of available operators
C- List of available values or modalities
D- List of inclusion criteria
E- List of available actions to exploit the data



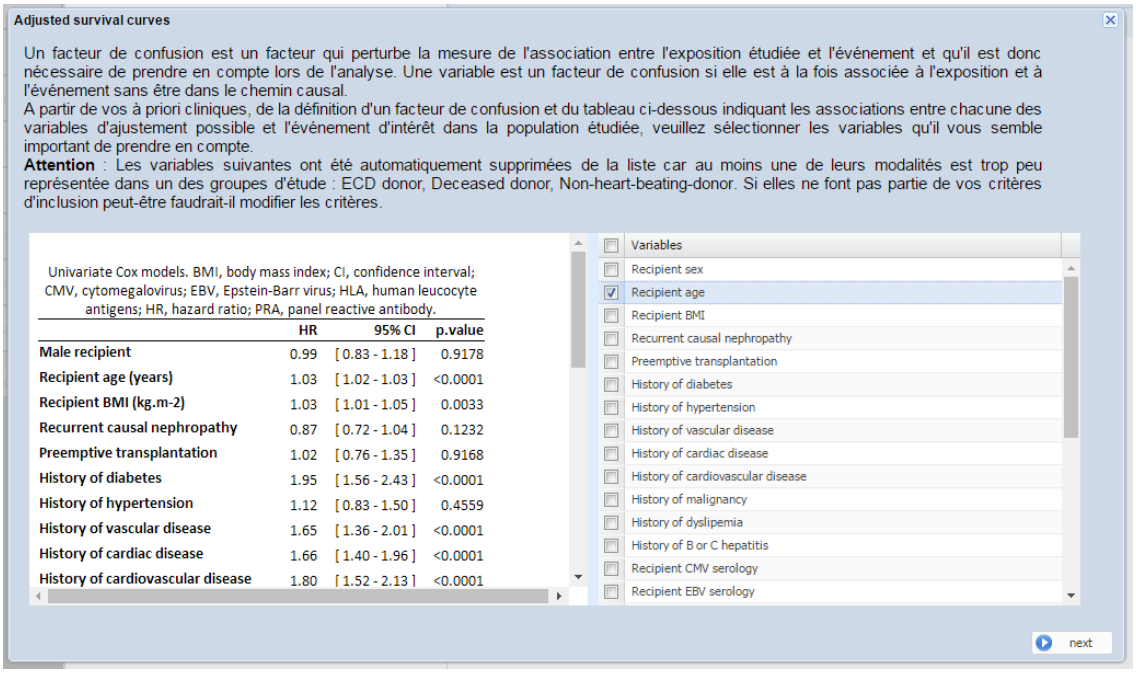
Descriptive table of the recipient age (in years) according to the type of donor (ECD/SCD)

	frequency	mean	SD	min	1st quartile	median	3rd quartile	max
ECD	752	62,28	9,33	21	57	63	68	84
SCD	948	44,32	11,95	18	35	45	53	78
Overall	1700	52,26	14,06	18	42	54	63	84

The frequency being higher or equal to 30 a Student test was performed to compare the means between the 2 groups. A significant p.value inferior to 0.0001 was obtained, the hypothesis of equality of means can therefore be rejected and we can conclude that the means between the two groups are significantly different at the risk alpha = 5%.

Figure 2: Results obtained in the DIVAT cohort by executing the functionality "Description of quantitative baseline variable" on the variable recipient age in the presence of two exposure groups (ECD/SCD)

A



B

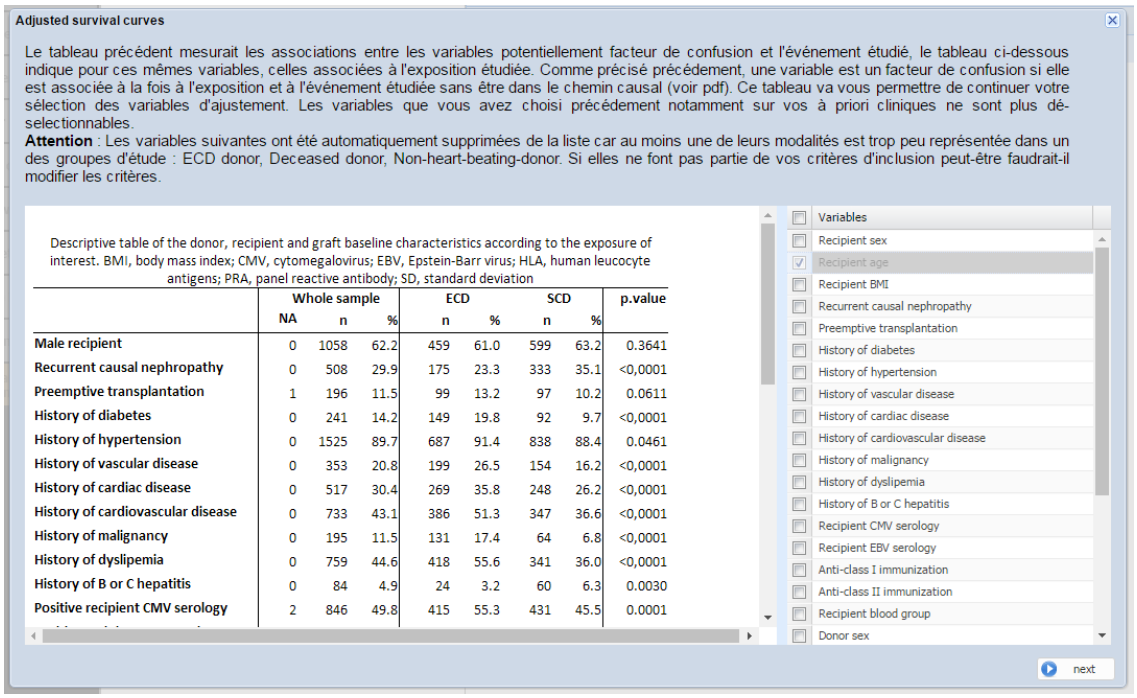


Figure 3: First and second dialog windows to guide the user in the choice of the adjustment factors

5.2 Discussions complémentaires

5.2.1 Avancement des travaux

Le manuscrit présenté dans la section précédente a été relu une première fois par les co-auteurs qui ont suggéré un certain nombre de modifications. Ces modifications ont été apportées au manuscrit qui va de nouveau être relu par les co-auteurs. Certains changements demandés impactent également l'application Plug-Stat®. Les développements sont en cours et expliquent que les Figures présentées à la fin du document ne correspondent pas toujours à ce qui est décrit dans le manuscrit. De plus, certaines Figures et certains résultats sont référés dans le manuscrit sous la lettre "X", ils seront ajoutés lorsque les développements seront achevés. Plus précisément, les méthodes marginales sont implémentées alors que les méthodes conditionnelles sont en cours d'implémentation. Les développements en cours concernent également l'ergonomie de Plug-Stat® et notamment sur les deux étapes que sont le choix de la méthode la plus adaptée à la question de recherche (i.e. conditionnelle ou marginale et ATE ou ATT) et la vérification des hypothèses du modèle.

5.2.2 Personnalisation de Plug-Stat®

Le travail d'adaptation de Plug-Stat® à la cohorte décrit dans la section 2 du manuscrit est une étape cruciale. C'est au cours de cette étape qu'est réalisé le paramétrage qui garantira la justesse des analyses qui pourront par la suite être réalisées dans l'application. Lorsque nous avons réalisé ce travail pour la cohorte DIVAT, nous avons rapidement constaté qu'il était indispensable de créer une nouvelle base de données formatée pour correspondre aux besoins de la recherche épidémiologique et des analyses statistiques. Réalisé lors de réunion de concertation multidisciplinaires (cliniciens, épidémiologistes, biostatisticiens, data-managers) ce travail a permis :

- De restreindre la base de données aux patients potentiellement incluables dans une étude. Par exemple, nous avons exclus les patients mineurs, ceux pour lesquels les données sont entièrement non renseignées, ceux ayant reçu un traitement expérimental dans le cadre d'un protocole, etc.
- De limiter le nombre de variables présentes afin de ne garder que des variables pouvant présenter un intérêt pour une étude de recherche. Plus de 300 variables sont collectées dans DIVAT dont plus de trente uniquement en lien avec l'interprétation des biopsies, au final nous en avons gardé une petite centaine.
- De recoder les variables conservées pour en faire des variables exploitables statistiquement. Par exemple, tous les antécédents des transplantés sont rentrés à l'aide d'une liste déroulante contenant près de 150 possibilités, nous avons transformé cette variable unique générant plusieurs lignes par patient en neuf indicatrices (0/1) correspondantes aux grandes familles d'antécédents (antécédent cardiovasculaire, antécédent d'hypertension artérielle, etc.).
- De distinguer les variables pré-greffes pouvant être utilisées par l'utilisateur pour construire ses groupes d'exposition et ses covariables des variables post-greffes pouvant être utilisées comme critère de jugement.

Indépendamment de la création d'une base de données adaptées aux analyses, c'est aussi lors de ces réunions de concertation que nous avons identifié les principaux critères de jugement qui seront paramétrés afin de pouvoir être utilisés comme critère de jugement dans les analyses réalisées dans Plug-Stat®. Nous avons également identifié les principaux facteurs de confusion que nous proposerons en tant que facteurs d'ajustement lors des analyses.

5.2.3 Choix des méthodes statistiques implémentées

Chronologiquement, les premières méthodes permettant de prendre en compte les facteurs de confusion implémentées dans Plug-Stat® ont été les courbes de survie ajustées par IPW et le test du Log-rank proposé par Xie et Liu (2005). Le choix de ces deux méthodes découle des résultats obtenus dans notre premier travail (Le Borgne et al., 2015). Nous avons vu dans le chapitre précédent que ces méthodes marginales nécessitent l'hypothèse de positivité et qu'elles ne sont par conséquent, pas adaptées à toutes les questions de recherche cliniques. Nous allons donc également implémenter le modèle de Cox multiple afin de proposer une méthode conditionnelle.

Après avoir ajouté la possibilité de réaliser des analyses de survie ajustées dans Plug-Stat®, il était important d'étendre le répertoire des méthodes disponibles dans Plug-Stat® afin de permettre l'étude de critères de jugement binaires ou continus non-censurés.

Pour étudier un critère de jugement binaire en prenant en compte les facteurs de confusion nous avons implémenté la régression logistique univariée pondérée par IPW et nous allons implémenter la régression logistique multiple. Cette dernière a été choisie pour sa capacité à estimer l'ATE ou l'ATT selon le poids utilisé et pour ses bonnes performances. Dans un travail de simulations en cours dans notre équipe, le modèle logistique pondéré par IPW a été comparé à deux autres méthodes permettant l'estimation d'OR marginaux : l'appariement sur le score de propension et une approche basée sur le modèle logistique multiple traditionnel. Ce dernier consiste à retrouver les effets marginaux en prédisant la probabilité de l'événement pour chaque individu constituant les populations ATE ou ATT et en faisant la moyenne. En termes de biais, les meilleurs résultats sont obtenus à partir du modèle logistique multiple et du modèle logistique pondéré. En termes de variances l'approche basée sur le modèle logistique multiple surpasse le modèle logistique pondéré qui est associé à des variances plus fortes et à une puissance statistique plus faible. Les résultats en faveurs du modèle de régression multiple (cette fois-ci utilisé correctement pour estimer les effets marginaux) sont cohérents avec les résultats que nous avions obtenus en présence de données de survie (Le Borgne et al., 2015) et la littérature (Austin, 2013; Austin et Schuster, 2014). Il est donc envisageable dans le futur que nous choisissions d'implémenter un seul modèle de régression multiple, permettant à lui seul d'estimer les effets conditionnels, ATE et ATT.

En ce qui concerne les critères de jugement continus, la régression linéaire multiple est en cours d'implémentation. Contrairement aux OR les coefficients estimés par les modèles linéaires sont *collapsibles*, autrement dit les effets conditionnels et marginaux sont égaux. Il n'y a donc pas d'intérêt à implémenter dans Plug-Stat® deux méthodes. Pour choisir entre la régression linéaire multiple et pondérée par IPW, nous avons réalisé une étude de simulations afin d'étudier les performances en termes de biais, de variances et de risques de première et de seconde espèces. Le plan de cette étude de simulations est repris de travaux précédents (Austin et al., 2007b; Gayat et al., 2012; Austin et al., 2007a; Austin, 2007a). Comme on pouvait s'y attendre les résultats sont proches de ceux obtenus en présence d'un critère de jugement de survie ou binaire. La méthode appariée est associée aux plus mauvais résultats en termes de biais, de racine carrée de l'erreur quadratique moyenne et d'erreurs standards empiriques et asymptotiques. Sous l'hypothèse nulle, le taux d'erreur de première espèce est bien respecté avec le modèle linéaire multiple mais plus fluctuant avec les deux autres méthodes. Sous l'hypothèse alternative, la plus forte puissance statistique est obtenue avec le modèle linéaire multiple. Concernant le choix des covariables à inclure dans le modèle, il semble préférable pour le modèle linéaire multiple d'inclure les covariables liées à l'événement, pour le modèle pondéré d'inclure les covariables associées à l'événement et à l'exposition. A la vue des biais associés à l'appariement, l'utilisation de cette méthode est à déconseiller. Finalement, les meilleurs résultats sont obtenus avec le modèle linéaire multiple, c'est pourquoi nous avons choisi d'implémenter cette méthode dans Plug-Stat®.

5.2.4 Vérification des hypothèses

La vérification des hypothèses est un point important des analyses et certainement la principale difficulté rencontrée lors de l'implémentation de modèles dans une solution logicielle telle que Plug-Stat®. C'est pourquoi nous destinons Plug-Stat® aux biostatisticiens, épidémiologistes, et cliniciens-chercheurs ayant une formation en Biostatistique.

Les premières méthodes à avoir été implémentées dans Plug-Stat® sont des méthodes basées sur le score de propension. Le choix initial d'implémenter ces méthodes avait été en partie dicté par le fait qu'elles facilitent la vérification des hypothèses. En effet, quel que soit le critère de jugement, le score de propension est estimé par un modèle de régression logistique, dont la seule hypothèse est celle de linéarité avec le logit pour les variables quantitatives. La vérification de cette hypothèse par l'utilisateur est difficilement envisageable. Nous avons choisi de modéliser les variables quantitatives avec des fonctions splines, les coefficients de ce modèle logistique n'ayant pas à être interprétés.

Concernant la vérification des hypothèses de positivité et d'équilibre ou de bonne spécification du modèle, une étape de validation est proposée à l'utilisateur après le choix des facteurs d'ajustement et avant l'obtention des résultats. Cette étape propose : i) un graphique montrant la distribution du score de propension dans les deux groupes d'exposition, ii) un tableau descriptif indiquant la distribution des facteurs d'ajustement retenus par l'utilisateur dans les deux échantillons pondérés d'exposés et de non-exposés et les différences standardisées et iii) un paragraphe explicatif visant à aider l'utilisateur dans leurs interprétations (voir Figure 5.1). Enfin, une dernière hypothèse doit être vérifiée lors de la comparaison de courbes de survie avant d'interpréter le test du Log-rank, il s'agit de l'absence d'interaction entre le temps et la différence de survie entre les deux groupes. Cette hypothèse sera vérifiée visuellement par l'utilisateur, le non-respect de l'hypothèse peut entrainer un manque de puissance pour mettre en évidence une différence de survie entre les deux groupes.

Lors de l'implémentation des méthodes multivariées nous garderont le même principe avec une ou plusieurs étape(s) de vérification des hypothèses avant l'obtention des résultats.

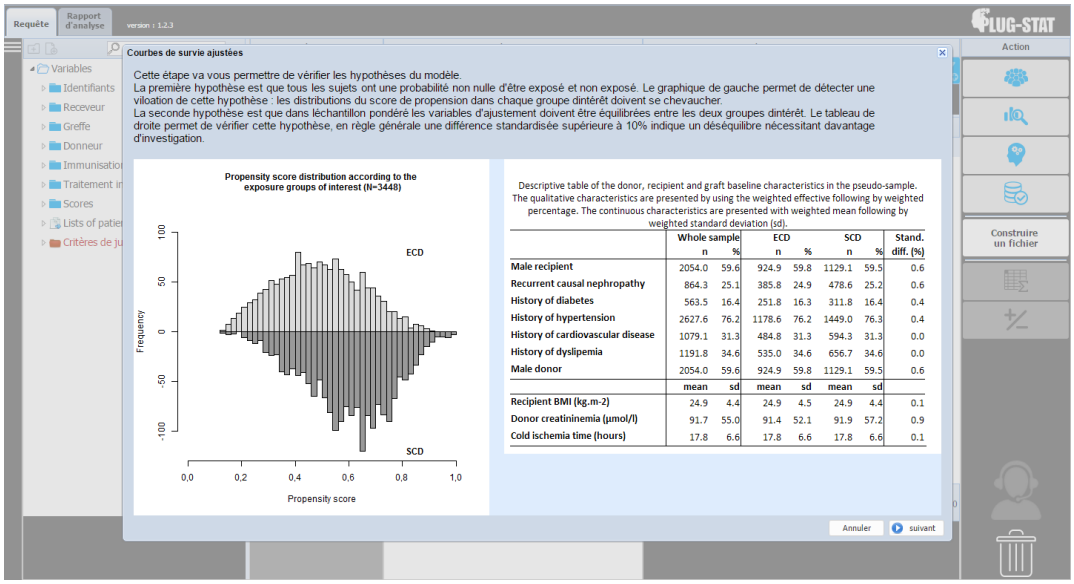


FIGURE 5.1 – Fenêtre de dialogue de Plug-Stat® proposée à l'utilisateur pour la vérification des hypothèses de positivité et de bonne spécification du modèle lors de l'utilisation d'une méthode basée sur le score de propension.

5.2.5 Module d'appariement

A la vue des résultats de nos différentes études de simulations, nous avons choisi de ne pas implémenter les méthodes appariées dans Plug-Stat® pour les études épidémiologiques de grande ampleur. Cependant, l'appariement reste une des approches les plus couramment utilisées en épidémiologie clinique et peut présenter des avantages pour des petits échantillons en recherche biologique fondamentale. Pour cette raison nous avons jugé préjudiciable de totalement écarter l'appariement des fonctionnalités proposées par Plug-Stat®, en particulier pour une cohorte comme DIVAT qui est associée à une biocollection. Nous avons donc choisi d'ajouter à l'outil la possibilité de générer les paires exposé/non-exposé puis de télécharger le fichier de données contenant les paires formées.
Pour la génération des paires, étant donné l'architecture de Plug-Stat® basée sur l'utilisation du logiciel R (R Development Core Team, 2010), deux solutions étaient envisageables : l'utilisation d'un package R existant ou le développement d'un nouveau programme. Afin de choisir entre ces deux possibilités, nous avons proposé dans le cadre du Master 1 Parcours Bioinformatique - Biostatistique de l'Université de Nantes un projet de travail d'étude et de recherche (TER). Trois objectifs étaient fixés pour ce projet : i) concevoir un programme en langage R permettant d'apparier les individus, ii) identifier les packages R existants permettant de répondre à la problématique et iii) comparer les performances du programme développé par rapport aux packages R existants.
Réalisé par trois étudiants ce TER a permis l'obtention d'un programme R permettant de générer des paires de sujets regroupant un sujet exposé et un sujet non-exposé. Volontairement le programme a été conçu pour maximiser le nombre de paires formées au détriment du temps d'exécution et de la distance entre les deux individus d'une même paire. Parmi les packages R existants, le package "Matching" (Sekhon, 2015) a été identifié comme pouvant répondre aux attentes fixées. Des simulations ont donc été réalisées pour comparer les performances des deux solutions possibles. Finalement, les résultats ont conduit à implémenter le programme développé par les étudiants pour son meilleur taux de paires formées et son temps d'exécution tout à fait acceptable bien que supérieur à celui du package "Matching".

5.3 Perspectives

Les perspectives autour de cet outil sont multiples. A ce jour, Plug-Stat® correspond à une version pilote. L'objectif est de continuer son développement dans le cadre d'autres cohortes. A cette fin, nous avons remporté un appel à projet de l'Agence Nationale de la Recherche (ANR) pour la création d'un Laboratoire Commun (LabCom) entre l'Université de Nantes et la société IDBC/A2com (plus de détails sur ces perspectives sont données dans la section 6.2.3).

Chapitre 6

Discussion

6.1 Conclusions

Le thème central de cette thèse a été le développement de Plug-Stat®, logiciel proposé par la société IDBC/A2com pour l'exploitation et la valorisation des données de cohortes observationnelles. Ces développements ont été réalisés dans le cadre de la cohorte DIVAT. Pour nous guider dans le choix des méthodes à implémenter dans Plug-Stat®, nous avons réalisé un premier travail afin de comparer les performances en termes d'erreurs de première et de seconde espèces de différents modèles permettant d'évaluer la causalité d'une exposition sur l'incidence d'un événement en présence de données censurées à droite.

Dans ce travail nous avons montré que parmi les méthodes permettant d'estimer un effet marginal, les meilleurs performances en termes d'erreurs de première et de seconde espèces sont obtenues avec le test du Log-rank ajusté proposé par Xie et Liu (2005). De plus, les courbes de survie ajustées marginales sont intéressantes car elles permettent d'illustrer précisément les différences absolues de survie entre les groupes d'exposition au cours du temps, évitant la perte d'information liée à l'utilisation de HR ajustés. Une erreur a été de comparer des méthodes marginales et conditionnelles alors qu'elles estiment des effets différents. Cependant, nous avons également souligné qu'il semble peu probable que les différences entre les HR théoriques marginaux et conditionnels remettent en cause nos conclusions. Ainsi, nous pensons que le modèle de Cox multivarié demeure le plus performant en termes de risques de première et seconde espèces.

L'ensemble des résultats de ce premier travail nous a conduit au choix d'implémenter dans Plug-Stat® à la fois des méthodes conditionnelles et des marginales afin d'offrir aux utilisateurs la méthode la plus adaptée à leurs questions de recherche. De plus, nous pensons qu'il est intéressant de présenter, lorsque possible, l'estimation des deux effets. L'effet individuel ou conditionnel peut intéresser un clinicien dans le cadre de sa pratique clinique alors que l'effet populationnel ou marginal peut intéresser un décideur dans sa mission de santé publique.

Le développement de Plug-Stat® est en cours et sera poursuivi durant les cinq prochaines années dans le cadre du LabCom RISCA (Recherche en Informatique et en Statistique pour l'Analyse de Cohortes) établissant un partenariat entre l'Université de Nantes et IDBC/A2com. A l'heure actuelle seules les méthodes marginales IPW ont été implémentées et Plug-Stat® n'a été adapté qu'à la cohorte DIVAT. La première étude de recherche clinique réalisée à partir de l'outil va débuter. Il s'agira d'étudier l'effet du type de dialyse (hémodialyse ou dialyse péritonéale) sur la survie du greffon. A terme, nous espérons que Plug-Stat® réduise le temps et le coût nécessaires à la réalisation d'une étude de recherche clinique et améliore la qualité des résultats. Le logiciel permettrait ainsi une meilleure exploitation et valorisation des bases de données de cohortes.

Dans un autre travail s'intéressant aux analyses pronostiques, nous avons rappelé l'importance de prendre en compte les covariables lorsque l'on souhaite évaluer les capacités discriminantes intrinsèques d'un biomarqueur. La prise en compte des covariables dans ce champs d'application reste peu courante, potentiellement due à la rareté et à la relative complexité des méthodes statistiques proposées pour

cela. Afin de tenter de partiellement combler ce manque, nous avons proposé un nouvel estimateur d'une courbe ROC dépendante du temps qui est standardisé et pondéré. Cet estimateur fournit une estimation des capacités discriminantes intrinsèques à un marqueur en tenant compte des facteurs potentiellement confondants. Nous avons également montré qu'une covariable peut influencer la mesure des capacités discriminantes du marqueur dès lors qu'elle est associée au biomarqueur ou à l'événement d'intérêt et qu'il est donc nécessaire de prendre en compte toutes les covariables associées soit au marqueur soit à l'événement. Pour les cohortes en lien avec une biocollection, il sera possible d'implémenter cette méthode dans Plug-Stat®.

En dehors des performances statistiques que nous avons résumées dans la section précédente, les méthodes marginales et conditionnelles présentent chacune des avantages et des inconvénients. Les méthodes marginales ont pour avantage principal leur mimétisme avec les ERC. Cela les rends plus facilement compréhensibles pour des non-spécialistes en Biostatistique et leur confère donc un avantage d'interprétation. Il s'agit également d'un argument convaincant pour les relecteurs et les éditeurs de journaux biomédicaux. Il faut cependant rester vigilant et ne pas sur-interpréter les résultats qui en découlent : les méthodes basées sur le score de propension ne garantissent pas l'équilibre des facteurs non observés entre les groupes d'exposition. Elles n'offrent donc pas un pouvoir de preuve supérieur à celui des méthodes multivariées plus traditionnelles et ne doivent pas être vues comme une alternative aux ERC lorsque ces derniers sont réalisables. Un autre avantage des méthodes basées sur le score de propension est qu'il est plus facile de vérifier la bonne spécification du modèle du score de propension que celle d'un modèle de régression multivarié (Austin, 2011a). La vérification de l'équilibre des covariables entre les groupes d'exposition dans l'échantillon pondéré (ou apparié) permet de facilement vérifier si les déséquilibres ont été éliminés.

L'inconvénient majeur des méthodes basées sur le score de propension réside dans l'estimation de la variance. Pour tenir compte de la nature pondérée ou appariée de l'échantillon, il est recommandé d'utiliser un estimateur robuste de la variance. Cet estimateur permet de maintenir un taux d'erreur de première espèce proche de la valeur nominale au prix d'une variance plus forte et par conséquent d'une plus faible puissance statistique.

Les articles publiés dans des revues médicales qui utilisent une méthode basée sur le score de propension connait une croissance exponentielle depuis quinze ans (Gayat et al., 2010). Cependant, l'application de ces méthodes en recherche médicale reste relativement récente et explique probablement leur utilisation non-optimale et les lacunes observées dans la présentation de la méthodologie utilisée. Plusieurs revues systématiques de la littérature (Austin, 2007b; Gayat et al., 2010; Ali et al., 2015; Lonjon et al., 2016) ont souligné que des points importants étaient fréquemment omis : la vérification de l'équilibre, la stratégie de sélection des covariables, l'algorithme utilisé dans le cadre de l'appariement, etc. L'hétérogénéité des méthodes utilisées (i.e. appariement, IPW, stratification, ajustement sur le score de propension) semble également refléter une méconnaissance des chercheurs qui appliquent ces méthodes. Par exemple, l'ajustement sur le score de propension estime l'effet conditionnel avec des performances inférieures aux modèles multivariés traditionnels (Austin, 2013). Son utilisation ne semble donc présenter aucun avantage, c'est pourtant la méthode la plus utilisée après l'appariement sur le score de propension (Ali et al., 2015; Lonjon et al., 2016). Au contraire, la pondération IPW est la méthode la moins fréquemment utilisée parmi les méthodes basées sur le score de propension. Pourtant, elle est associée aux meilleurs performances et permet d'estimer l'ATE ou l'ATT selon le système de pondération utilisé. De nombreux travaux méthodologiques autour du score de propension ont été publiés durant les quinze dernières années, ils sont certainement à l'origine de la forte popularité actuelle de ces méthodes. Aujourd'hui, le défi majeur est pédagogique afin d'évoluer vers une meilleure utilisation, en pratique, des méthodes basées sur le score de propension. Nous réfléchissons justement dans Plug-Stat® à implémenter directement

les méthodes les plus adaptées, pour éviter ces écueils.

6.2 Perspectives

6.2.1 Application des courbes ROC standardisées et pondérées

Afin d'optimiser l'allocation des greffons aux patients atteints d'IRCT, Rao et al. (2009) ont proposé un score nommé KDRI indiquant le profil du greffon rénal. Ce score a été construit à partir d'un modèle de Cox multivarié incluant des variables du donneur et de la greffe mais aussi des variables du receveur. Bien sûr, seules les variables du donneur et de la greffe composent le score KDRI, mais les variables du receveur ont permis d'obtenir un score ajusté sur ces caractéristiques. Cependant, bien que les auteurs semblent conscient de l'importance de prendre en compte les caractéristiques des receveurs, ils ne les ont pas prises en compte lors de l'évaluation des capacités pronostiques du score. Aussi, le C-index présenté est possiblement biaisé par le fait que le système d'allocation des greffons à conduit à attribuer les greffons avec un KDRI élevé à des receveurs plus à risque. Toutefois, notons que depuis 2015, ce score est utilisé aux États-Unis dans le système d'allocation des greffons. Un travail en cours au sein de notre équipe a pour objectif d'évaluer les capacités pronostiques de ce score sur les données Françaises recueillies dans la base DIVAT. En cas de non-validation du score, il sera envisagé d'en proposer un nouveau. Pour cela, l'estimateur des courbes ROC dépendantes du temps standardisées et pondérées sera utilisé afin d'évaluer les capacités discriminantes du score KDRI indépendamment des caractéristiques des receveurs. Dans ce travail, l'ajustement sur les caractéristiques du receveur sera primordial.
Un autre champs d'application possible pour l'estimateur des courbes ROC standardisées et pondérées est celui de la génomique et du séquençage à haut débit. Les capacités des gènes à pronostiquer la survenue d'un événement de santé sont étudiées et les gènes sont souvent classés selon leurs capacités pronostiques. Dans ce domaine, il est important de prendre en compte les facteurs de risque cliniques et environnementaux car les gènes étant coûteux et complexes à mesurer ils ne sont intéressants que s'ils apportent une information supplémentaire à celle apportée par les facteurs de risque cliniques et environnementaux plus facilement mesurables. Afin d'améliorer la compréhension de la mécanistique de la maladie, identifier les gènes ayant une bonne AUC ajustée présente également un intérêt. Yu et al. (2012) ont proposé trois méthodes pour obtenir des AUC ajustées. Ces méthodes présentent plusieurs limites que nous avons discutées dans la section 4.4. Entre autres, elles sont toutes basées sur la construction d'un score incluant le gène d'intérêt et les facteurs de risque cliniques et environnementaux ce qui ne répond pas à l'objectif d'estimer une AUC ajustée. Nous envisageons de reprendre ces travaux de Yu et al. (2012) afin de montrer que dans ce contexte, l'utilisation des courbes ROC dépendantes du temps standardisées et pondérées est préférable.

6.2.2 Les modèles multivariés pour estimer l'effet marginal

Dans un travail en cours dans notre équipe, nous montrons que l'effet marginal d'une exposition sur un événement binaire non censuré peut être estimé par une approche basée sur le modèle logistique multivarié traditionnel. A partir de ce dernier, on peut en effet prédire pour chaque individu constituant les populations ATE ou ATT sa probabilité de subir l'événement dans les deux situations hypothétiques où : i) il aurait reçu l'exposition, et ii) il n'aurait pas reçu l'exposition. Ensuite, les probabilités marginales de subir l'événement dans les populations ATE ou ATT selon si les sujets ont tous été exposés ou tous été non-exposés sont simplement obtenues en faisant la moyenne des probabilités individuelles prédites. A partir de ces probabilités, on peut estimer l'OR marginal (ATE ou ATT). Dans une étude de simulations, nous avons comparé cette approche au modèle logistique pondéré par IPW et à l'appariement sur le score

de propension. En termes de biais et de variances les meilleurs résultats sont obtenus avec l'approche basée sur le modèle logistique multivarié.
Nous envisageons de continuer nos recherches dans cette thématique afin de proposer une approche basée sur le modèle de Cox multivarié pour estimer le HR marginal, ATE ou ATT. Les performances des différentes méthodes en termes de biais, de variances et de risques de première et seconde espèces devront également être comparées. A terme, il est donc possible que nous choisissions d'implémenter un seul modèle dans Plug-Stat® : le modèle multivarié permettant à lui seul d'estimer les effets conditionnels et marginaux ATE et ATT.

6.2.3 Laboratoire Commun RISCA

Afin de continuer le développement de Plug-Stat® dans le cadre d'autres cohortes, nous avons remporté un appel à projet de l'ANR pour la création d'un LabCom entre l'équipe d'accueil EA4275-SPHERE de l'Université de Nantes et la société IDBC/A2com. Forts de l'expérience acquise à partir de la cohorte DIVAT, l'objectif est d'adapter Plug-Stat® à quatre nouvelles cohortes : ATLANREA (cohorte multicentrique prospective de patients admis en réanimation), CordaBASE (cohorte monocentrique prospective de patients après implantation d'une valve cardiaque bioprothétique), EVALADD (cohorte prospective de patients débutant des soins pour une addiction comportementale), et EKiTE (cohorte multicentrique et européenne de patients transplantés rénaux). Ce LabCom traduit aussi la volonté des deux partenaires de continuer à accroître et pérenniser leur collaboration historique, principalement formalisée dans le cadre de ma thèse CIFRE.
Afin de répondre aux nouveaux défis méthodologiques qui devront être relevés, une feuille de route stratégique sur 5 ans comportant 3 axes de recherche a été établie.

Axe 1 - Développer un module d'importation des données
Lorsqu'un client choisit les 2 solutions proposées par IDBC/A2com, à savoir i) Integr@lis® pour la saisie et le stockage des données et ii) Plug-Stat® pour l'extraction et l'analyse des données, les outils sont conçus pour être naturellement compatibles sans extraction de données. Cependant, beaucoup de cohortes possèdent déjà leur propre solution de saisie et de stockage. Plug-Stat® doit donc être équipé d'un module d'importation de données. Le module développé se présentera sous la forme d'une interface graphique qui permettra de définir le libellé et le type de chaque variable, de regrouper des modalités, d'organiser la liste des variables telle qu'elle apparaitra aux utilisateurs dans Plug-Stat® et de définir l'ensemble des paramètres nécessaires au fonctionnement du logiciel. Cet outil devra aussi permettre à l'utilisateur de définir des contrôles de cohérences, pour alerter ou empêcher l'import d'erreurs dans l'outil d'analyse. Ce module sera utilisé une seule fois pour chaque cohorte, les importations suivantes de données se feront automatiquement à partir du premier paramétrage du module.
La seconde tâche de cet axe, qui sera menée en parallèle du développement du module d'importation, consiste à identifier les principaux critères de jugement et facteurs de confusion associés à la pathologie. Pour chaque cohorte séparément, ces paramètres seront définis au travers d'une revue de la littérature et lors de réunions de concertation multidisciplinaires (cliniciens, chercheurs, épidémiologistes, biostatisticiens et data-managers, par exemple).

Axe 2 - Proposer des modèles statistiques adaptés à chaque cohorte
L'étape principale de cet axe consiste à étudier quels modèles sont les plus adaptés pour chaque cohorte. Cela dépendra notamment des critères de jugement identifiés dans l'axe 1 et des contraintes méthodologiques propres à chaque cohorte. Nous devrons, entre autres, tenir compte pour les cohortes

ATLANREA et CordaBASE des risques compétitifs et pour la cohorte EVALADD de la censure par intervalle. Ces modèles seront implémentés dans Plug-Stat® dans le cadre de l'axe 3.
Les facteurs de confusion sont la principale source de biais dans les études observationnelles. Une réflexion forte autour de leur prise en compte alimentera donc cet axe. Par exemple, les travaux de recherche décrits dans la section 6.2.2 visant à proposer une approche basée sur le modèle multivarié pour estimer un effet marginal, intégrerons cet axe.
Enfin, des travaux d'épidémiologie appliquée seront réalisés. Nous montrerons l'utilité des modèles choisis en les appliquant aux données extraites des cohortes et en répondant à une problématique clinique. Dans sa thèse, Gillaizeau (2015) a montré l'intérêt d'utiliser un modèle *illness-death* semi-Markovien pour estimer l'incidence globale cumulée de la dégénérescence valvulaire (en anglais, *structural valve deterioration*, SVD) chez les patients après implantation d'une valve cardiaque bioprothétique. Ce modèle permet de tenir compte du risque compétitif entre la SVD et le décès ainsi que de la censure par intervalle, le diagnostic de SVD étant posé à partir de critères échocardiographiques. Au final, l'incidence globale cumulée de SVD à 5 ans estimée par ce modèle était de 22,3% (intervalle de confiance à 95% (IC95%) de 16,3% à 28,9%) alors qu'elle était estimée à 8,4% (IC95% de 5,3% à 11,3%) avec l'estimateur de Kaplan-Meier.

Axe 3 - Travailler sur l'ergonomie de Plug-Stat® et y implémenter les modèles statistiques
La finalité du LabCom est de proposer un outil d'analyse performant mais facilement accessible. Cela suppose la conception d'interfaces ergonomiques pour guider l'utilisateur dans la réalisation de son étude (choix du critère de jugement, des facteurs de confusion, vérification des hypothèses, etc.) et contrôler la validité des résultats obtenus. Une tâche constante de cet axe sera d'améliorer l'ergonomie générale de l'application Plug-Stat®. Pour cela nous envisageons de réaliser des enquêtes auprès des utilisateurs ainsi que des tests utilisateurs permettant une mise en situation qui vise à étudier les comportements des utilisateurs face à l'interface. Selon les résultats, les interfaces seront remodelées. Enfin, pour aider les utilisateurs dans leur prise en main de Plug-Stat® et dans la réalisation de leurs études, nous développerons de courts tutoriels vidéos ciblés sur des points précis.
Une autre tâche de cet axe consiste à implémenter dans Plug-Stat® les modèles statistiques adaptés à chaque pathologie. Pour chacun de ces modèles, les interfaces nécessaires seront conçues et programmées par un groupe de travail pluridisciplinaire (biostatisticiens, épidémiologistes, cliniciens, ergonomes, informaticiens). Puis les modèles statistiques seront implémentés de sorte qu'ils puissent être lancés à partir des interfaces développées.

A terme, nous souhaitons que Plug-Stat® réponde aux principales problématiques méthodologiques couramment rencontrées dans les études à partir de données de cohortes observationnelles. Le financement ANR obtenu à hauteur de 300 000€ et l'investissement d'IDBC/A2com à la même hauteur vont nous permettre d'aborder pour la première fois la plupart de ces problématiques dans le cadre des 4 cohortes incluses dans le LabCom et de développer un module d'importation et des interfaces ergonomiques qui seront ensuite ré-utilisés pour chaque nouvelle cohorte. La phase de personnalisation du logiciel restera inévitable pour toutes les futures cohortes qui choisiront Plug-Stat®, mais grâce aux outils développés dans le cadre du LabCom, son coût sera réduit et pourra être supporté par chaque cohorte.

Bibliographie

Agence Nationale d'Accréditation et d'Évaluation en Santé. Diagnostic de l'insuffisance rénale chronique chez l'adulte. Recommandations. Technical report, Paris, 2002.

Akritas M. G. Nearest Neighbor Estimation of a Bivariate Distribution Under Random Censoring. *The Annals of Statistics*, 22(3) :1299–1327, Sept. 1994. ISSN 0090-5364, 2168-8966. doi : 10.1214/aos/1176325630. URL http://projecteuclid.org/euclid.aos/1176325630.

Ali M. S., Groenwold R. H. H., Belitser S. V., Pestman W. R., Hoes A. W., Roes K. C. B., Boer A. d., et Klungel O. H. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal : a systematic review. *Journal of Clinical Epidemiology*, 68(2) :112–121, Feb. 2015. ISSN 1878-5921. doi : 10.1016/j.jclinepi.2014.08.011.

Austin P. C. The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine*, 26(16) :3078–3094, July 2007a. ISSN 0277-6715. doi : 10.1002/sim.2781. 00129.

Austin P. C. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006 : a systematic review and suggestions for improvement. *The Journal of Thoracic and Cardiovascular Surgery*, 134(5) :1128–1135, Nov. 2007b. ISSN 1097-685X. doi : 10.1016/j.jtcvs.2007.07.021.

Austin P. C. Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model. *Journal of Clinical Epidemiology*, 63(1) :2–6, Jan. 2010a. ISSN 1878-5921. doi : 10.1016/j.jclinepi.2008.11.004.

Austin P. C. Absolute risk reductions and numbers needed to treat can be obtained from adjusted survival models for time-to-event outcomes. *Journal of Clinical Epidemiology*, 63(1) :46–55, Jan. 2010b. ISSN 1878-5921. doi : 10.1016/j.jclinepi.2009.03.012.

Austin P. C. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine*, 29(20) :2137–2148, Sept. 2010c. ISSN 1097-0258. doi : 10.1002/sim.3854.

Austin P. C. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res*, 46(3) :399–424, May 2011a. ISSN 0027-3171. doi : 10.1080/00273171.2011.568786.

Austin P. C. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat*, 10(2) :150–161, Mar. 2011b. ISSN 1539-1604. doi : 10.1002/pst.433.

Austin P. C. A Tutorial and Case Study in Propensity Score Analysis : An Application to Estimating the Effect of In-Hospital Smoking Cessation Counseling on Mortality. *Multivariate Behav Res*, 46(1) : 119–151, Jan. 2011c. ISSN 0027-3171. doi : 10.1080/00273171.2011.540480.

Austin P. C. The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in Medicine*, 32(16) :2837–2849, July 2013. ISSN 1097-0258. doi : 10.1002/sim.5705.

Austin P. C. The use of propensity score methods with survival or time-to-event outcomes : reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine*, 33(7) : 1242–1258, Mar. 2014. ISSN 1097-0258. doi : 10.1002/sim.5984. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.5984/abstract.

Austin P. C. et Schuster T. The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes : A simulation study. *Stat Methods Med Res*, Feb. 2014. ISSN 1477-0334. doi : 10.1177/0962280213519716.

Austin P. C. et Stuart E. A. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34(28) :3661–3679, 2015. ISSN 1097-0258. doi : 10.1002/sim.6607. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.6607/abstract. 00001.

Austin P. C., Grootendorst P., et Anderson G. M. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects : a Monte Carlo study. *Stat Med*, 26(4) :734–753, Feb. 2007a. ISSN 0277-6715. doi : 10.1002/sim.2580.

Austin P. C., Grootendorst P., Normand S.-L. T., et Anderson G. M. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect : a Monte Carlo study. *Statistics in Medicine*, 26(4) :754–768, Feb. 2007b. ISSN 0277-6715. doi : 10.1002/sim.2618.

Bender R. et Blettner M. Calculating the "number needed to be exposed" with adjustment for confounding variables in epidemiological studies. *Journal of Clinical Epidemiology*, 55(5) :525–530, May 2002. ISSN 0895-4356.

Binder D. A. Fitting Cox's Proportional Hazards Models from Survey Data. *Biometrika*, 79(1) :139–147, 1992. ISSN 0006-3444. doi : 10.2307/2337154. URL http://www.jstor.org/stable/2337154.

Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ (Clinical research ed.)*, 312(7040) :1215–1218, May 1996. ISSN 0959-8138.

Blanche P., Dartigues J.-F., et Jacqmin-Gadda H. Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biom J*, 55(5) :687–704, Sept. 2013. ISSN 1521-4036. doi : 10.1002/bimj.201200045.

Blotière P.-O., Tuppin P., Weill A., Ricordeau P., et Allemand H. [The cost of dialysis and kidney transplantation in France in 2007, impact of an increase of peritoneal dialysis and transplantation]. *Néphrologie & Thérapeutique*, 6(4) :240–247, July 2010. ISSN 1872-9177. doi : 10.1016/j.nephro.2010.04.005.

Bouyer J. *Épidémiologie : principes et méthodes quantitatives*. Lavoisier, 2009. ISBN 978-2-7430-1167-3.

Briançon S., Jacquelinet C., Merle S., et Lassalle M. Prévalence 2013 de l'IRCT. In *Rapport annuel REIN*, chapter 3, pages 86–126. 2013.

Brookhart M. A., Schneeweiss S., Rothman K. J., Glynn R. J., Avorn J., et Stürmer T. Variable Selection for Propensity Score Models. *Am. J. Epidemiol.*, 163(12) :1149–1156, June 2006. ISSN 0002-9262, 1476-6256. doi : 10.1093/aje/kwj149. URL http://aje.oxfordjournals.org/content/163/12/1149.

Cai T., Pepe M. S., Zheng Y., Lumley T., et Jenny N. S. The sensitivity and specificity of markers for event times. *Biostat*, 7(2) :182–197, Apr. 2006. ISSN 1465-4644, 1468-4357. doi : 10.1093/biostatistics/kxi047. URL http://biostatistics.oxfordjournals.org/content/7/2/182.

Cameron J. I., Whiteside C., Katz J., et Devins G. M. Differences in quality of life across renal replacement therapies : a meta-analytic comparison. *American Journal of Kidney Diseases : The Official Journal of the National Kidney Foundation*, 35(4) :629–637, Apr. 2000. ISSN 1523-6838.

Chambless L. E. et Diao G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in Medicine*, 25(20) :3474–3486, Oct. 2006. ISSN 1097-0258. doi : 10.1002/sim.2299. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.2299/abstract.

Cole S. R. et Hernán M. A. Adjusted survival curves with inverse probability weights. *Comput Methods Programs Biomed*, 75(1) :45–49, July 2004. ISSN 0169-2607. doi : 10.1016/j.cmpb.2003.10.004.

Cole S. R. et Hernán M. A. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6) :656–664, Sept. 2008. ISSN 1476-6256. doi : 10.1093/aje/kwn164.

Cox D. Regression Models and Life-Tables (with discussion). *Journal of the Royal Statistical Society*, B(24) :187–220, 1972.

Dodd L. E. et Pepe M. S. Semiparametric Regression for the Area under the Receiver Operating Characteristic Curve. *Journal of the American Statistical Association*, 98(462) :409–417, 2003. ISSN 0162-1459. URL http://www.jstor.org/stable/30045250.

Gail M. H., Wieand S., et Piantadosi S. Biased Estimates of Treatment Effect in Randomized Experiments with Nonlinear Regressions and Omitted Covariates. *Biometrika*, 71(3) :431–444, 1984. ISSN 0006-3444. doi : 10.2307/2336553. URL http://www.jstor.org/stable/2336553. 00465.

Gayat E., Pirracchio R., Resche-Rigon M., Mebazaa A., Mary J.-Y., et Porcher R. Propensity scores in intensive care and anaesthesiology literature : a systematic review. *Intensive Care Medicine*, 36 (12) :1993–2003, Aug. 2010. ISSN 0342-4642, 1432-1238. doi : 10.1007/s00134-010-1991-5. URL http://link.springer.com.gate2.inist.fr/article/10.1007/s00134-010-1991-5. 00055.

Gayat E., Resche-Rigon M., Mary J.-Y., et Porcher R. Propensity score applied to survival data analysis through proportional hazards models : a Monte Carlo study. *Pharmaceutical Statistics*, 11(3) :222–229, June 2012. ISSN 1539-1612. doi : 10.1002/pst.537.

Gillaizeau F. *Développement et validation de modèles multi-états semi-Markoviens pour le pronostic de patients atteints de maladies chroniques*. PhD thesis, Université de Nantes, 2015.

Goehrs J.-M., Borel T., Costagliola D., et les participants à la table ronde N 6 de Giens XXVII. La mise en place des cohortes en France : pourquoi, pour qui, comment et avec quels moyens ? *Thérapie*, 67(4) :375–380, Aug. 2012. ISSN 0040-5957. doi : 10.2515/therapie/2012049.

Goldberg M. et Zins M. Les cohortes généralistes en population : l'exemple des cohortes Gazel et Constances. *Bulletin de l'académie nationale de médecine*, 197(2) :299–313, 2013.

Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology*, 125(5) :761–768, May 1987. ISSN 0002-9262.

Groupe de travail de la Société de Néphrologie. [Evaluation of glomerular filtration rate and proteinuria for the diagnosis of chronic kidney disease]. *Néphrologie & Thérapeutique*, 5(4) :302–305, July 2009. ISSN 1769-7255. doi : 10.1016/j.nephro.2009.02.011.

Heagerty P. J., Lumley T., et Pepe M. S. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2) :337–344, June 2000. ISSN 0006-341X.

Hernán M. A., Clayton D., et Keiding N. The Simpson's paradox unraveled. *International Journal of Epidemiology*, 40(3) :780–785, June 2011. ISSN 1464-3685. doi : 10.1093/ije/dyr041.

Horvitz D. G. et Thompson D. J. A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, 47(260) :663–685, 1952. ISSN 0162-1459. doi : 10.2307/2280784. URL http://www.jstor.org/stable/2280784.

Hourmant M., Chantrel F., Pavaday K., Couchoud C., et Jacquelinet C. Accès à la liste d'attente et à la greffe rénale. In *Rapport annuel REIN*, chapter 7, pages 225–255. 2013.

Hung H. et Chiang C.-T. Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics*, 38(1) :8–26, Mar. 2010. ISSN 1708-945X. doi : 10.1002/cjs.10046. URL http://onlinelibrary.wiley.com/doi/10.1002/cjs.10046/abstract.

Imbens G. Nonparametric Estimation of Average Treatment Effects under Exogeneity : A Review. *Review of Economics and Statistics*, 2004.

Ioannidis J. P. A., Greenland S., Hlatky M. A., Khoury M. J., Macleod M. R., Moher D., Schulz K. F., et Tibshirani R. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet (London, England)*, 383(9912) :166–175, Jan. 2014. ISSN 1474-547X. doi : 10.1016/S0140-6736(13)62227-8.

Jager K. J., Zoccali C., Macleod A., et Dekker F. W. Confounding : what it is and how to deal with it. *Kidney International*, 73(3) :256–260, Feb. 2008. ISSN 0085-2538. doi : 10.1038/sj.ki.5002650.

Janes H. et Pepe M. S. Adjusting for Covariates in Studies of Diagnostic, Screening, or Prognostic Markers : An Old Concept in a New Setting. *Am. J. Epidemiol.*, 168(1) :89–97, July 2008. ISSN 0002-9262, 1476-6256. doi : 10.1093/aje/kwn099. URL http://aje.oxfordjournals.org/content/168/1/89.

Janes H. et Pepe M. S. Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika*, 96(2) :371–382, June 2009. ISSN 0006-3444. doi : 10.1093/biomet/asp002.

Jepsen P., Johnsen S. P., Gillman M. W., et Sørensen H. T. Interpretation of observational studies. *Heart*, 90(8) :956–960, Aug. 2004. ISSN 1355-6037. doi : 10.1136/hrt.2003.017269.

Joffe M. M., Have T. R. T., Feldman H. I., et Kimmel S. E. Model Selection, Confounder Control, and Marginal Structural Models : Review and New Applications. *The American Statistician*, 58(4) :272–279, 2004. ISSN 0003-1305. URL http://www.jstor.org/stable/27643582.

Kaplan E. L. et Meier P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282) :457–481, 1958. ISSN 0162-1459. doi : 10.2307/2281868. URL http://www.jstor.org/stable/2281868.

Lassalle M., Hannedouche T., Briançon S., Boime S., Monnet E., et Stengel B. Incidence 2013 de l'IRCT. In *Rapport annuel REIN*, chapter 2, pages 39–84. 2013.

Le Borgne F., Giraudeau B., Querard A. H., Giral M., et Foucher Y. Comparisons of the performance of different statistical tests for time-to-event analysis with confounding factors : practical illustrations in kidney transplantation. *Stat Med*, Oct. 2015. ISSN 1097-0258. doi : 10.1002/sim.6777.

Levey A. S., Stevens L. A., Schmid C. H., Zhang Y. L., Castro A. F., Feldman H. I., Kusek J. W., Eggers P., Van Lente F., Greene T., Coresh J., et CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration). A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine*, 150(9) :604–612, May 2009. ISSN 1539-3704.

Lin D. Y. et Wei L. J. The Robust Inference for the Cox Proportional Hazards Model. *Journal of the American Statistical Association*, 84(408) :1074–1078, Dec. 1989. ISSN 0162-1459. doi : 10.1080/01621459.1989.10478874.

Liu C., Xie J., et Zhang Y. Weighted nonparametric maximum likelihood estimate of a mixing distribution in nonrandomized clinical trials. *Statistics in Medicine*, 26(29) :5303–5319, 2007. ISSN 1097-0258. doi : 10.1002/sim.2914.

Lonjon G., Porcher R., Ergina P., Fouet M., et Boutron I. Potential Pitfalls of Reporting and Bias in Observational Studies With Propensity Score Analysis Assessing a Surgical Procedure : A Methodological Systematic Review. *Annals of Surgery*, May 2016. ISSN 1528-1140. doi : 10.1097/SLA.0000000000001797.

Lumley T. et Scott A. Partial likelihood ratio tests for the Cox model under complex sampling. *Statistics in Medicine*, 32(1) :110–123, Jan. 2013. ISSN 1097-0258. doi : 10.1002/sim.5492.

Macleod M. R., Michie S., Roberts I., Dirnagl U., Chalmers I., Ioannidis J. P. A., Al-Shahi Salman R., Chan A.-W., et Glasziou P. Biomedical research : increasing value, reducing waste. *Lancet (London, England)*, 383(9912) :101–104, Jan. 2014. ISSN 1474-547X. doi : 10.1016/S0140-6736(13)62329-6.

Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports. Part 1*, 50(3) :163–170, Mar. 1966. ISSN 0069-0112.

McNamee R. Regression modelling and other methods to control confounding. *Occupational and Environmental Medicine*, 62(7) :500–506, July 2005. ISSN , 1470-7926. doi : 10.1136/oem.2002.001115. URL http://oem.bmj.com/content/62/7/500.

Ministère des Affaires sociales et de la Santé. Plan 2007-2011 pour l'amélioration de la qualité de vie des personnes atteintes de maladies chroniques, 2007. URL hhttp://social-sante.gouv.fr/IMG/pdf/plan2007_2011.pdf.

Morgan S. L. et Todd J. J. A Diagnostic Routine for the Detection of Consequential Heterogeneity of Causal Effects. *Sociological Methodology*, 38(1) :231–281, Dec. 2008. ISSN 1467-9531. doi : 10.1111/j.1467-9531.2008.00204.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9531.2008.00204.x/abstract.

Normand S.-L. T., Sykora K., Li P., Mamdani M., Rochon P. A., et Anderson G. M. Readers guide to critical appraisal of cohort studies : 3. Analytical strategies to reduce confounding. *BMJ : British Medical Journal*, 330(7498) :1021–1023, Apr. 2005. ISSN 0959-8138. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC557157/.

Normand S. T., Landrum M. B., Guadagnoli E., Ayanian J. Z., Ryan T. J., Cleary P. D., et McNeil B. J. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly : a matched analysis using propensity scores. *Journal of Clinical Epidemiology*, 54(4) :387–398, Apr. 2001. ISSN 0895-4356.

Pardo-Fernandez J. C. A review on ROC curves in the presence of covariates. *Revstat Statistical Journal*, 12(1) :21–41, 2014. ISSN 1645-6726.

Pepe M. S. et Cai T. The analysis of placement values for evaluating discriminatory measures. *Biometrics*, 60(2) :528–535, June 2004. ISSN 0006-341X. doi : 10.1111/j.0006-341X.2004.00200.x.

Pepe M. S. et Longton G. Standardizing diagnostic markers to evaluate and compare their performance. *Epidemiology*, 16(5) :598–603, Sept. 2005. ISSN 1044-3983.

Peto R. et Peto J. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society, Series A (Blackwell Publishing)*, 135(2) :185–207, 1972. doi : 10.2307/2344317.

Pirracchio R., Carone M., Rigon M. R., Caruana E., Mebazaa A., et Chevret S. Propensity score estimators for the average treatment effect and the average treatment effect on the treated may yield very different estimates. *Statistical Methods in Medical Research*, Nov. 2013. ISSN 1477-0334. doi : 10.1177/0962280213507034.

Pocock S. J., Assmann S. E., Enos L. E., et Kasten L. E. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting : current practice and problems. *Statistics in Medicine*, 21(19) :2917–2930, Oct. 2002. ISSN 0277-6715. doi : 10.1002/sim.1296.

R Development Core Team . *R : A Language and Environment for Statistical Computing*. Vienna, Austria, 2010. ISBN 3-900051-07-0. URL http://www.R-project.org/.

Rao P. S., Schaubel D. E., Guidinger M. K., Andreoni K. A., Wolfe R. A., Merion R. M., Port F. K., et Sung R. S. A comprehensive risk quantification score for deceased donor kidneys : the kidney donor risk index. *Transplantation*, 88(2) :231–236, July 2009. ISSN 1534-6080. doi : 10.1097/TP.0b013e3181ac620b.

Robins J. Marginal structural models, in : 1997 Proceedings of the American Statistical Association, Section on Bayesian Statistical Science. pages 1–10, 1998.

Robins J. M., Hernán M. A., et Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5) :550–560, Sept. 2000. ISSN 1044-3983.

Rochon P. A., Gurwitz J. H., Sykora K., Mamdani M., Streiner D. L., Garfinkel S., Normand S.-L. T., et Anderson G. M. Reader's guide to critical appraisal of cohort studies : 1. Role and design. *BMJ (Clinical research ed.)*, 330(7496) :895–897, Apr. 2005. ISSN 1756-1833. doi : 10.1136/bmj.330.7496.895.

Rodríguez-Álvarez M. X., Meira-Machado L., Abu-Assi E., et Raposeiras-Roubín S. Nonparametric estimation of time-dependent ROC curves conditional on a continuous covariate. *Stat Med*, Oct. 2015. ISSN 1097-0258. doi : 10.1002/sim.6769.

Rosenbaum P. R. Model-Based Direct Adjustment. *Journal of the American Statistical Association*, 82 (398) :387, June 1987. ISSN 01621459. doi : 10.2307/2289440.

Rosenbaum P. R. et Rubin D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1) :41–55, Apr. 1983. ISSN 0006-3444, 1464-3510. doi : 10.1093/biomet/70.1.41.

Rothwell P. M. External validity of randomised controlled trials : "to whom do the results of this trial apply ?". *Lancet (London, England)*, 365(9453) :82–93, Jan. 2005. ISSN 1474-547X. doi : 10.1016/S0140-6736(04)17670-8.

Sanson-Fisher R. W., Bonevski B., Green L. W., et D'Este C. Limitations of the randomized controlled trial in evaluating population-based health interventions. *American Journal of Preventive Medicine*, 33(2) :155–161, Aug. 2007. ISSN 0749-3797. doi : 10.1016/j.amepre.2007.04.007.

Schoenfeld D. Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1) :239–241, Apr. 1982. ISSN 0006-3444, 1464-3510. doi : 10.1093/biomet/69.1.239. URL http://biomet.oxfordjournals.org/content/69/1/239.

Sekhon J. S. Multivariate and propensity score matching with balance optimization. https://cran.r-project.org/web/packages/Matching/Matching.pdf, 2015. Accessed : 2016-07-01.

Shah B. R., Laupacis A., Hux J. E., et Austin P. C. Propensity score methods gave similar results to traditional regression modeling in observational studies : a systematic review. *Journal of Clinical Epidemiology*, 58(6) :550–559, June 2005. ISSN 0895-4356. doi : 10.1016/j.jclinepi.2004.10.016.

Song X. et Zhou X.-H. A semiparametric approach for the covariate specific roc curve with survival outcome. *Statistica Sinica*, 18(947-965) :84, 2008.

Stürmer T., Joshi M., Glynn R. J., Avorn J., Rothman K. J., et Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology*, 59(5) :437–447, May 2006. ISSN 0895-4356. doi : 10.1016/j.jclinepi.2005.07.004.

Stuart E. A. Matching methods for causal inference : A review and a look forward. *Statistical Science : A Review Journal of the Institute of Mathematical Statistics*, 25(1) :1–21, Feb. 2010. ISSN 0883-4237. doi : 10.1214/09-STS313.

Sugihara M. Survival analysis using inverse probability of treatment weighted methods based on the generalized propensity score. *Pharm Stat*, 9(1) :21–34, Mar. 2010. ISSN 1539-1612. doi : 10.1002/pst.365.

Tonelli M., Wiebe N., Knoll G., Bello A., Browne S., Jadhav D., Klarenbach S., et Gill J. Systematic review : kidney transplantation compared with dialysis in clinically relevant outcomes. *American Journal of Transplantation : Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons*, 11(10) :2093–2109, Oct. 2011. ISSN 1600-6143. doi : 10.1111/j.1600-6143.2011.03686.x.

Tripepi G., Jager K. J., Dekker F. W., et Zoccali C. Testing for causality and prognosis : etiological and prognostic models. *Kidney International*, 74(12) :1512–1515, 2008. ISSN 0085-2538. doi : 10.1038/ki.2008.416. URL http://www.nature.com/ki/journal/v74/n12/full/ki2008416a.html.

Uno H., Cai T., Tian L., et Wei L. J. Evaluating Prediction Rules for t-Year Survivors With Censored Regression Models. *Journal of the American Statistical Association*, 102 :527–537, 2007. URL http://econpapers.repec.org/article/besjnlasa/v_3a102_3ay_3a2007_3am_3ajune_3ap_3a527-537.htm.

Van Spall H. G. C., Toren A., Kiss A., et Fowler R. A. Eligibility criteria of randomized controlled trials published in high-impact general medical journals : a systematic sampling review. *JAMA*, 297(11) : 1233–1240, Mar. 2007. ISSN 1538-3598. doi : 10.1001/jama.297.11.1233.

Westreich D. et Cole S. R. Invited Commentary : Positivity in Practice. *American Journal of Epidemiology*, 171(6) :674–677, Jan. 2010. ISSN 0002-9262, 1476-6256. doi : 10.1093/aje/kwp436. URL http://aje.oxfordjournals.org/content/early/2010/02/05/aje.kwp436.

Winkelmayer W. C., Weinstein M. C., Mittleman M. A., Glynn R. J., et Pliskin J. S. Health economic evaluations : the special case of end-stage renal disease treatment. *Medical Decision Making : An International Journal of the Society for Medical Decision Making*, 22(5) :417–430, Oct. 2002. ISSN 0272-989X.

Wolfe R. A., Ashby V. B., Milford E. L., Ojo A. O., Ettenger R. E., Agodoa L. Y., Held P. J., et Port F. K. Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. *The New England Journal of Medicine*, 341(23) :1725–1730, Dec. 1999. ISSN 0028-4793. doi : 10.1056/NEJM199912023412303.

Wong G., Howard K., Chapman J. R., Chadban S., Cross N., Tong A., Webster A. C., et Craig J. C. Comparative survival and economic benefits of deceased donor kidney transplantation and dialysis in people with varying ages and co-morbidities. *PloS One*, 7(1) :e29591, 2012. ISSN 1932-6203. doi : 10.1371/journal.pone.0029591.

World Health Organization. Preparing a Health Care Workforce for the 21st Century :
The Challenge of Chronic Conditions, 2005. URL http://www.who.int/chp/knowledge/publications/workforce_report/en/.

World Health Organization. Global status report on noncommunicable diseases, 2014. URL http://www.who.int/nmh/publications/ncd-status-report-2014/en/.

Xie J. et Liu C. Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Stat Med*, 24(20) :3089–3110, Oct. 2005. ISSN 0277-6715. doi : 10.1002/sim.2174.

Xu S., Ross C., Raebel M. A., Shetterly S., Blanchette C., et Smith D. Use of Stabilized Inverse Propensity Scores as Weights to Directly Estimate Relative Risk and Its Confidence Intervals. *Value in Health*, 13(2) :273–277, Mar. 2010. ISSN 1098-3015. doi : 10.1111/j.1524-4733.2009.00671.x. URL http://www.sciencedirect.com/science/article/pii/S1098301510603725.

Yu T., Li J., et Ma S. Adjusting confounders in ranking biomarkers : a model-based ROC approach. *Briefings in Bioinformatics*, 13(5) :513–523, Sept. 2012. ISSN 1477-4054. doi : 10.1093/bib/bbs008.

Zheng Y., Cai T., Stanford J. L., et Feng Z. Semiparametric models of time-dependent predictive values of prognostic biomarkers. *Biometrics*, 66(1) :50–60, Mar. 2010. ISSN 1541-0420. doi : 10.1111/j.1541-0420.2009.01246.x.

Zheng Y., Cai T., Jin Y., et Feng Z. Evaluating prognostic accuracy of biomarkers under competing risk. *Biometrics*, 68(2) :388–396, June 2012. ISSN 1541-0420. doi : 10.1111/j.1541-0420.2011.01671.x.

Zwarenstein M. et Treweek S. What kind of randomized trials do we need ? *Journal of Clinical Epidemiology*, 62(5) :461–463, May 2009. ISSN 1878-5921. doi : 10.1016/j.jclinepi.2009.01.011.

Table des figures

UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Florent LE BORGNE

Conception d'un outil simple d'utilisation pour réaliser des analyses statistiques ajustées valorisant les données de cohortes observationnelles de pathologies chroniques : application à la cohorte DIVAT.

Conception of an easy to use application allowing to perform adjusted statistical analysis for the valorization of observational data from cohorts of chronic disease: application to the DIVAT cohort.

Résumé

En recherche médicale, les cohortes permettent de mieux comprendre l'évolution d'une pathologie et d'améliorer la prise en charge des patients. La mise en évidence de liens de causalité entre certains facteurs de risque et l'évolution de l'état de santé des patients est possible grâce à des études étiologiques. L'analyse de cohortes permet aussi d'identifier des marqueurs pronostiques de l'évolution d'un état de santé. Cependant, les facteurs de confusion constituent souvent une source de biais importante dans l'interprétation des résultats des études étiologiques ou pronostiques.
Dans ce manuscrit, nous présentons deux travaux de recherche en Biostatistique dans la thématique des scores de propension. Dans le premier travail, nous comparons les performances de différents modèles permettant d'évaluer la causalité d'une exposition sur l'incidence d'un événement en présence de données censurées à droite. Dans le second travail, nous proposons un estimateur de courbes ROC dépendantes du temps standardisées et pondérées permettant d'estimer la capacité prédictive d'un marqueur en prenant en compte les facteurs de confusion potentiels. En cohérence avec l'objectif de fournir des outils statistiques adaptés, nous présentons également dans ce manuscrit une application nommée Plug-Stat®. En lien direct avec la base de données, elle permet de réaliser des analyses statistiques adaptées à la pathologie afin de faciliter la recherche épidémiologique et de mieux valoriser les données de cohortes observationnelles.

Mots clés
Scores de propension, Facteurs de confusion, Données censurées, Courbes ROC, Transplantation rénale, Cohorte.

Abstract

In medical research, cohorts help to better understand the evolution of a pathology and improve the care of patients. Causal associations between risk factors and outcomes are regularly studied through etiological studies. Cohorts analysis also allow the identification of new markers for the prediction of the patient evolution. However, confounding factors are often source of bias in the interpretation of the results of etiologic or prognostic studies.
In this manuscript, we presented two research works in Biostatistics, the common topic being propensity scores. In the first work, we compared the performances of different models allowing to evaluate the causality of an exposure on an outcome in the presence of right censored data. In the second work, we proposed an estimator of standardized and weighted time-dependent ROC curves. This estimator provides a measure of the prognostic capacities of a marker by taking into account the possible confounding factors. Consistent with our objective to provide adapted statistical tools, we also present in this manuscript an application, so-called Plug-Stat®. Directly linked with the database, it allows to perform statistical analyses adapted to the pathology in order to facilitate epidemiological studies and improve the valorization of data from observational cohorts.

Key Words
Propensity score, Confounding factors, Censored data, ROC curves, Kidney transplantation, Cohort.