

Évolution de pathologies chroniques : Modélisation et médecine de précision centrée sur le patient

Etienne Dantan

▶ **To cite this version:**

Etienne Dantan. Évolution de pathologies chroniques : Modélisation et médecine de précision centrée sur le patient. Statistiques [stat]. Université de Nantes, 2019. tel-03248520

HAL Id: tel-03248520

https://nantes-universite.hal.science/tel-03248520

Submitted on 3 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université de Nantes

Ecole Doctorale Biologie Santé

INSERM UMR 1246 - SPHERE

methodS in Patient-centered outcomes and HEalth REsearch

Habilitation à Diriger des Recherches

Etienne Dantan

Maître de Conférences en Biostatistique

Evolution de pathologies chroniques : Modélisation et médecine de précision centrée sur le patient

Travail évalué devant le jury suivant

Virginie Rondeau	Directrice de Recherche, Inserm CR1219 (Bordeaux)
Aurélien Latouche	Professeur des Universités, Institut Curie (Paris)
Francis Guillemin	Professeur des Universités et Praticien Hospitalier, EA4360 (Nancy)
Cécile Couchoud	Coordinatrice nationale du registre REIN, Agence de la Biomédecine (Paris)
Caroline Mollevi	Coordinatrice Recherche méthodologique, Institut régional du Cancer (Montpellier)
Véronique Sébille-Rivain	Professeur des Universités et Praticien Hospitalier, Inserm UMR1246 (Nantes)

Je dédie ce travail à Félix, Coline et Raphaëlle.

Remerciements

La rédaction de ce manuscrit d'Habilitation à Diriger des Recherches a été l'occasion de faire la synthèse de mes travaux passés, actuels et à venir. Le tout premier constat qui me vient à l'esprit est le suivant. L'ensemble de mes activités de recherche est le fruit d'un travail plus que jamais collaboratif. L'occasion m'est donc donnée de remercier très chaleureusement l'ensemble de mes collègues de travail. En premier lieu, je souhaite exprimer ma considération à Yohann. Tes remarques toujours pertinentes, ton dynamisme et ton franc-parler rendent nos échanges extrêmement riches et intéressants et je ne cesse d'apprendre à tes côtés. Dans ton sillage, bon nombre de travaux ont pu voir le jour. J'espère sincèrement que ceci est loin d'être terminé.

Je tiens également à remercier Magali. Tu as toujours soutenu nos développpements méthodologiques, convaincue de leur apport en épidémiologie de la transplantation rénale. Ton enthousiasme, ta motivation et ton écoute sont extrêmement précieux. Travailler à tes côtés est un réel plaisir.

Mes collègues proches (Marine, Florent, Philippe, Paul, Denis, Maxime, Thomas, Anne-Hélène, Emmanuel, Özgür, etc.), biostatisticiens, épidémiogistes ou cliniciens, ont très largement contribué aux différents travaux présentés dans ce document. Je les en remercie. J'adresse une mention très spéciale aux étudiants que j'ai encadrés ou que j'accompagne à présent, acteurs essentiels de ces recherches et sans qui ces travaux ne pourraient pas voir le jour (Marie-Cécile, Florence, Jeanne). De l'ensemble de ces collaborations, j'apprécie la teneur des échanges scientifiques et retiens également la très bonne ambiance de travail.

Un grand merci à Véronique qui m'a ouvert les portes de l'équipe SPHERE à Nantes. L'équipe que tu diriges avec Bruno nous permet de développer nos recherches en toute sérénité et en toute liberté. Ta contribution est donc immense au travers tout ce que tu réalises pour faire vivre l'équipe. J'en profite pour adresser mes remerciements à l'ensemble des membres de SPHERE (Jean-Benoît, Fanny, Myriam, Bruno, Agnès, Solène, Julie, Antoine, Line, Bastien, Arthur, et tous les autres). La très bonne QoL au travail offerte par SPHERE est appréciable chaque jour.

J'ai également une pensée pour Hélène Jacqmin-Gadda. Je crois avoir parcouru un petit bout de chemin depuis que j'ai quitté Bordeaux et l'ISPED, mais je n'en serais pas là aujourd'hui sans la confiance que tu m'as accordée en début de thèse.

Je tiens à remercier l'ensemble des membres du jury, Mmes Virginie Rondeau, Cécile Couchoud, Caroline Mollevi, Véronique Sébille-Rivain et MM. Aurélien Latouche et Francis Guillemin, de me faire l'honneur d'évaluer ce travail d'Habilitation à Diriger des Recherches. Je remercie particulièrement mes rapporteurs pour leur relecture et commentaires qui éclaireront très certainement mon travail de recherche.

Enfin, je pense sincèrement que mes activités de recherche et mon investissement professionnel sont étroitement lié à mon équilibre personnel. J'aimerais donc remercier mes proches (amis, familles Charruyer et Dantan) d'être présents à mes côtés. J'ai une pensée très particulière pour mes parents (impliqués une fois de plus) qui m'ont toujours soutenu et accompagné dans mes choix professionnels et personnels.

Merci Chloé. Mes travaux sont un peu les tiens tant ton soutien indéfectible, ta patience, ton accompagnement me permettent de m'épanouir.

♡♡♡ J'embrasse également très fort mes petits bouts : Félix, Coline et Raphaëlle ♡♡♡. Merci d'illuminer notre vie tous les jours avec votre fraîcheur. Je vous aime !

Table des matières

Avant-propos **1**

I Bilan de mes activités de recherche **5**

1 Curriculum Vitae **7**
1.1 Diplômes . . . 7
1.2 Expériences professionnelles . . . 8
1.3 Production scientifique . . . 8
1.3.1 Publications internationales avec comité de lecture . . . 8
1.3.2 Publications nationales . . . 11
1.3.3 Publications sous presse, en révision, soumises, ou en préparation . . . 11
1.3.4 Communications orales dans un congrès international . . . 11
1.3.5 Communications orales dans un congrès national . . . 12
1.3.6 Communications affichées . . . 14
1.4 Logiciels . . . 15
1.4.1 Calculateurs en ligne . . . 15
1.4.2 Packages R . . . 16
1.5 Financements obtenus en tant que porteur ou co-porteur de projet . . . 16
1.6 Responsabilités pédagogiques . . . 17
1.7 Rayonnement scientifique . . . 17

2 Synthèse des encadrements **19**
2.1 Étudiants en thèse de sciences . . . 19
2.1.1 Jeanne Simon-Pimmel . . . 19
2.1.2 Marie-Cécile Fournier . . . 20
2.1.3 Florence Gillaizeau . . . 21
2.2 Étudiants en Master 2 . . . 23
2.2.1 Julie Havas . . . 23
2.2.2 Angelina Dion . . . 23

2.2.3 Amina Blidi . . . 24
2.2.4 Damien Drubay . . . 25
2.3 Étudiants en Master 1 . . . 25
2.3.1 Alpha Barry . . . 25
2.3.2 Alexandre Huneau . . . 25
2.3.3 Alexandra Bruel . . . 26
2.3.4 Olivier Merdy . . . 26

3 Synthèse des activités de recherche passées 27
3.1 Histoire naturelle de l'évolution de maladies chroniques . . . 27
3.2 Recherches tournées vers les outils prédictifs . . . 30
3.3 Vers des approches méthodologiques centrées sur le patient . . . 31
3.4 Travaux de recherche appliqués . . . 33

II Cinq publications représentatives 35

4 Evaluation des capacités pronostiques d'un marqueur à partir de courbes de survie publiées 37

5 Modèle conjoint pour l'évolution des mesures de sérum de créatinine et le risque d'échec de greffe rénale 47

6 Evaluation des capacités prédictives d'un score pronostique dynamique 61

7 Stratification du risque de survenue d'un événement : une prise de décision médicale centrée sur le patient 73

8 Comparaison du risque de survenue d'un échec de greffe entre les receveurs de reins marginaux et standards : une approche par score de propension 89

III Projets de recherche 99

9 Pronostic en transplantation rénale 101
9.1 Introduction . . . 101
9.2 Prédictions dynamiques de l'échec de greffe rénale . . . 102
9.2.1 Matériels et méthodes . . . 103
9.2.2 Résultats . . . 104
9.2.3 Discussion . . . 112
9.3 Prédictions dynamiques en présence de mesures longitudinales extrêmes . . . 114
9.3.1 Modèle conjoint pour données longitudinales et données de survie . . . 116
9.3.2 Application au pronostic en transplantation rénale . . . 119
9.4 Prédiction dynamiques : prise en considération du délai entre deux visites de suivi 120

10 Modélisation en réanimation **123**
10.1 Revue de la littérature : Pronostic en réanimation . . . 123
10.2 La cohorte AtlanREA . . . 124
10.3 Modélisation de la survie nette . . . 124
10.4 GOS pondéré . . . 126

11 Modèle conjoint pour l'évolution longitudinale d'un score d'utilité et du risque de survenue d'un événement **129**
11.1 Introduction . . . 129
11.2 Méthodes . . . 132
11.3 Applications en cancérologie . . . 136
11.4 Applications en transplantation rénale . . . 137

IV Conclusions et Perspectives **139**

Liste des figures **145**

Liste des tableaux **147**

Bibliographie **149**

Avant-propos

Voila près de dix ans que j'ai entamé une carrière d'enseignant-chercheur en biostatistique, j'arrive à présent à un moment important de mon parcours professionnel : la reconnaissance d'une Habilitation à Diriger des Recherches (HDR). La rédaction de mon dossier de candidature à l'HDR me donne l'occasion de m'arrêter un instant sur le chemin accompli. Mais il s'agit également de me projeter en évoquant les différents travaux de recherche que je souhaiterai mener dans un avenir proche.

Ce manuscrit comporte trois parties. La première détaille mon curriculum vitae, met l'accent sur la production scientifique et mes activités d'encadrement. Elle se termine par une synthèse de mes activités de recherche passées. A la suite d'un parcours universitaire tourné vers les mathématiques, je me suis spécialisé en terminant ma formation par un Master suivi d'un doctorat en biostatistique. Cette orientation s'est peu à peu révélée par ma motivation à appliquer des développements méthodologiques dans le domaine de la santé. Initialement ancrées dans la modélisation biostatistique, mes activités de recherche sont de plus en plus imprégnées par la nécessité de les voir utiles, non seulement pour la communauté scientifique en biostatistique et médicale, mais également pour les patients.

Dans une seconde partie, conformément aux recommandations de rédaction, j'ai choisi cinq publications que j'estime être les plus représentatives de mes activités de recherche. Ces publications illustrent la volonté que j'ai de diffuser mes travaux dans différents domaines de spécialités en publiant notamment dans des journaux de biostatistique, d'épidémiologie et de spécialités médicales.

De fait, mes recherches se veulent pluridisciplinaires et progressent grâce aux compétences diverses et variées de mes collègues. Le travail collaboratif est ainsi devenu prépondérant. Dans une troisième et dernière partie, j'aborde les projets de recherche que nous avons récemment démarrés et ceux encore à l'état embryonnaire qui représenteront le futur proche de notre groupe de travail.

Liste des symboles

1-RBI 1-year Renal Biopsy Index

ATER Attaché Temporaire d'Enseignement et de Recherche

ATG Anti-Thymocyte Globuline

AUC Area Under the ROC Curve

AVC hémorragiques Accident Vasculaire Cérébral hémorragique

DIVAT Données Informatisées et VAlidées en Transplantation

DynPG Dynamic prediction of Patient and Graft survival

ECD Expanded Criteria Donor

ENSAI École Nationale de la Statistique et de l'Analyse de l'Information

GOS Glasgow Outcome Scale

HDR Habilitation à Diriger des Recherches

HM Hémorragie Méningée

IC95% Intervalle de confiance à 95%

IF Impact Factor

INSERM Institut National de la Santé Et de la Recherche Médicale

IPW Inverse Probability Weighting

ISPED Institut de Santé Publique, d'Epidémiologie et de Développement

KDRI Kidney Donor Risk Index

mRS modified Rankin Scale

QALY Quality Adjusted Life-Years

ROC Receiver Operating Characteristic

SCD Standard Criteria Donor

SPHERE methodS in Patient-centered outcomes and HEalth REsearch

TC Traumatisme Crânien

TRIPOD Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis

UE Unité d'Enseignement

UFR Unité de Formation et de Recherche

UW-mRS Utility-Weighted modified Rankin Scale

Première partie

Bilan de mes activités de recherche

- Chapitre 1 -

Curriculum Vitae

DANTAN Etienne

- Lieu et date de naissance : 11 Novembre 1982, à Niort (79)
- Nationalité : Français
- Situation familiale : Marié, 3 enfants
- Statut : Maître de Conférences en Biostatistique
- CNU : 85 - Sciences physico-chimiques et ingénierie appliquée à la santé
- Laboratoire : Equipe INSERM UMR 1246 - SPHERE (methodS in Patient-centered outcomes and HEalth REsearch), à l'Université de Nantes
- Adresse professionnelle : Institut de Recherche en Santé 2 (IRS2), 22 boulevard Bénoni Goullin - 44200 Nantes Cedex 1 - FRANCE
- Téléphone : +33 (0)2 53 00 91 23
- Email : Etienne.Dantan@univ-nantes.fr

1.1 Diplômes

2009 Doctorat en Biostatistique de l'Université de Bordeaux 2, soutenu le 8 décembre 2009, mention très honorable
Titre : *Modèles conjoints pour données longitudinales et données de survie incomplètes appliqués à l'étude du vieillissement cognitif*
Directrice : Hélène Jacqmin-Gadda

2006 Master 2 Recherche (ex. Diplôme d'Etude Approfondies, DEA) en Epidémiologie et Biostatistique de l'Université de Bordeaux 2 (Institut de Santé Publique, d'Epidémiologie et de Développement, ISPED), mention assez bien

2005 Master 2 Professionnel (ex. Diplôme d'Etude Supérieures Spécialisées, DESS) en Sciences Humaines Appliqués, option Modélisation mathématique, de l'Université de Bordeaux 2, mention bien

2004 Master 1 (ex. Maîtrise) en Sciences Humaines Appliqués, option Modélisation mathématique, de l'Université de Bordeaux 2, mention bien

2003 Licence de Mathématiques Appliquées en Sciences Sociales (MASS) de l'Université de La Rochelle, mention assez bien

2002 Diplôme d'Etudes Universitaires Générales (DEUG) de Mathématiques Appliquées en Sciences Sociales (MASS) de l'Université de La Rochelle

2000 Baccalauréat Scientifique au Lycée Jean Macé à Niort (79), mention assez bien

1.2 Expériences professionnelles

2011 - ... Maître de conférences en Biostatistique au sein de l'unité INSERM UMR 1246 - SPHERE, rattaché à l'UFR de Pharmacie de l'université de Nantes

2014 - ... Enseignant contractuel extérieur à l'École Nationale de la Statistique et de l'Analyse de l'Information (ENSAI), à Bruz

2010 - 2011 Attaché Temporaire d'Enseignement et de Recherche (ATER) en Biostatistique au sein de l'équipe EA 4275 - SPHERE (nouvellement Unité INSERM UMR 1246), rattaché à l'UFR de Pharmacie de l'université de Nantes

2006 - 2009 Doctorant à l'Université de Bordeaux 2 au sein de l'unité INSERM U897

2006 - 2009 Enseignant vacataire à l'École du Service de Santé des Armées (ESSA) de Bordeaux

2006 - 2008 Enseignant vacataire à l'Université de Bordeaux 2

1.3 Production scientifique

1.3.1 Publications internationales avec comité de lecture

1. Fournier MC, Foucher Y, Blanche P, Legendre C, Girerd S, Ladrière M, Morelon E, Buron F, Rostaing L, Kamar N, Mourad G, Garrigue V, Couvrat-Desvergnes G, Giral M, and

Dantan E; DIVAT Consortium. Dynamic predictions of long-term kidney graft failure : an information tool promoting patient-centered care. *Nephrology Dialysis Transplantation*. 2019 Mar 11, doi : 10.1093/ndt/gfz027. [IF=4.600]

2. Lorent M, Maalmi H, Tessier P, Supiot S, **Dantan E** and Foucher Y. Meta-analysis of predictive models to assess the clinical validity and utility for patient-centered medical decision making : application to the CAncer of the Prostate Risk Assessment (CAPRA). *BMC Medical Informatics and Decision Making*. 2019 Jan 7 ;19(1) :2. [IF=2.134]

3. Caruana E, Foucher Y, Tessier P, Frenel JS, Classe JM and **Dantan E**. Patient-centered simulations to assess the usefulness of the 70-gene signature for adjuvant chemotherapy administration in early-stage breast cancer. *Breast Cancer Research and Treatment*. 2019 Jan 2. [IF=3.605]

4. Couvrat-Desvergnes G, Foucher Y, Le Borgne F, Dion A, Mourad G, Garrigue V, Legendre C, Rostaing L, Kamar N, Kessler M, Ladrière M, Morelon E, Buron F, Giral M and **Dantan E**. Comparison of graft and patient survival according to the transplantation center policy for 1-year screening biopsy among stable kidney recipients : a propensity score-based study. *Nephrology Dialysis Transplantation*. 2018 July 14. [IF=4.600]

5. Giral M, Renaudin, K, Naesens M, Luning R, Anglicheau D, Morelon E, Huneau A, Paul C, Brouard S, Couvrat-Desvergnes G, Foucher Y and **Dantan E**. The 1-year Renal Biopsy Index (1-RBI) : a scoring system to drive biopsy indication at 1-year post-kidney transplantation. *Transpl Int*. 2018 June 7 ;31 :947-955. [IF=3.196]

6. Foucher Y, Lorent M, Tessier P, Supiot S, Sébille V and **Dantan E**. A mini-review of quality of life as an outcome in prostate cancer trials : patient-centered approaches are needed to propose appropriate treatments on behalf of patients. *Health Qual Life Outcomes*. 2018 Mar 5 ;16(1) :40. [IF=2.278]

7. Quérard AH, Le Borgne F, Dion A, Giral M, Mourad G, Garrigue V, Rostaing L, Kamar N, Loupy A, Legendre C, Morelon E, Buron F, Foucher Y and **Dantan E**. Propensity score-based comparison of the graft failure risk between kidney transplant recipients of standard and expanded criteria donor grafts : Toward increasing the pool of marginal donors. *Am J Transplant*. 2018 May ;18(5) :1151-1157. Epub 2018 Feb 8. [IF=6.493]

8. Fournier MC, **Dantan E** and Blanche P. An R2 -curve for evaluating the accuracy of dynamic predictions. *Stat Med*. 2018 Mar 30 ;37(7) :1125-1133. Epub 2017 Dec 4. [IF=1.932]

9. Gillaizeau F, Sénage T, Le Borgne F, Le Tourneau T, Roussel JC, Leffondré K, Porcher R, Giraudeau B, **Dantan E** and Foucher Y. Inverse probability weighting to control confounding in an illness-death model for interval-censored data. *Stat Med*. 2018 Apr 15 ;37(8) :1245-1258. Epub 2017 Dec 4. [IF=1.932]

10. **Dantan E**, Foucher Y, Lorent M, Giral M and Tessier P. Optimal threshold estimator of a prognostic marker by maximizing a time-dependent expected utility function for a

patient-centered stratified medicine. *Stat Methods Med Res*. 2018 Jun ;27(6) :1847-1859. Epub 2016 Oct 20. [IF=2.284]

11. Fournier MC, Foucher Y, Blanche P, Buron F, Giral M and **Dantan E**. A joint model for longitudinal and time-to-event data to better assess the specific role of donor and recipient factors on long-term kidney transplantation outcomes. *Eur J Epidemiol*. 2016 May ;31(5) :469-79. Epub 2016 Feb 1. [IF=7.023]
12. Quérard AH, Foucher Y, Combescure C, **Dantan E**, Larmet D, Lorent M, Pouteau LM, Giral M and Gillaizeau F. Comparison of survival outcomes between Expanded Criteria Donor and Standard Criteria Donor kidney transplant recipients : a systematic review and meta-analysis. *Transpl Int*. 2016 Apr ;29(4) :403-15. Epub 2016 Feb 12. [IF=3.196]
13. Gillaizeau F, **Dantan E**, Giral M and Foucher Y. A multistate additive relative survival semi-Markov model. *Stat Methods Med Res*. 2017 Aug ;26(4) :1700-1711. Epub 2015 Jun 7. [IF=2.284]
14. Debout A, Foucher Y, Trébern-Launay K, Legendre C, Kreis H, Mourad G, Garrigue V, Morelon E, Buron F, Rostaing L, Kamar N, Kessler M, Ladrière M, Poignas A, Blidi A, Soulillou JP, Giral M and **Dantan E**. Each additional hour of cold ischemia time significantly increases the risk of graft failure and mortality following renal transplantation. *Kidney Int*. 2015 Feb ;87(2) :343-9. Epub 2014 Sep 17. [IF=8.429]
15. **Dantan E**, Combescure C, Lorent M, Ashton-Chess J, Daguin P, Classe JM, Giral M and Foucher Y. An original approach was used to better evaluate the capacity of a prognostic marker using published survival curves. *J Clin Epidemiol*. 2014 Apr ;67(4) :441-8. [IF=4.245]
16. Couvrat-Desvergnes G, Masseau A, Benveniste O, Bruel A, Hervier B, Mussini JM, Buob D, Hachulla E, Matignon M, Azar R, Mac Namara E, MacGregor B, Daniel L, Lacraz A, De Broucker T, Rouvier P, Carli P, Laville M, **Dantan E**, Hammidou M, Moreau A and Fakhouri F. The Spectrum of Renal Involvement in Patients With Inflammatory Myopathies. *Medicine*. 2014 Jan ;93(1) :33-41. [IF=2.028]
17. De Bock E, Hardouin JB, Blanchin M, Le Neel T, Kubis G, Bonnaud-Antignac A, **Dantan E** and Sébille V. Rasch-family models are more valuable than score-based approaches for analysing longitudinal patient-reported outcomes with missing data. *Stat Methods Med Res*. 2016 Oct ;25(5) :2067-2087. Epub 2013 Dec 16. [IF=2.284]
18. Cantarovich D, Hodemon-Corne B, Trébern-Launay K, Giral M, Foucher Y and **Dantan E**. Early Steroid Withdrawal Compared With Steroid Avoidance Correlates With Graft Failure Among Kidney Transplant Recipients With an History of Diabetes. *Transplantation Proceedings*. 2013 ;45(4) :1497-1502. [IF=0.806]
19. **Dantan E**, Joly P, Dartigues JF and Jacqmin-Gadda H. Joint model with latent state for longitudinal and multistate data. *Biostatistics*. 2011 Oct ;12(4) :723-36. Epub 2011 Mar 17. [IF=1.504]

20. **Dantan E**, Proust-Lima C, Letenneur L and Jacqmin-Gadda H. Pattern Mixture Models and Latent Class Models for the Analysis of Multivariate Longitudinal Data with Informative Dropouts. *The International Journal of Biostatistics*. 2008 ;4(1) : Article 14. [IF=0.840]

1.3.2 Publications nationales

1. Gillaizeau F, Giral M, **Dantan E**, Dragun D, Soulillou JP and Foucher Y. Multi-state analysis of kidney transplant recipients outcome : a semi-Markov model for studying the role of pre-transplant sensitization against Angiotensin II Type 1 receptor. *Journal de la Société Française de Statistique*. 2014 ;155(1) :117-133.

1.3.3 Publications sous presse, en révision, soumises, ou en préparation

1. Lenain R, **Dantan E**, Giral M, Foucher Y, Hazzan M, Fournier MC. External validation of dynamic predictions of long-term kidney graft failure. *En cours de rédaction.*
2. Quérard AH, Le Borgne F, Dion A, Giral M, Mourad G, Garrigue V, Rostaing L, Kamar N, Loupy A, Legendre C, Morelon E, Buron F, Foucher Y and **Dantan E**. Validation of the Kidney Donor Risk Index to graft failure prognostic from a French kidney transplant recipient cohort. *En cours de rédaction.*
3. Asar O, Mc Fournier MC and **Dantan E**. Dynamic predictions of kidney graft survival in the presence of longitudinal outliers. *Soumis à Journal of the Royal Statistical Society : Series C.*
4. Simon-Pimmel J, Foucher Y, Leger M, Feuillet F, Contentin L, Cinotti R, Frasca D and **Dantan E**. Hemorrhagic brain injuries in critical care units : A systematic review of prognostic models. *Soumis à Stroke.*
5. Simon-Pimmel J, Foucher Y, Leger M, Feuillet F, Contentin L, Cinotti R, Frasca D and **Dantan E**. Systematic review of prognostic models in intensive care admitted traumatic brain injuries. *En cours de rédaction.*

1.3.4 Communications orales dans un congrès international

1. Quérard AH and **Dantan E**. Propensity score-based comparison of the graft failure risk between kidney transplant recipients of standard and expanded criteria donor grafts : towards increasing the pool of marginal donors. European Society for Organ Transplantation, Barcelone, Spain, 2017.
2. Blanche P, Fournier MC and **Dantan E**. R2-type Curves for Dynamic Predictions from Joint Longitudinal-Survival Models : Inference & application to prediction of kidney graft

failure. Workshop on Flexible Models for Longitudinal and Survival Data with Applications in Biostatistics, Warwick, England, 2015.

3. Merdy O, Lachaud L, Marc L, Chardigny S, Tavernier E, Labour B, Toumaniantz G, Niel S and **Dantan E**. The frequency of pain associated with repeated OMT training in an osteopathic undergraduate population : a cross-sectional study. 10th International Conference on Advances in Osteopathic Research, Sao Paulo, Brazil, 2014.
4. Gillaizeau F, **Dantan E**, Giral M and Foucher Y. Additive relative survival multistate semi-Markov model. International Society for Clinical Biostatistics, Vienna, Austria, 2014.
5. Gillaizeau F, Foucher Y, Dragun D, Soulillou JP, Dufay A, Van Huyen J, Renaudin K, Moreau A, **Dantan E** and Giral M. A multistate model to investigate the relationship between pre-graft level of Angiotensin II Type 1 Receptor (AT1R) antibodies and kidney transplant recipients outcome. American Transplant Congress, Seattle, USA, 2013.
6. **Dantan E**, Lorent M and Foucher Y. Confusing correlation and prediction : a simple approach to evaluate marker accuracy for predicting time-to-event. 39th Annual Meeting of the European Group for Blood and Marrow Transplantation. London, England, 2013.
7. **Dantan E**, Combescure C, Ashton-Chess J, Lorent M, Giral M and Foucher Y. How to evaluate the prognostic capacity of surrogate makers ? A methodological review in transplantation literature. Congress of the British Transplant Society, Glasgow, Scotland, 2012.
8. de Bock E, Hardouin JB, Blanchin M, Le Neel T, Kubis G, **Dantan E** and Sébille V. Analysis of longitudinal Patient-Reported Outcomes with intermittent missing data (informative and non informative) : comparison of score and Rasch-based methods. 19th Annual Conference of ISOQOL, Budapest, Hungary, 2012.
9. Cantarovich D, Hodemon-Corne B, **Dantan E**, Trébern-Launay K, Karam G and Giral M. Corticosteroid avoidance or withdrawal : which immunosuppression is better after renal transplantation ? American Transplant Congress, Philadelphia, USA, 2011.
10. **Dantan E** and Jacqmin-Gadda H. Joint model with latent state for the pre-diagnosis phase of dementia. International Society for Clinical Biostatistics, Pragues, Czech republic, 2009.
11. **Dantan E**, Proust-Lima C and Jacqmin-Gadda H. Pattern Mixture Models and Latent Class Models for the Analysis of Multivariate Longitudinal Data with informative Dropouts. International Society for Clinical Biostatistics, Alexandroupolis, Greece, 2007.

1.3.5 Communications orales dans un congrès national

1. Simon-Pimmel J, Bodet L, Cinotti R, Feuillet F, Foucher Y, Frasca Y, Leger M and **Dantan E**. Retour d'expérience d'une revue de la littérature de scores pronostiques : Application à la neuro-réanimation. 13e conférence francophone d'épidémiologie clinique (EPICLIN) et 26e journées des statisticiens des centres de lutte contre le cancer, Toulouse, France, 2019.

2. Fournier MC, Foucher Y, Giral M and **Dantan E**. Dynamic predictions of long-term kidney graft failure : a tool to better inform the recipient in a shared decision making approach. Ecole d'Automne du DHU 2020, Nantes, 2018.
3. Fournier MC, Foucher Y, Giral M and **Dantan E**. Une application Shiny pour prédire dynamiquement la survie patient-greffon d'un patient transplanté rénal : vers une décision médicale plus partagée. 7ème rencontres R, Rennes, France, 2018.
4. Fournier MC, Foucher Y, Giral M and **Dantan E**. Prédictions dynamiques du risque d'échec de greffe : un outil pour aider à l'information des patients transplantés rénaux et une prise en charge centrée sur le patient et ses attentes. 12e conférence francophone d'épidémiologie clinique (EPICLIN) et 25e journées des statisticiens des centres de lutte contre le cancer, Nice, France, 2018.
5. Fournier MC, Blanche P, Foucher Y, Giral M and **Dantan E**. Construction d'un score pronostique dynamique et évaluation de ses performances : application en transplantation rénale. 10e conférence francophone d'épidémiologie clinique (EPICLIN 2016) et 23e journées des statisticiens des centres de lutte contre le cancer, Strasbourg, France, 2016.
6. Fournier MC, Foucher Y, Giral M and **Dantan E**. Etude des déterminants de santé associés à l'évolution de la créatininémie etou du risque d'échec de greffe chez des patients transplantés rénaux : une approche par modélisation conjointe. GDR Statistique & Santé, Paris, France, 2015.
7. Fournier MC, Foucher Y, Giral M and **Dantan E**. Etude des déterminants de santé associés à l'évolution de la créatininémie et / ou du risque d'échec de greffe chez des patients transplantés rénaux : une approche par modélisation conjointe. EPICLIN, Montpellier, France, 2015.
8. **Dantan E**, Foucher Y, Giral M and Tessier P. Quel seuil de discrimination pour un test pronostique à des fins de médecine stratifiée ? Une approche pratique d'une théorie de la décision basée sur les QALYs. EPICLIN, Montpellier, France, 2015.
9. Maalmi H, Lorent M, **Dantan E**, Supiot S, Classe JM and Foucher Y. Une méthode non-paramétrique pour évaluer les capacités pronostiques d'un marqueur à partir des données agrégées de la littérature : une illustration par une méta-analyse du score CAPRA en cancer de la prostate. EPICLIN, Montpellier, France, 2015.
10. Gillaizeau F, **Dantan E**, Giral M and Foucher Y. A multistate additive relative survival semi-Markov model. 46e Journées de Statistique de la SFdS, Rennes, France, 2014.
11. **Dantan E**, Combescure C, Lorent M, Ashton-Chess J, Daguin P, Classe JM, Giral M and Foucher Y. An original approach to evaluate the prognostic marker capacity using published survival curves. 46e Journées de Statistique de la SFdS, Rennes, France, 2014.
12. Gillaizeau F, Giral M, **Dantan E**, Dragun D, Soulillou JP and Foucher Y. A Semi-Markov model with Generalized Weibull distribution for multistate data in kidney transplant recipients. Journées de la Société Française de Biométrie, Paris, France, 2012.

13. de Bock E, Hardouin JB, Blanchin M, Le Neel T, Kubis G, **Dantan E**, and Sébille V. Analyse de PRO recueillis de façon longitudinale et comportant des données manquantes intermittentes potentiellement informatives : comparaison de modèles issus de la théorie classique des tests et de la théorie de réponse aux items. GDR statistique et santé, Rennes, France, 2012.
14. **Dantan E**, Proust-Lima C and Jacqmin-Gadda H. Modèles par mélange de schémas d'observation et modèles à classes latentes pour le traitement des sorties d'étude informative. GDR Statistiques et santé, Paris, France, 2007.
15. **Dantan E**, Proust-Lima C and Jacqmin-Gadda H. Modèles par mélange de schémas d'observation et modèles à classes latentes pour le traitement des sorties d'étude informatives. 39e Journées de Statistique de la SFdS, Angers, France, 2007.

1.3.6 Communications affichées

1. Fournier MC, **Dantan E** and Blanche P. R2-type curve to evaluate dynamic predictions. International Society for Clinical Biostatistics, Birmingham, England, 2016.
2. Couvrat-Desvergnes G, Giral M and **Dantan E**. Clinical utility of surveillance kidney transplant biopsies at 1-year post-transplantation : a long-term survival comparison among renal transplant recipients. 26th International Congress of The Transplantation Society, Hong-Kong, China, 2016.
3. Quérard AH, Foucher Y, Combescure C, **Dantan E**, Larmet D, Lorent M, Pouteau LM, Giral M and Gillaizeau F. Comparison of survival outcomes between Expanded Criteria Donors and Standard Criteria Donors kidney transplant recipients : a systematic review and meta-analysis. American Transplant Congress, Philadelphia, USA, 2015.
4. Fournier MC, **Dantan E**, Foucher Y and Giral M. A joint survival longitudinal approach to better assess the etiological role of risk factors of kidney transplant graft failure. American transplant congress, Philadelphia, USA, 2015.
5. **Dantan E**, Renaudin K, Brouard S, Huneau A, Paul C, Couvrat-Desvergnes G and Giral M. A clinical signature surrogate of histological lesions to avoid unnecessary one year surveillance biopsy after renal transplantation. American Transplant Congress, Philadelphia, USA, 2015.
6. **Dantan E**, Foucher Y, Combescure C, Lorent M, Ashton-Chess J, Daguin P, Giral M and Classe JM. Évaluation a posteriori des capacités d'un marqueur génétique à pronostiquer la survenue de métastases distantes chez des femmes atteintes d'un cancer du sein. 8ème Journées du Cancéropôle du Grand-Ouest, Les Sables d'Olonnes, France, 2014.
7. Quérard AH, Gillaizeau F, Pouteau LM, Foucher Y, Larmet D, **Dantan E**, Lorent M, Combescure C and Giral M. Expanded Criteria Donor in Kidney Transplantation : a meta-analysis. European Society for Organ Transplantation, Madrid, Spain, 2014.

8. **Dantan E**, Daveluy A, Laribière A, Laborde B, Couallier V, Miremont-Salamé G, Moore N, Haramburu F. Health professional users of a regional pharmacovigilance centre : trends over time. 5th ISoP Annual Meeting Pharmacovigilance - A global Concern : East meets West, Manila, Philippines, 2005.

1.4 Logiciels

A travers différents projets de recherche, j'ai participé à l'élaboration de scores prédictifs ainsi qu'au développement de méthodes pour une aide à la prise de décision médicale. Ces outils sont disponibles en ligne (`www.labcom-risca.com`) sous forme d'applications pour smartphones, tablettes et ordinateurs, ou implémentés dans des packages R.

1.4.1 Calculateurs en ligne

1-RBI Pour aider à l'indication d'une biopsie du greffon après une transplantation rénale, nous avons proposé un score diagnostique d'anormalité histologique à 1 an post-transplantation : le 1-RBI (*1-year Renal Biopsy Index*). Il s'agit d'un score calculable à 1 an post-transplantation à partir de 5 variables collectées en pré-greffe et au cours de la première année de greffe rénale. Ce travail vient d'être publié dans *Transplantation International* [Giral et al., 2018].

DynPG Afin d'obtenir des prédictions dynamiques de l'échec de greffe rénale, nous proposons de considérer les mesures répétées de sérum de créatinine au cours du suivi du patient. Disponible en ligne sous la forme d'une application SHINY, le DynPG (*dynamic prediction of Patient and Graft survival*) est un outil de prédiction de l'échec de greffe qui se met à jour au cours du suivi du patient. Il est calculable à partir de 6 variables collectées en pré-greffe et du suivi longitudinal de serum de créatinine. Ce travail vient d'être accepté dans *Nephrology Dialysis Transplantation*.

EVALBIOM La distance entre les courbes de survie est souvent mal interprétée en terme de capacités pronostiques. EVALBIOM est une application en ligne permettant de calculer différents indicateurs dépendants du temps à partir de courbes de survie déjà publiées afin de mieux évaluer les capacités pronostiques d'un marqueur. En fournissant les effectifs initiaux de deux groupes de patients ainsi que leurs probabilités de survie respectives à un certain temps de suivi, EVALBIOM retourne, pour ce temps de pronostic, les indicateurs objectivement interprétables comme la sensibilité, la spécificité, les valeurs prédictives, etc. Ce travail a fait l'objet d'une publication dans le *Journal of Clinical Epidemiology* [Dantan et al., 2014].

1.4.2 Packages R

R est un logiciel libre d'accès (*open source*) proposant un langage et un environnement pour les calculs statistiques et graphiques. Il offre une grande variété d'outils statistiques, auxquels peuvent s'ajouter des éléments complémentaires, appelés *packages*. Tous les packages sont disponibles à l'adresse suivante : `https://cran.r-project.org`.

ROCt Développé par Yohann Foucher, Marine Lorent et Florent Le Borgne, ROCt regroupe un ensemble de fonctions permettant d'estimer les capacités diagnostiques/pronostiques d'un marqueur quantitatif continu, avec notamment l'estimation de courbes ROC (Receiver Operating Characteristic) dépendantes du temps pour la gestion d'événements censurés à droite [Heagerty et al., 2000]. L'estimateur de la courbe ROC dépendante du temps a été développé pour les modèles de survie traditionnels (toutes causes confondues) et pour les modèles de survie nette (mortalité en excès liée à une pathologie) [Lorent et al., 2014]. Il a également été proposé d'estimer les capacités ajustées sur d'éventuels facteurs de confusion [Le Borgne et al., 2017].

Parmi les fonctions disponibles dans ce package, j'ai contribué au développement et à l'implémentation de fonctions permettant d'estimer un seuil de décision optimal pour la mise en place d'une prise en charge stratifiée. Nous proposons l'estimation d'une fonction d'utilité dépendante du temps. La maximisation de cette fonction permet de déterminer un seuil optimal de discrimination pour un marqueur quantitatif en tenant compte de l'espérance de vie ainsi que de la qualité de vie dans chacun des bras de traitement [Dantan et al., 2018].

1.5 Financements obtenus en tant que porteur ou co-porteur de projet

- Validation du score KDRI : étude française multicentrique comparative de la survie à long terme de patients transplantés rénaux (Investigateur principal : E.Dantan, Appel d'Offre "Recherche et Greffe" 2016 de l'Agence de la Biomédecine).
- Validation of Expanded Criteria Donor (ECD), a comparative multicenter french study of long-term survival of kidney transplant patients (Investigateur principal : AH. Quérard, Financement de l'entreprise Chiesi France).
- Score diagnostique de la normalite anatomo-pathologique sur les biopsies de surveillance des greffons renaux a un an post-transplantation (Investigateur principal : Y. Foucher, Appel d'Offre "Recherche et Greffe" 2014 de l'Agence de la Biomédecine).
- Méta-Analyses de marqueur pronostique en cancérologie : projet MAP-MARKER (Investigateur principal : Y. Foucher, Appel à projets 2013 de recherche en Sciences Humaines

et Sociales, Epidémiologie et Santé Publique de l'Institut National du Cancer).

1.6 Responsabilités pédagogiques

- **Ecole doctorale** Biologie Santé de l'Université Bretagne-Loire : Responsable de l'Unité d'Enseignement (UE) *Initiation à la biostatistique*
- **Master 2** Modélisation en Pharmacologie Clinique et Epidémiologie (MPCE), à l'Université de Nantes : Responsable de l'UE *Méthodes statistiques avancées en Epidemiologie*, Intervenant au sein des UE *Données longitudinales et données de survie*, *Principes d'estimation en biostatistique* et *Epidémiologie clinique*.
- **Master 2** Biologie, Biotechnologies et Recherche Thérapeutique (BBRT), à l'Université de Nantes : Responsable de l'UE *Biostatistique I*
- **ENSAI :** Responsable de l'UE *Modèles mixtes* en troisième année filière Statistique pour les Sciences de la Vie
- **Master 1** Biologie Santé, à l'Université de Nantes : Intervenant au sein de l'UE *Introduction à la biostatistique*
- **Deuxième année de pharmacie**, à l'Université de Nantes : Intervenant au sein de l'UE *Biomathématiques appliquées à la pharmacocinétique*
- **PACES** (Première Année Commune des Etudes de Santé), à l'Université de Nantes : Intervenant au sein de l'UE *Biostatistique*

1.7 Rayonnement scientifique

- **Reviewing d'articles scientifiques** pour les revues *Journal of Clinical Epidemiology*, *International Journal of Biostatistics*, *Scientific Reports*
- **Accueil d'un chercheur étranger :** Özgür Asar, enseignant-chercheur à l'Université d'Acibadem en Turquie, a été en visite au sein de l'Unité INSERM UMR 1246 - SPHERE durant le mois de Septembre 2018 avec l'aide d'une bourse de mobilité obtenue auprès de l'ambassade de France.
- **Communications orales invitées dans un colloque national**
  - Tessier P and Dantan E. Stratification des patients transplantés rénaux à des fins de médecine personnalisée : une application des théories de la décision dans un contexte pronostique. Journée scientifique "Les 20 ans de DIVAT", Nantes, 2015.
  - Tessier P and Dantan E. Du pronostic au choix médical : quelles utilisations des théories de la décision dans la mise en place d'une stratégie thérapeutique ? Cohortes

et biocollections au service de la médecine prédictive - La Biostatistique au cœur de cette recherche translationnelle, Journées scientifiques de l'Université de Nantes, Nantes, 2012.

— Dantan E, Lorent M and Foucher Y. A simple approach to evaluate biomarker accuracy for predicting time-to-event from published paper. Journée Méthodologie de la DiRC Grand-Ouest, Nantes, 2012.

- **Co-organisation**
  - — du colloque *Cohortes et biocollections au service de la médecine predictive : la Biostatistique au cœur de cette recherche translationnelle* dans le cadre des Journées Scientifiques de l'Université de Nantes en 2012,
  - — d'une formation en Biostatistique sur le thème de l'inférence en statistique dans le cadre du projet européen ShareBiotech, Nantes, 2012.
- **Membre du comité de suivi** du Labcom RISCA - Research in Informatics and Statistics for Cohort Analyses (Investigateur principal : Y. Foucher, Entreprise partenaire : IDBC-A2COM, Programme Labcom 2016 de l'Agence Nationale de la Recherche).

- Chapitre 2 -

Synthèse des encadrements

2.1 Étudiants en thèse de sciences

2.1.1 Jeanne Simon-Pimmel

Jeanne Simon-Pimmel a débuté sa thèse en 2017 après l'obtention d'une allocation doctorale financée par l'Université de Nantes (concours de l'Ecole Doctorale Biologie-Santé). Son travail de thèse s'intitule *Pronostic des patients cérébro-lésés en réanimation : d'un modèle de prédiction à la mise en place d'une médecine stratifiée*. Sa thèse est dirigée par le Dr. Yohann Foucher et le Pr. Denis Frasca (INSERM UMR 1246 - SPHERE). J'en suis le co-encadrant à hauteur de 40%.

Ce travail de thèse en Epidémiologie et Biostatistique sera appliqué au contexte de la réanimation à partir des données de la cohorte AtlanRéa (`www.atlanrea.org`). Un premier travail en cours consiste en une revue systématique des outils pronostiques existants pour la prise en charge des patients cérébro-lésés sévères en réanimation. Les résultats de cette revue systématique orienteront la suite du travail de recherche pour l'élaboration d'un modèle de prédiction chez les patients cérébro-lésés. Plusieurs pistes de réflexion ont déjà été évoquées. Par exemple, l'évolution longitudinale précoce de la pression intracrânienne comme marqueur de sévérité de l'agression cérébrale pourrait contribuer au développement d'un outil de prédiction dynamique à moyen terme. Les modèles de survie relative initialement développés pour prendre en considération la mortalité spécifiquement liée à la pathologie étudiée pourrait être utilisés dans le contexte de la réanimation afin d'étudier le délai de retour à un risque de mortalité considéré comme normal. Ce type d'approche méthodologique pourrait s'avérer pertinent dans les problématiques du droit à l'oubli pour des patients ayant subi une atteinte cérébrale.

Une première communication orale a été réalisée à l'occasion du congrès EPICLIN 2019 :

- Simon-Pimmel J, Bodet L, Cinotti R, Feuillet F, Foucher Y, Frasca Y, Leger M and **Dantan E**. Retour d'expérience d'une revue de la littérature de scores pronostiques : Application à la neuro-réanimation. 13e conférence francophone d'épidémiologie clinique (EPICLIN) et 26e journées des statisticiens des centres de lutte contre le cancer, Toulouse, France, 2019.

2.1.2 Marie-Cécile Fournier

En 2013, Marie-Cécile Fournier avait aussi réussi le concours de l'Ecole Doctorale Biologie-Santé pour l'obtention d'un contrat doctoral de l'Université de Nantes. Son travail de thèse s'intitule *Pronostic dynamique de l'évolution de l'état de santé de patients atteints d'une maladie chronique*. Sa thèse a été soutenue le 10 octobre 2016. Elle a été dirigée par le Dr. Yohann Foucher (INSERM UMR 1246 - SPHERE) et le Pr. Magali Giral (INSERM UMR 1064). J'en fus le co-encadrant à hauteur de 50%.

Ce travail de thèse en Epidémiologie et Biostatistique appliqué à la transplantation rénale s'intéresse aux modèles conjoints pour données longitudinales et données de survie [Rizopoulos, 2012]. Dans un premier travail [Fournier et al., 2016], nous proposons un modèle conjoint à effets aléatoires partagés afin d'étudier le rôle spécifique des déterminants de santé sur l'évolution longitudinale du sérum de créatinine et/ou sur le risque d'échec de greffe rénale (voir le chapitre 5 pour plus de détails). Dans une seconde partie du travail de recherche, nous nous sommes intéressés aux prédictions dynamiques. En collaboration avec le Dr. Paul Blanche, nous proposons un indicateur de type R^2 [Fournier et al., 2018] afin d'évaluer les capacités pronostiques d'un score dynamique (voir le chapitre 6 pour plus de détails). Nous avons également développé un outil de prédiction dynamique, *DynPG* disponible en ligne (`www.labcom-risca.com`), permettant de mettre à jour la prédiction d'un échec de greffe rénale en fonction du suivi longitudinal de sérum de créatinine (voir la section 9.2 pour plus de détails) [Fournier et al., 2019].

Le travail de thèse du Dr. Marie-Cécile Fournier a fait l'objet de trois articles publiés :

- **Fournier MC**, Foucher Y, Blanche P, Buron F, Giral M and **Dantan E**. A joint model for longitudinal and time-to-event data to better assess the specific role of donor and recipient factors on long-term kidney transplantation outcomes. *Eur J Epidemiol.* 2016 May ;31(5) :469-79. Epub 2016 Feb 1.
- **Fournier MC**, **Dantan E** and Blanche P. An R2 -curve for evaluating the accuracy of dynamic predictions. *Stat Med.* 2018 Mar 30 ;37(7) :1125-1133. Epub 2017 Dec 4.
- **Fournier MC**, Foucher Y, Blanche P, Legendre C, Girerd S, Ladrière M, Morelon E, Buron F, Rostaing L, Kamar N, Mourad G, Garrigue V, Couvrat-Desvergnes G, Giral M, and **Dantan E**. Dynamic predictions of long-term kidney graft failure : a cohort-based validation study. *Nephrology Dialysis Transplantation.* 2019 Mar 11, doi : 10.1093/ndt/gfz027.

Par ailleurs, plusieurs communications orales et affichées ont également été associées au

travail de Marie-Cécile Fournier :

- **Fournier MC**, Foucher Y, Giral M and **Dantan E**. Dynamic predictions of long-term kidney graft failure : a tool to better inform the recipient in a shared decision making approach. Ecole d'Automne du DHU 2020, Nantes, 2018.
- **Fournier MC**, Foucher Y, Giral M and **Dantan E**. Une application Shiny pour prédire dynamiquement la survie patient-greffon d'un patient transplanté rénal : vers une décision médicale plus partagée. 7ème rencontres R, Rennes, France, 2018.
- **Fournier MC**, Foucher Y, Giral M and **Dantan E**. Prédictions dynamiques du risque d'échec de greffe : un outil pour aider à l'information des patients transplantés rénaux et une prise en charge centrée sur le patient et ses attentes. 12e conférence francophone d'épidémiologie clinique (EPICLIN) et 25e journées des statisticiens des centres de lutte contre le cancer, Nice, France, 2018.
- **Fournier MC**, Blanche P, Foucher Y, Giral M and **Dantan E**. Construction d'un score pronostique dynamique et évaluation de ses performances : application en transplantation rénale. 10e conférence francophone d'épidémiologie clinique (EPICLIN 2016) et 23e journées des statisticiens des centres de lutte contre le cancer, Strasbourg, France, 2016.
- **Fournier MC**, **Dantan E** and Blanche P. R2-type curve to evaluate dynamic predictions. International Society for Clinical Biostatistics, Birmingham, England, 2016.
- Blanche P, **Fournier MC** and **Dantan E**. R2-type Curves for Dynamic Predictions from Joint Longitudinal-Survival Models : Inference & application to prediction of kidney graft failure. Workshop on Flexible Models for Longitudinal and Survival Data with Applications in Biostatistics, Warwick, England, 2015.
- **Fournier MC**, Foucher Y, Giral M and **Dantan E**. Etude des déterminants de santé associés à l'évolution de la créatininémie et/ou du risque d'échec de greffe chez des patients transplantés rénaux : une approche par modélisation conjointe. GDR Statistique & Santé, Paris, France, 2015.
- **Fournier MC**, Foucher Y, Giral M and **Dantan E**. Etude des déterminants de santé associés à l'évolution de la créatininémie et/ou du risque d'échec de greffe chez des patients transplantés rénaux : une approche par modélisation conjointe. EPICLIN, Montpellier, France, 2015.
- **Fournier MC**, **Dantan E**, Foucher Y and Giral M. A joint survival longitudinal approach to better assess the etiological role of risk factors of kidney transplant graft failure. American transplant congress, Philadelphia, USA, 2015.

2.1.3 Florence Gillaizeau

Débutée en septembre 2012, Florence Gillaizeau a soutenu sa thèse de Biostatistique intitulée *Développement et validation de modèles multi-états semi-Markoviens pour le pronostic de*

patients atteints de maladies chroniques le 15 octobre 2015. Son projet de recherche a été financé par le PHRC national MaKiT (*Mortality after Kidney Transplantation*, PROG/11/85, 2011). Sa thèse a été dirigée par le Dr. Yohann Foucher (INSERM UMR 1246 - SPHERE) et le Pr. Magali Giral (INSERM UMR 1064). J'en fus le co-encadrant à hauteur de 20%.

L'objectif était de poursuivre les développements méthodologiques initiés par Yohann Foucher sur les modèles multi-états [Foucher et al., 2007, 2010]. Deux extensions ont été proposées aux modèles semi-Markoviens. Dans un premier travail, Florence Gillaizeau a développé un modèle multi-états semi-markovien de survie relative à risques additifs afin d'évaluer si certains déterminants de santé étaient associés à un excès de risque de décès spécifiquement lié à la pathologie [Gillaizeau et al., 2017]. Dans un second travail, il a été proposé de simplifier la mise en œuvre des modèles multi-états en utilisant la méthode de pondération des individus par le score de propension (Inverse Probability Weighting, IPW) comme méthode d'ajustement [Gillaizeau et al., 2018]. Ces extensions ont été évaluées à partir de simulations et appliquées sur données réelles en transplantation rénale et en maladies cardiaques valvulaires.

Le travail de thèse du Dr. Florence Gillaizeau a fait l'objet de trois articles :

- **Gillaizeau F**, Sénage T, Le Borgne F, Le Tourneau T, Roussel JC, Leffondré K, Porcher R, Giraudeau B, **Dantan E** and Foucher Y. Inverse probability weighting to control confounding in an illness-death model for interval-censored data. *Stat Med.* 2018 Apr 15 ;37(8) :1245-1258. Epub 2017 Dec 4.
- **Gillaizeau F**, **Dantan E**, Giral M and Foucher Y. A multistate additive relative survival semi-Markov model. *Stat Methods Med Res.* 2017 Aug ;26(4) :1700-1711. Epub 2015 Jun 7.
- **Gillaizeau F**, Giral M, **Dantan E**, Dragun D, Soulillou JP and Foucher Y. Multi-state analysis of kidney transplant recipients outcome : a semi-Markov model for studying the role of pre-transplant sensitization against Angiotensin II Type 1 receptor. *Journal de la Société Française de Statistique.* 2014 ;155(1) :117-133.

Plusieurs communications orales ont également été associées au travail de Florence Gillaizeau :

- **Gillaizeau F**, Sénage T, Le Tourneau T, Roussel JC and Foucher Y. La détérioration de bioprothèse valvulaire : un problème majeur de santé publique largement sous-estimé par la non-prise en compte de la censure par intervalle et l'évolution multi-états de la maladie. EPICLIN, Montpellier, France, 2015.
- **Gillaizeau F**, **Dantan E**, Giral M and Foucher Y. Additive relative survival multistate semi-Markov model. International Society for Clinical Biostatistics, Vienna, Austria, 2014.
- **Gillaizeau F**, **Dantan E**, Giral M and Foucher Y. A multistate additive relative survival semi-Markov model. 46e Journées de Statistique de la SFdS, Rennes, France, 2014.
- **Gillaizeau F**, Foucher Y, Dragun D, Soulillou JP, Dufay A, Van Huyen J, Renaudin K, Moreau A, **Dantan E** and Giral M. A multistate model to investigate the relationship

between pre-graft level of Angiotensin II Type 1 Receptor (AT1R) antibodies and kidney transplant recipients outcome. American Transplant Congress, Seattle, USA, 2013.

- **Gillaizeau F**, Giral M, **Dantan E**, Dragun D, Soulillou JP and Foucher Y. A Semi-Markov model with Generalized Weibull distribution for multistate data in kidney transplant recipients. Journées de la Société Française de Biométrie, Paris, France, 2012.
- **Gillaizeau F**, Dragun D, Soulillou JP, Dufay A, Van Huyen J, Renaudin K, Moreau A, Giral M and Foucher Y. A Semi-Markov model with Generalized Weibull distribution for multistate data in kidney transplant recipients. Congrès annuel de la Société Francophone de Transplantation (SFT), Nantes, France, 2012.

2.2 Étudiants en Master 2

2.2.1 Julie Havas - Master 2 de Biostatistique de l'Université de Bordeaux (ISPED), 2017

L'objectif de ce stage était de modéliser conjointement l'évolution longitudinale des performances hémodynamiques à partir des gradients transprothétiques et le risque de survenue d'un décès chez des patients porteurs d'une bioprothèse valvulaire aortique. A l'aide d'un modèle conjoint à effets aléatoires partagés, nous nous sommes particulièrement intéréssés au rôle spécifique d'une inadéquation de taille entre la prothèse et le patient. Ce travail a été réalisé à partir de la cohorte Cordabase de 613 patients ayant eu un remplacement valvulaire aortique dans le service de chirurgie cardiaque du CHU de Nantes entre 2002 et 2007.

2.2.2 Angelina Dion - Master 2 Modélisation en Pharmacologie Clinique et Epidémiologie de l'Université de Nantes, 2015

En transplantation rénale, l'utilité clinique des biopsies de surveillance à 1 an post-transplantation reste débattue. Les pratiques des centres de transplantation restent très hétérogènes vis à vis de cet acte diagnostique invasif. L'objectif du stage d'Angelina Dion a été de comparer la survenue d'un échec de greffe chez des patients transplantés rénaux stables à 1 an post-greffe entre les centres transplanteurs ayant une politique de biopsie de surveillance et ceux n'en ayant pas. A partir de la cohorte de patients transplantés rénaux DIVAT (Données Informatisées et VAlidées en Transplantation, `www.divat.fr`), nous n'avons pas pu observer de différence significative du risque d'échec de greffe entre les deux types de centres de transplantation [Couvrat-Desvergnes et al., 2018].

- Couvrat-Desvergnes G, Foucher Y, Le Borgne F, **Dion A**, Mourad G, Garrigue V, Legendre C, Rostaing L, Kamar N, Kessler M, Ladrière M, Morelon E, Buron F, Giral M and **Dantan E**. Comparison of graft and patient survival according to the transplantation

center policy for 1-year screening biopsy among stable kidney recipients : a propensity score-based study. *Nephrology Dialysis Transplantation*. 2018 July 14.

A la suite de son stage, Angelina Dion a pu être embauché comme ingénieur d'étude pour une durée de 3 mois à l'aide d'un financement de l'entreprise pharmaceutique Chiesi. Elle a ainsi pu contribuer aux travaux de recherche initiés en collaboration avec le Dr. Anne-Hélène Quérard sur la problématique des reins greffés marginaux [Querard et al., 2016]. Dans une approche multivariée classique, nous validons l'estimation conditionnelle du critère ECD sur le risque d'échec de greffe. A l'aide d'une approche par pondération IPW, nous estimons finalement un effet absolu (marginal) du critère ECD sur l'échec de greffe rénale moins fort que celui initialement mis en évidence [Querard et al., 2018]. Cela pourrait constituer un argument en faveur d'une extension des critères de marginalité du greffon permettant ainsi un élargissement du pool de greffon (voir le chapitre 8 pour plus de détails).

- Quérard AH, Le Borgne F, **Dion A**, Giral M, Mourad G, Garrigue V, Rostaing L, Kamar N, Loupy A, Legendre C, Morelon E, Buron F, Foucher Y and **Dantan E**. Propensity score-based comparison of the graft failure risk between kidney transplant recipients of standard and expanded criteria donor grafts : Toward increasing the pool of marginal donors. *Am J Transplant*. 2018 May ;18(5) :1151-1157. Epub 2018 Feb 8.
- Quérard AH, Le Borgne F, **Dion A**, Giral M, Mourad G, Garrigue V, Rostaing L, Kamar N, Loupy A, Legendre C, Morelon E, Buron F, Foucher Y and **Dantan E**. Validation of the Kidney Donor Risk Index to graft failure prognostic from a French kidney transplant recipient cohort. *En cours de rédaction*.

2.2.3 Amina Blidi - Master 2 de Santé Publique (parcours Biostatistique) de l'Université de Paris Descartes, 2012

Le temps d'ischémie froide représente l'intervalle de temps entre le prélèvement du greffon et la greffe. De nombreuses études en transplantation rénale ont démontré un lien entre le temps d'ischémie froide et l'incidence du retard au démarrage du rein greffé (Delayed Graft Function). L'objectif du stage d'Amina Blidi a été de préciser l'effet du temps d'ischémie froide sur la survie patient/greffon. A partir de la cohorte de patients transplantés rénaux DIVAT, nous avons pu conclure que chaque heure supplémentaire d'ischémie froide semble augmenter le risque d'échec de greffe [Debout et al., 2015].

- Debout A, Foucher Y, Trébern-Launay K, Legendre C, Kreis H, Mourad G, Garrigue V, Morelon E, Buron F, Rostaing L, Kamar N, Kessler M, Ladrière M, Poignas A, **Blidi A**, Soulillou JP, Giral M and **Dantan E**. Each additional hour of cold ischemia time significantly increases the risk of graft failure and mortality following renal transplantation. *Kidney Int*. 2015 Feb ;87(2) :343-9. Epub 2014 Sep 17.

2.2.4 Damien Drubay - Master 2 de Biostatistique de l'Université de Bordeaux (ISPED), 2011

L'objectif du stage de Damien Drubay était de modéliser conjointement l'évolution longitudinale de la fonction rénale et la survie patient/greffon chez des patients transplantés rénaux. A l'aide d'un modèle conjoint à effets aléatoires partagés, nous nous sommes particulièrement intéréssés au rôle spécifique du poids du rein. Ce travail a été réalisé à partir de la cohorte observationnelle et prospective de patients transplantés rénaux DIVAT.

2.3 Étudiants en Master 1

2.3.1 Alpha Barry - Master 1 de Bioinformatique/Biostatistique de l'Université de Nantes, 2017

L'objectif de ce stage était de décrire le niveau de protéinurie à un an post-transplantation rénale, ainsi que d'étudier les paramètres du receveur, du donneur et de la transplantation qui pourraient être associés à la protéinurie. Dans la cohorte DIVAT, ce paramètre est mal renseigné. Un objectif secondaire a été d'étudier les caractéristiques de la transplantation rénale pouvant être associées au risque d'observer une donnée manquante pour la protéinurie à un an post-transplantation.

2.3.2 Alexandre Huneau - Master 1 de Bioinformatique/Biostatistique de l'Université de Nantes, 2014

L'objectif de ce stage était de construire un score diagnostique des lésions histologiques anormales sur la biopsie du greffon à un an post-transplantation rénale. Ce travail a fait l'objet d'une publication [Giral et al., 2018] et du développement d'une application en ligne (`www.labcom-risca.com`) pour une aide à l'indication d'une biopsie du greffon à un an de greffe.

- Giral M, Renaudin, K, Naesens M, Luning R, Anglicheau D, Morelon E, **Huneau A**, Paul C, Brouard S, Couvrat-Desvergnes G, Foucher Y and **Dantan E**. The 1-year Renal Biopsy Index (1-RBI) : a scoring system to drive biopsy indication at 1-year post-kidney transplantation. *Transpl Int.* 2018 June 7 ;31 :947-955.

2.3.3 Alexandra Bruel - Master 1 de Biologie Santé (parcours Sciences médicales) de l'Université de Nantes, 2013

Le stage d'Alexandra Bruel a porté sur l'étude des facteurs de risque d'insuffisance rénale chronique chez les patients porteurs de myosites inflammatoires. Ce travail a été réalisé en collaboration avec le Dr. Couvrat-Desvergnes et a donné lieu à une publication [Couvrat-desvergnes et al., 2014].

- Couvrat-Desvergnes G, Masseau A, Benveniste O, **Bruel A**, Hervier B, Mussini JM, Buob D, Hachulla E, Matignon M, Azar R, Mac Namara E, MacGregor B, Daniel L, Lacraz A, De Broucker T, Rouvier P, Carli P, Laville M, **Dantan E**, Hammidou M, Moreau A and Fakhouri F. The Spectrum of Renal Involvement in Patients With Inflammatory Myopathies. *Medicine*. 2014 Jan ;93(1) :33-41.

2.3.4 Olivier Merdy - Master 1 de Biologie Santé (parcours Sciences médicales) de l'Université de Nantes, 2013

L'objectif du stage d'Olivier Merdy a été de réaliser une étude transversale de la fréquence des douleurs musculosquelettiques attribuées aux techniques manuelles reçues en cours d'apprentissage chez les étudiants en ostéopathie.

- Chapitre 3 -

Synthèse des activités de recherche passées

Après avoir obtenu en 2005 un Master en Modélisation mathématique, j'ai souhaité rapprocher mon parcours universitaire des applications en santé. C'est pourquoi j'ai choisi de poursuivre et compléter ma formation en validant en 2006 un second Master en Epidémiologie et Biostatistique. C'est véritablement lors du stage de six mois clôturant cette formation que j'ai débuté mes travaux de recherche et pris goût aux problématiques méthodologiques dans l'analyse de données. Avec l'obtention d'une bourse ministérielle au concours de l'école doctorale, j'ai réalisé une thèse de Biostatistique dirigée par Hélène Jacqmin-Gadda, au sein de l'équipe INSERM U897 à l'Université de Bordeaux, et soutenue le 8 décembre 2009. D'abord ATER, puis nommé sur un poste de maître de conférence, je poursuis, depuis septembre 2010, mes activités de recherche à l'Université de Nantes au sein de l'équipe INSERM UMR 1246 - SPHERE.

3.1 Histoire naturelle de l'évolution de maladies chroniques

Je m'intéresse depuis le début de mes activités de recherche en biostatistique aux méthodes permettant d'étudier l'histoire naturelle de l'évolution de maladies chroniques. Deux approches en particulier ont fait l'objet de développements auxquels j'ai pu contribuer : les modèles conjoints pour données longitudinales et données de survie et les modèles multi-états.

Mon travail de thèse s'intitule *Modèles conjoints pour données longitudinales et données de survie incomplètes appliqués à l'étude du vieillissement cognitif*. Dans l'étude du vieillissement cérébral, le suivi des personnes âgées est soumis à une forte sélection avec un risque de décès associé à de faibles performances cognitives. La modélisation de l'histoire naturelle du vieillissement cognitif est complexe du fait de données longitudinales et données de survie incomplètes. Par ailleurs, un déclin accru des performances cognitives est souvent observé avant le diagnostic de démence sénile, mais le début de cette accélération n'est pas facile à identifier. Les profils

d'évolution peuvent être variés et associés à des risques différents de survenue d'un événement ; cette hétérogénéité des déclins cognitifs de la population des personnes âgées doit être prise en compte. Mon travail de thèse avait pour objectif d'étudier des modèles conjoints pour données longitudinales et données de survie incomplètes afin de décrire l'évolution cognitive chez les personnes âgées. L'utilisation d'approches à variables latentes a permis de tenir compte de ces phénomènes sous-jacents au vieillissement cognitif que sont l'hétérogénéité et l'accélération du déclin. Au cours d'un premier travail, j'ai pu comparer deux approches afin de prendre en considération les sorties d'étude informatives dans l'analyse longitudinale du vieillissement cérébral mesuré au travers de différents tests psychométriques [Dantan et al., 2008]. Dans un second travail, nous proposons un modèle conjoint à état latent pour modéliser simultanément l'évolution cognitive et son accélération pré-démentielle, le risque de démence et le risque de décès [Dantan et al., 2011] (Figure 3.1).

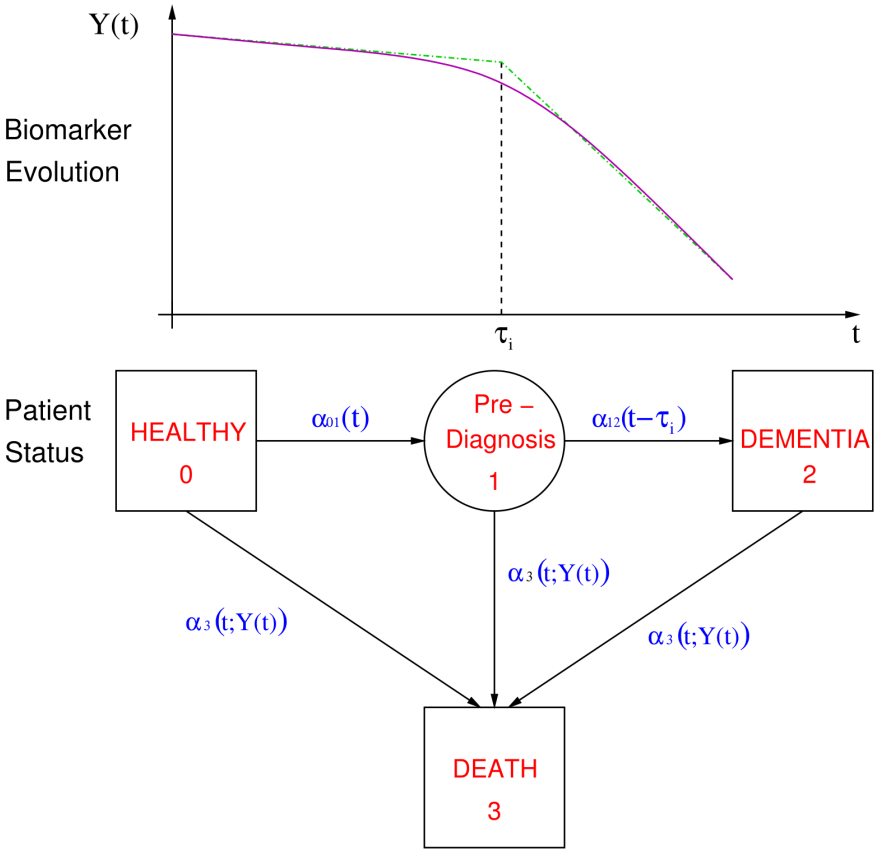


Figure 3.1: *Modèle conjoint à état latent pour l'accélération du déclin cognitif, la survenue d'une démence et la survenue d'un décès, avec α_{01}, α_{12} et α_3 les intensités de transitions entre les differents états (voir Dantan et al. [2011] pour plus de détails).*

L'expertise acquise sur les modèles conjoints pour données longitudinales et données de survie m'a permis d'encadrer la thèse de Marie-Cécile Fournier sur cette thématique appliquée à la transplantation rénale (voir la section 2.1.2 et le chapitre 5 pour plus de détails). Nous nous sommes notamment intéressés à l'utilisation d'un modèle conjoint à effets aléatoires partagés

dans une perspective étiologique. Plus précisément, il s'agissait d'identifier les facteurs associés à l'évolution longitudinale du sérum de créatinine et/ou au risque d'échec de greffe [Fournier et al., 2016]. J'ai également encadré les stages de Master de Damien Drubay et Julie Havas portant sur l'utilisation de modèles conjoints à effets aléatoires partagés dans les domaines de la transplantation rénale et de la chirurgie cardiaque, respectivement (voir la section 2.2 pour plus de détails).

Dans le domaine médical, un critère de jugement fréquemment rencontré est la survenue d'un événement clinique d'intérêt. Le modèle de Cox est le modèle de survie le plus fréquemment utilisé pour étudier les facteurs de risque associés à la probabilité de survenue de l'événement [Cox, 1972]. En présence de risques compétitifs où plusieurs événements sont en concurrence [Gail, 1975], plusieurs approches peuvent être envisagées. Les modèles multi-états sont une généralisation des modèles de survie permettant d'étudier l'évolution des patients au travers différents états de santé et de se focaliser sur les risques de transition entre les états. Ils ont notamment pour avantage d'estimer sans biais la probabilité cumulée de survenue d'un événement. Dans le cadre du vieillissement cérébral, nous avons développé un modèle conjoint pour données longitudinales et données de survie pour lequel le sous-modèle de survie était un modèle multi-état (Figure 3.1) afin d'étudier l'effet de covariables sur chacune des transitions vers une démence d'Alzheimer ou un décès. Deux particularités méthodologiques de ce modèle multi-état sont à souligner : l'état pré-démence est un état latent, la vraisemblance de ce modèle permet de tenir compte de la censure par intervalle pour la survenue d'une démence d'Alzheimer.

En transplantation rénale, l'un des objectifs est de retarder la survenue d'un échec de greffe défini par le retour en dialyse du patient transplanté et d'étudier les facteurs associés à la perte du greffon. Dans ce contexte, le décès constitue un événement en compétition avec la survenue d'un retour en dialyse. Plusieurs travaux de Yohann Foucher ont montré l'intérêt des modèles multi-états semi-markoviens en transplantation rénale. Les modèles semi-markoviens font l'hypothèse de risques de transition entre deux états qui dépendent principalement du temps passé dans le premier état [Foucher et al., 2007, 2010]. Dans le cadre de la thèse de Florence Gillaizeau co-encadrée avec Yohann Foucher et Magali Giral, nous avons pu proposer deux extensions aux modèles multi-état semi-markoviens. Dans un premier travail, Florence Gillaizeau a développé un modèle multi-états semi-markovien de survie relative à risques additifs afin d'évaluer si certains déterminants de santé sont associés à un excès de risque de décès spécifiquement lié à la pathologie [Gillaizeau et al., 2017]. Dans un second travail, il a été proposé de simplifier la mise en œuvre des modèles multi-états en utilisant la méthode de pondération des individus par le score de propension (IPW) comme méthode d'ajustement [Gillaizeau et al., 2018].

3.2 Recherches tournées vers les outils prédictifs

Dans le contexte de maladies chroniques, les outils de prédictions d'événements cliniques deviennent extrêmement populaires. Ces outils pourraient aider les patients et médecins dans une décision médicale partagée et faciliter la mise en place d'une médecine dite prédictive, personnalisée, préventive et participative (i.e. médecine 4P) [Flores et al., 2013]. Contrairement aux modèles étiologiques qui s'attachent à l'identification des relations entre un ensemble de covariables et un ensemble de variables réponses, les modèles de prédiction ont pour seul objectif d'estimer la probabilité d'avoir un événement d'intérêt [Tripepi et al., 2008]. Cette probabilité estimée constitue une information importante pour les patients désireux de connaître les risques qu'ils encourent. Fournir une telle estimation aux patients pourraient leur permettre d'améliorer la connaissance de leur pathologie, de leur possible devenir et par conséquence les inciter à s'investir d'avantage dans le suivi de leur maladie en étant par exemple plus adhérants aux traitements proposés [Gordon et al., 2013; Jamieson et al., 2016]. Dans une telle approche, l'interprétation étiologique des facteurs ayant été retenus comme facteurs prédictifs n'est que secondaire. Toutefois, l'utilisation en pratique clinique de tels outils diagnostiques/pronostiques repose en partie sur leur capacité à correctement prédire l'événement d'intérêt. Un score prédictif, qu'il soit diagnostique ou pronostique, doit avoir de bonnes propriétés de calibration et de discrimination [Graf et al., 1999; Steyerberg et al., 2010].

La discrimination est la capacité d'un outil de prédiction à différencier les sujets susceptibles ou non de vivre l'événement d'intérêt. En d'autres termes, un outil hautement discriminant estimera des prédictions de survenue de l'événement très différentes entre des individus susceptibles ou non de connaître l'événement. L'indicateur le plus classique évaluant les performances discriminantes est l'aire sous la courbe ROC (i.e. AUC under the ROC curve). Au sein de notre groupe de travail, plusieurs travaux de recherche en Biostatistique et Epidémiologie se sont intéressés à l'évaluation des capacités pronostiques discriminantes d'un score ou marqueur, avec notamment le développement des courbes ROC dépendantes du temps, nettes, ajustées ou agrégées [Foucher et al., 2010; Foucher and Danger, 2012; Lorent et al., 2014; Le Borgne et al., 2017; Combescure et al., 2014, 2016]. La calibration est la capacité d'un outil à faire coïncider prédictions et observations. Elle est généralement étudiée à l'aide d'approches graphiques représentant les probabilités observées en fonction de quantiles de prédictions. D'autres indicateurs évaluent les performances globales d'un outils de prédiction. Par exemple, le Brier Score est défini comme une erreur quadratique entre l'observation et la prédiction [Gerds and Schumacher, 2006; Blanche et al., 2015].

Concernant mes travaux dans ce domaine, j'ai d'abord proposé un article publié dans une revue d'épidémiologie [Dantan et al., 2014] faisant le point sur la confusion répandue entre corrélation et capacité prédictive. Dans cet article, nous proposons une application en ligne, EVALBIOM (`www.divat.fr/en/online-calculators/evalbiom`), permettant de calculer diffé-

rents indicateurs pronostiques dépendants du temps (sensibilité, spécificité, valeurs prédictives) à partir de courbes de survie déjà publiées (voir le chapitre 4 pour plus de détails). Dans un autre travail publié dans *Statistics in Medicine*, nous nous intéressons à l'évaluation des capacités pronostiques d'un score dynamique. Nous avons proposé un critère de type R^2 [Fournier et al., 2018] qui permet de solutionner l'une des limites du Brier Score : sa dépendance à la probabilité marginale de survenue de l'événement (voir le chapitre 6 pour plus de détails).

En lien avec mes domaines de recherche, j'ai réalisé plusieurs travaux appliqués. J'ai, par exemple, proposé un score diagnostique d'anormalité histologique sur la biopsie à un an post-transplantation rénale, le 1-RBI [Giral et al., 2018]. Cet outil pourrait permettre de proposer une biopsie aux patients les plus susceptibles de présenter les lésions histologiques majeures. Toujours en transplantation rénale, la thèse de Marie-Cécile Fournier avait pour objectif final de proposer un score pronostique dynamique de l'échec de greffe. Nous proposons d'inclure les mesures répétées de sérum de créatinine en plus de caractéristiques initiales (âge du receveur, rang de greffe, immunisation pré-greffe anti-HLA, survenue d'au moins un rejet aigu dans la première année post-transplantation, antécédents cardiovasculaires) afin d'obtenir des prédictions dynamiques de l'échec de greffe pouvant être mises à jour tout au long du suivi du patient (voir la section 9.2 pour plus de détails). Un modèle conjoint pour données répétées de sérum de créatinine et risque de survenue d'un échec de greffe a été estimé sur une cohorte d'apprentissage de 2584 patients transplantés rénaux. A partir du modèle estimé, nous pouvons calculer les prédictions individuelles d'échec de greffe à différents temps de suivi. Les capacités pronostiques ont été étudiées à partir d'une cohorte de validation constituée de 2589 patients. Pour les différents temps de prédiction, les capacités de discrimination et de calibration apparaissent raisonnables. Une application SHINY est disponible en ligne afin de calculer ces prédictions dynamiques (`https://shiny.idbc.fr/dynpg`). Ce travail vient d'être accepté dans dans une revue de transplantation [Fournier et al., 2019].

3.3 Vers des approches méthodologiques centrées sur le patient

Mon activité de recherche s'inscrit au sein de l'équipe INSERM UMR 1246 - SPHERE, équipe pluridisciplinaire constituée de biostatisticiens, méthodologistes, épidémiologistes, économistes de la santé, cliniciens. Le projet de recherche porté par notre équipe vise à renforcer et développer des méthodes dans le domaine de la recherche en santé centrée sur le patient.

Comme énoncé dans un article publié dans le *Journal of the American Medical Association* en 2012, les "patients ne se distinguent pas seulement par leur variabilité biologique, ils diffèrent également selon la manière dont la maladie peut affecter leur vie" [Methodology Committee of the Patient-Centered Outcomes Research Institute (PCORI), 2012; Frank et al., 2014]. Les

dimensions pychologiques, culturelles, comportementales, environnementales et économiques apparaissent être des facteurs influençant l'évolution de l'état de santé des personnes. Mais prendre en considération ces caractéristiques du point de vue du patient constitue un challenge dans la recherche en santé. Les chercheurs, décideurs en santé et cliniciens font face à différents problèmes conceptuels, méthodologiques et pratiques notamment concernant la mesure et l'interprétation du vécu des patients et de leur adaptation à la pathologie, leur relation avec l'environnement, leurs préférences et sentiments, ou encore l'impact de leur perception de la maladie sur leur pronostic. Des travaux de recherche sur la planification d'études pour comprendre ces concepts complexes et sur l'identification des approches méthodologiques adéquates pour l'analyse et l'interprétation de ces données semblent nécessaires. Le projet scientifique de SPHERE repose sur trois axes de recherche : 1) la mesure et l'interprétation de la perception et le vécu du patient, 2) l'évaluation en vie réelle, et 3) la médecine de précision. Mes travaux de recherche s'inscrivent plus particulièrement dans l'axe 3.

La *médecine de précision* a pour objectif de fournir le bon traitement, pour la bonne personne, au bon moment. Cela repose en partie sur la capacité d'un score, marqueur ou modèle de prédiction à stratifier les patients selon les meilleures options thérapeutiques. Toutefois, alors que le développement d'outils de prédiction ne cesse de croître, un très faible pourcentage est finalement transféré en pratique clinique [Joyner and Paneth, 2015]. Ce problème pourrait être dû en partie au manque de considération des patients eux-mêmes. L'un des objectifs des recherches centrées sur le patient est de permettre aux voix des patients et aidants d'être entendues notamment dans l'évaluation des options thérapeutiques et de leurs conséquences [Ziegelstein, 2015]. Plusieurs de nos travaux soutiennent la mise en place d'une prise en charge stratifiée et centrée sur le patient. En particulier, nous nous intéressons à la notion d'utilité d'une décision médicale en termes de QALYs (Quality Adjusted Life-Years). Les QALYs sont des indicateurs combinant en une seule mesure des informations sur la quantité de vie et la qualité de vie. Ils se calculent en pondérant chaque période de temps par un score d'utilité reflétant les préférences des patients pour les différents états de santé [Weinstein et al., 2009]. Les utilités peuvent être obtenues après conversion des résultats de questionnaires de qualité de vie tel que le SF36 ou EQ5D. Un score de 0 correspond au décès et un score de 1 à l'état de parfaite santé. Bien que les QALYs aient été initialement proposés pour l'évaluation médico-économique, leur nature composite les rends très intéressants pour une prise de décision plus proche de préférences des patients, comme nous le suggérons dans une mini-revue de la littérature des critères de jugement des essais randomisés en cancer de la prostate [Foucher et al., 2018].

Dans un travail méthodologique publié dans *Statistical Methods and Medical Research*, nous définissons une fonction d'utilité dépendante du temps [Dantan et al., 2018]. La maximisation de cette fonction permet l'estimation d'un seuil de décision sur un marqueur quantitatif pour une prise en charge stratifiée. Plus précisément, la fonction d'utilité représente le nombre de QALYs espéré de la population en fonction du seuil utilisé pour définir les strates de traitements. Notre approche intègre à la fois les conséquences des prises en charge en terme de gain d'espérance

de vie ainsi que les conséquences en terme de qualité de vie. Nous avons appliqué ce travail en cancérologie, en collaboration avec l'Institut de Cancérologie de l'Ouest (ICO, Dr S. Supiot & Pr. JM Classe). En cancer de la prostate, même si nous validons le score CAPRA (Cancer of the Prostate Risk Assessment) pour ses bonnes aptitudes à prédire la survie libre de récidive, son utilité est toutefois discutable pour aider à la prise de décision entre surveillance active (pour les patients à faible risque), prostatectomie (pour les patients à risque intermédiaire), ou hormonothérapie (pour les patients à fort risque) [Lorent et al., 2019]. En revanche, en cancer du sein, nous validons l'utilité de la signature génomique Mammaprint®. Celle-ci pourrait éviter la chimiothérapie adjuvante chez les femmes ayant un test génomique négatif alors qu'elles sont à haut risque clinique [Caruana et al., 2019].

3.4 Travaux de recherche appliqués

Mes recherches en biostatistique et épidémiologie se veulent appliquées et tentent d'apporter des éléments de réponse à des problématiques méthodologiques rencontrées à partir de données réelles. Il est nécessaire d'avoir une compréhension assez fine des pathologies auxquelles seront appliqués ces développements. Depuis mon arrivée à Nantes, je développe donc une activité de recherche clinique. J'ai notamment pu acquérir des connaissances en épidémiologie de la transplantation rénale à partir de plusieurs travaux appliqués à la cohorte DIVAT [Cantarovich et al., 2013; Debout et al., 2015; Querard et al., 2016, 2018; Giral et al., 2018; Couvrat-Desvergnes et al., 2018].

La cohorte observationnelle DIVAT (Données Informatisées et VAlidées en Transplantation, `www.divat.fr`) est prospective et composée de patients transplantés rénaux. Il s'agit d'une cohorte multicentrique française avec 8 centres de transplantation (Nantes, Nancy, Toulouse, Montpellier, Paris Necker, Paris Saint-Louis, Lyon, Nice). Plus de 25000 patients ont été inclus depuis l'initiation du réseau en 1990 et DIVAT représente près de 40% des greffes rénales en France. Les paramètres du donneur, du receveur et plusieurs caractéristiques de la transplantation sont collectés. Des données sont aussi collectées au cours du suivi du patients à 3 et 6 mois post-transplantation puis à toutes les dates anniversaires de la greffe (serum de créatinine, protéinurie, etc.). Les patients sont suivis jusqu'à leur date de retour en dialyse ou décès. DIVAT regroupe ainsi l'ensemble des critères cliniques et biologiques utiles à la prise en charge et au suivi médical du patient transplanté rénal. Cette cohorte observationnelle s'avère être un formidable outil pour développer des travaux en épidémiologie de la transplantation rénale. Au sein d'un groupe de travail né de la collaboration historique entre Yohann Foucher et Magali Giral, néphrologue au sein de l'équipe INSERM 1064 et responsable scientifique de DIVAT, j'ai notamment travaillé sur la thématique des donneurs marginaux. Nous avons réalisé une revue de la littérature du critère ECD en transplantation rénale concluant au manque de validation externe de ce critère [Querard et al., 2016]. A partir de la cohorte DIVAT, nous validons l'estima-

tion conditionnelle du critère ECD sur le risque d'échec de greffe à l'aide d'un modèle de survie multivarié. A l'aide d'une approche par score de propension, nous estimons un effet marginal du critère ECD sur l'échec de greffe rénale moins fort que celui initialement mis en évidence [Querard et al., 2018], ce qui pourrait constituer un argument en faveur d'une extension des critères de marginalité du greffon permettant ainsi un élargissement du pool de greffon (voir le chapitre 8 pour plus de détails). Le critère ECD présente la limite d'être un critère binaire défini à partir de 4 variables du donneur (âge, antécédent d'hypertension, serum de créatinine, cause de décès). Un critère continu de marginalité du greffon composé de 10 variables du donneur, le Kidney Donor Risk Index (KDRI), a été récemment proposé aux Etats-Unis [Rao et al., 2009]. Un travail de validation du KDRI est en cours.

Dans le suivi du patient transplanté, les biopsies de surveillance peuvent être proposées à différents temps post-transplantation afin de détecter fortuitement des lésions histologiques du greffon. Leur intérêt clinique reste toutefois discutable pour plusieurs raisons. Tout d'abord, il ne semble pas y avoir actuellement de consensus sur les conduites thérapeutiques à tenir face aux différents types de lésions histologiques. Par ailleurs, une proportion importante de biopsies de surveillance ne révèle finalement aucune lésion histologique : le patient a donc été exposé inutilement à un acte médical invasif et dangereux pour son greffon. Cela explique l'hétérogénéité des pratiques de biopsie de surveillance entre les centres de transplantation, comme nous l'observons au sein du réseau DIVAT. A partir de DIVAT, nous n'avons pas mis en évidence de différence de risque de survenue d'un échec de greffe entre les centres transplanteurs ayant une politique de pratique de biopsie de surveillance à un an post-transplantation et ceux n'en ayant pas [Couvrat-Desvergnes et al., 2018]. D'un point de vue méthodologique, nous avons estimé un effet marginal d'une telle politique en réalisant une analyse de survie pondérée par un score de propension incluant les possibles facteurs de confusion. Le bénéfice d'une politique de dépistage systématique des lésions histologiques à partir de biopsie de surveillance ne paraît ainsi pas évident. Se pose la question de l'intérêt d'une politique de dépistage ciblée à partir d'une biopsie qui pourrait être proposée aux patients sans signes cliniques visibles mais tout de même susceptibles de présenter des lésions histologiques. Pour aider à l'indication d'une biopsie du greffon après une transplantation rénale, nous avons proposé un score d'anormalité histologique à 1 an post-transplantation, le 1-RBI (*1-year Renal Biospy Index*), présentant des capacités diagnostiques raisonnables [Giral et al., 2018].

Mes travaux de recherche appliqués ont essentiellement porté sur la transplantation rénale. Mais, je m'intéresse de plus en plus à d'autres pathologies. J'ai par exemple pu travailler sur une cohorte de patients ayant eu un remplacement valvulaire aortique (CORDABASE) dans le cadre de la thèse de Florence Gillaizeau ou le stage de Master 2 de Julie Havas. Nous débutons une collaboration sur le thème de la réanimation avec l'encadrement de la thèse de Jeanne Simon-Pimmel. Par ailleurs, certains de mes travaux *centrés sur le patient (i.e. patient-centered)* ont été développés dans le contexte de la cancérologie (voir la section 3.3 pour plus de détails).

Deuxième partie

Cinq publications représentatives

- Chapitre 4 -

Evaluation des capacités pronostiques d'un marqueur à partir de courbes de survie publiées

Dantan E, Combescure C, Lorent M, Ashton-Chess J, Daguin P, Classe JM, Giral M and Foucher Y. An original approach was used to better evaluate the capacity of a prognostic marker using published survival curves. *J Clin Epidemiol.* 2014 Apr ;67(4) :441-8.

Préambule

Prédire l'évolution de maladies chroniques est un important sujet de recherche en épidémiologie clinique. Dans la littérature médicale, un marqueur est souvent considéré comme pronostique s'il est significativement associé au risque de survenue d'un événement. Par exemple, la distance entre deux courbes de survie définies en fonction d'un marqueur binaire ou encore la significativité d'un rapport de risque dans un modèle de Cox sont souvent interprétées en terme de capacités pronostiques. Pour autant, ces indicateurs ne sont que le reflet d'une corrélation significative et ne peuvent suffire pour valider les capacités pronostiques du marqueur [Ware, 2006]. Cette confusion entre corrélation et pronostic est répandue [Pepe et al., 2004; Spruance et al., 2004].

Dans ce travail, nous proposons une réinterprétation des résultats d'articles déjà publiés présentant un marqueur binaire comme étant pronostique sans pour autant utiliser une méthodologie adéquate. Plus précisément, à partir d'équations simples, nous proposons de reconstruire les indicateurs de sensibilité et spécificité dépendantes du temps à l'aide d'informations disponibles sur deux courbes de survie. Il suffit de connaître les effectifs à l'inclusion des deux groupes, i.e. sujets à bas et haut risques de connaître l'événement, ainsi que les probabilités de survie de chacun des groupes pour un temps d'horizon choisi. Les valeurs prédictives positive et négative ainsi que les rapports de vraisemblance et les probabilités post-tests peuvent être également estimés. Nous pouvons ainsi compléter l'étude de marqueurs binaires sur leurs apti-

tudes à pronostiquer un événement. Nous avons proposés un outil disponible en ligne, EVALBIOM (`www.divat.fr/en/online-calculators/evalbiom`), pour rendre plus facilement calculable ces indicateurs dépendants du temps. Notre approche pourrait aider les lecteurs de ces articles publiés à mieux apprécier les performances pronostiques d'un marqueur binaire.

Trop souvent, dans des articles de recherche clinique, la méthodologie utilisée ne supporte par les conclusions avancées. Il est facile d'imaginer que ceci est lié en partie à une méconnaissance des méthodologies récentes adéquates de la part des investigateurs. Mon sentiment sur ce sujet est que les domaines de la Biostatistique, de l'Epidémiologie traditionnelle et de la Recherche clinique restent encore trop cloisonnés les uns par rapport aux autres. Par exemple, en 2000, Heagerty et al. [2000] a proposé dans *Biometrics* les courbes ROC dépendantes du temps pour évaluer les aptitudes discriminantes d'un marqueur en présence de données de survie censurées. Pour autant, comme illustré dans l'une des applications présentées dans notre article, 10 ans plus tard, on retrouve toujours des articles médicaux qui n'utilisent pas cet indicateur pour évaluer les capacités discriminantes de leur marqueur à l'étude [Mook et al., 2010]. Au-delà du développement de nouvelles méthodes statistiques parfois complexes, il me semble y avoir un besoin important de démocratisation et diffusion de ces approches récentes auprès de la communauté des épidémiologistes et méthodologistes qui seront en charge des analyses statistiques en recherche clinique. En ce sens, l'article publié dans le *Journal of Clinical Epidemiology* me semble intéressant à mettre en avant dans mes travaux de recherche [Dantan et al., 2014].

Journal of Clinical Epidemiology 67 (2014) 441−448

Journal of Clinical Epidemiology

An original approach was used to better evaluate the capacity of a prognostic marker using published survival curves

Etienne Dantan[a,*], Christophe Combescure[b], Marine Lorent[a], Joanna Ashton-Chess[c], Pascal Daguin[d], Jean-Marc Classe[e], Magali Giral[d,f], Yohann Foucher[a]

[a]*Department of Biostatistics, Pharmacoepidemiology and Subjective Measures in Health Sciences, EA 4275, Nantes University, 1 rue Gaston Veil, 44035 Nantes, France*

[b]*CRC & Division of Clinical Epidemiology, Department of Health and Community Medicine, University of Geneva and University Hospitals of Geneva, rue Gabrielle Perret-Gentil 4, 1211 Geneva, Switzerland*

[c]*TcLand Expression, 21 rue Lanoue Bras de Fer, 44200 Nantes, France*

[d]*Institute of Transplantation, Urology and Nephrology (ITUN), CHU Nantes and INSERM U1064, 30 bd Jean-Monnet, 44093 Nantes, France*

[e]*Department of Surgical Oncology, Institut de Cancérologie de l'Ouest—Cancer Center René Gauducheau, Boulevard Jacques Monod, 44805 Saint-Herblain, France*

[f]*CIC Biotherapy, CHU Nantes, 30 bd Jean-Monnet, 44093 Nantes, France*

Accepted 21 October 2013

Abstract

Objectives: Predicting chronic disease evolution from a prognostic marker is a key field of research in clinical epidemiology. However, the prognostic capacity of a marker is not systematically evaluated using the appropriate methodology. We proposed the use of simple equations to calculate time-dependent sensitivity and specificity based on published survival curves and other time-dependent indicators as predictive values, likelihood ratios, and posttest probability ratios to reappraise prognostic marker accuracy.

Study Design and Setting: The methodology is illustrated by back calculating time-dependent indicators from published articles presenting a marker as highly correlated with the time to event, concluding on the high prognostic capacity of the marker, and presenting the Kaplan−Meier survival curves. The tools necessary to run these direct and simple computations are available online at http://www.divat.fr/en/online-calculators/evalbiom.

Results: Our examples illustrate that published conclusions about prognostic marker accuracy may be overoptimistic, thus giving potential for major mistakes in therapeutic decisions.

Conclusion: Our approach should help readers better evaluate clinical articles reporting on prognostic markers. Time-dependent sensitivity and specificity inform on the inherent prognostic capacity of a marker for a defined prognostic time. Time-dependent predictive values, likelihood ratios, and posttest probability ratios may additionally contribute to interpret the marker's prognostic capacity. © 2014 Elsevier Inc. All rights reserved.

Keywords: Prognostic factor; Sensitivity; Specificity; Predictive values; Survival analysis; Likelihood ratios

1. Introduction

In many therapeutic areas, predicting health events is a real challenge to improve the long-term medical management of patients affected by chronic disease. In characterizing a biological or pathologic process, a surrogate marker may help to forecast a future event [1,2]. A marker is often claimed as prognostic if it is significantly associated with the time-to-event distribution. This is valid at a population level, but this does not imply that the marker is a useful tool for individual decision making. Rather, the clinical relevance of the marker should depend on its accuracy to predict a patient's evolution. Therefore, clinical and biological prognostic markers that can qualify a patient's likelihood of experiencing the event under consideration [3] are usually used to identify ''at risk'' patients who require more attentive follow-up or treatment adaptation. Understanding the term ''prediction'' as described in the literature can be baffling as this term can refer to the posterior estimation of a regression model such as the survival probability 5 years after treatment or the prognosis of a future event up to a certain prognostic time such as the death of a patient within the first 5 years after treatment. The present article concerns this

Conflict of interest: None.

Funding: This work was supported by a grant from the French National Research Agency ANR-11-JSV1-0008-01.

* Corresponding author. Tel.: +33 2 40 41 28 24; fax: +33 2 40 41 29 96.
E-mail address: Etienne.Dantan@univ-nantes.fr (E. Dantan).

0895-4356/$ - see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jclinepi.2013.10.022

What is new?

- Time-dependent sensitivity and specificity as well as time-dependent predictive values can be found from published survival curves.
- Time-dependent likelihood ratios and time-dependent posttest probability ratios are new indicators of prognostic marker accuracy.
- The online application available at http://www.divat.fr/en/online-calculators/evalbiom allows back calculating these time-dependent indicators.
- The proposed time-dependent indicators should help readers better evaluate clinical articles reporting on prognostic markers.

second definition. In the remainder of the article, the term ''time-dependent'' will be used for such prognosis up to a given prognostic time.

A widespread mistake when interpreting time-dependent data is confusing the notions of ''prediction'' and ''correlation'' [4]. The distance(s) between survival curves, the corresponding hazard ratio, and the associated *P*-value are often presented as the most popular indicators of predictive capacity. However, *P*-values only demonstrate that the relationship is not a result of sample-to-sample fluctuation. Moreover, in a diagnostic context, Pepe et al. [5] demonstrated that the magnitude of an odds ratio has to be huge for it to inform on predictive capacity. Ware [6] pursued this line of investigation in a prognostic context and reported that the hazard ratio is not synonymous of prognostic capacity. According to Spruance et al. [7], the hazard ratio between patients at high and low risk of experiencing an event is also often wrongly interpreted because it only reflects the magnitude of the changes in risk. Nevertheless, they demonstrated that the probability of experiencing the event sooner in the high-risk (HR) group compared with the low-risk (LR) group is equal to the hazard ratio divided by the hazard ratio plus one. Thus, the hazard ratio has to be massive to obtain a probability close to one, that is, the reference value, qualifying the prognostic abilities of a marker on a time-scale change. It is quite common to find that markers, defined by authors as prognostic, are in fact only correlated with the survival outcome. Such common presentation of results can lead to overoptimistic and even erroneous conclusions and consequently misinterpretations concerning the potential utility of the marker [8].

In contrast with the diagnostic context, sensitivity and specificity are not often used in prognostic studies. The reason for this may be the difficulty in calculating these values when dealing with the time-to-event censoring process involved in long-term studies, in which the study patients do not have the same follow-up time. Although Heagerty et al. [9] have described time-dependent estimators of sensitivity and specificity, these have not been widely adopted. Some authors estimate sensitivity and specificity based on the patients who have a follow-up at least equivalent to the prognostic time [10−12]. In this way, all patients who are censored before the prognostic time, that is, who do not have this minimum follow-up, are excluded from the analysis. This results in considerable selection bias, leading to an overrepresentation of patients with failure or patients who reached this minimum follow-up period.

As highlighted by Riley et al. [13], improving all aspects of prognosis research is necessary to give better clinically relevant evidence for clinicians and health policy makers. Over the past few years, new statistical methods have been developed to evaluate the prognostic capacity of markers. However, these methods are not systematically used in medical publications to justify conclusions relating to the prognostic capacity of the reported markers. Back calculation may be a solution to reappraise the conclusions reported in the published literature. For instance, Simel et al. [14] proposed an approach to calculate a posteriori sensitivity, specificity, and likelihood ratios when the odds ratio and marginal numbers of a contingency table are the only information available in a published article focusing on a diagnostic test. The aim of the present article was to enable readers to reinterpret the prognostic capacity of markers based on survival curves already published in the literature. Our proposed approach would enable the reader to determine the marker's true prognostic capacity. We use simple equations to calculate time-dependent sensitivity and specificity from survival curves. Other time-dependent indicators such as predictive values, likelihood ratios, or posttest probability ratios are also described. We additionally provide simple illustrations to interpret the true meaning of results relating to the prediction of long-term outcome.

2. Methods

2.1. Available information on published survival curves

The large majority of articles reporting on prognostic markers illustrate the results by plotting the unadjusted Kaplan−Meier survival curves. Fig. 1 illustrates the case of two survival curves for HR and LR groups defined for a binary marker. Here, NHR and NLR are the numbers of individuals at baseline classified in the HR and LR groups, respectively. SHR(t) and SLR(t) are the corresponding survival probabilities at time t for these same groups. If more than two marker-based groups are studied, the binary assumption of the prognostic test no longer meets the definition of traditional sensitivity and specificity in which the outcome is binary, for example, presence/absence of the disease. Nevertheless, in this context, it is straightforward to merge groups such as the HR and medium risk (MR). If the SMR(t) and NMR represent the survival probability at time t and the number of individuals at baseline in the

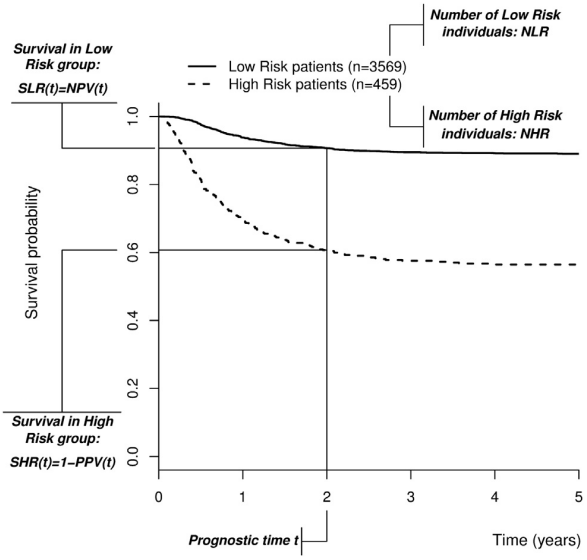


Fig. 1. Available information provided by most Kaplan–Meier survival curves. The four key parameters to be extracted are the baseline number of HR patients (NHR), the baseline number of LR patients (NLR), the survival at time t in the HR group (SHR), and the survival at time t in the LR group (SLR).

MR group, respectively, this new HR group can be characterized by a survival that is equal to [NHR × SHR(t) + NMR × SMR(t)]/(NHR + NMR). The MR group can also be merged with the LR group, which results in the calculation of two pairs of sensitivity and specificity values, one based on a strict and the other based on a lenient marker threshold.

2.2. Time-dependent sensitivity and specificity

The prognosis is made up to the prognostic time t. Then, $D(t)$ is the time-dependent indicator of the event with $D(t) = 1$ if the event occurs before t and $D(t) = 0$ otherwise. Based on the definitions of Heagerty et al. [9], the sensitivity at time t represents the proportion of patients who are correctly classified as HR among all patients who experience the event before time t, that is, $\mathrm{Se}(t) = P[\mathrm{HR}|D(t) = 1]$. The specificity at time t represents the proportion of patients who are correctly classified as LR among all patients who do not experience the event before t, that is, $\mathrm{Sp}(t) = P[\mathrm{LR}|D(t) = 0]$. Using the Bayes theorem (demonstration provided in the Appendix at www.jclinepi.com), time-dependent sensitivity and specificity for a prognosis up to time t can be easily calculated from survival at time t and the baseline numbers of individuals in the HR and LR groups:

$$\mathrm{Se}(t) = \frac{[1 - \mathrm{SHR}(t)] \times \mathrm{NHR}}{[1 - \mathrm{SHR}(t)] \times \mathrm{NHR} + [1 - \mathrm{SLR}(t)] \times \mathrm{NLR}} \quad (1)$$

$$\mathrm{Sp}(t) = \frac{\mathrm{SLR}(t) \times \mathrm{NLR}}{\mathrm{SHR}(t) \times \mathrm{NHR} + \mathrm{SLR}(t) \times \mathrm{NLR}} \quad (2)$$

2.3. Time-dependent predictive value

Although sensitivity and specificity indicate the intrinsic qualities of the marker by indicating risk group probabilities based on the future event status, predictive values are useful indicators for practical clinical interpretation and decision making. In fact, the time-dependent positive and negative predictive values, denoted PPV(t) and NPV(t), are directly obtained from the survival at time t. The positive predictive value is the probability that HR patients will experience the event before time t, that is, $\mathrm{PPV(t)} = P[D(t) = 1|\mathrm{HR}] = 1 - \mathrm{SHR}(t)$. The negative predictive value is the probability that LR patients will not experience the event before time t, that is, $\mathrm{NPV}(t) = P[D(t) = 0|\mathrm{LR}] = \mathrm{SLR}(t)$. The confidence intervals of each survival probability may sometimes be presented in the survival graphs or specified in the text for different time points. Therefore, the corresponding confidence interval for both predictive values can be obtained directly.

Time-dependent predictive values can also be defined as functions of time-dependent sensitivity and specificity (demonstration provided in the Appendix at www.jclinepi.com):

$$\mathrm{PPV}(t) = \frac{\mathrm{Se}(t) \times P(D(t) = 1)}{\mathrm{Se}(t) \times P[D(t) = 1] + [1 - \mathrm{Sp}(t)] \times \{1 - P[D(t) = 1]\}} \quad (3)$$

$$\mathrm{NPV}(t) = \frac{\mathrm{Sp}(t) \times \{1 - P[D(t) = 1]\}}{\mathrm{Sp}(t) \times \{1 - P[D(t) = 1]\} + [1 - \mathrm{Se}(t)] \times P[D(t) = 1]} \quad (4)$$

Regarding the previous equations, the main limitation of predictive values is their dependence on the population frailty, that is, the event probability. Therefore, without the calculation of previous time-dependent sensitivity and specificity values, the conclusions about prognostic capacity from a specific study cannot be directly generalized to other populations in which the event probability is different. This significant limitation of predictive values is well accepted in traditional diagnostic medicine but is always ignored in prognostic analyses.

2.4. Time-dependent likelihood ratios

In the diagnostic context, positive and negative likelihood ratios are well defined [15,16] and often used in practice. To our knowledge, their definitions have never been adapted to a prognostic context. Positive and negative

likelihood ratios for a prognosis up to time t, denoted $\mathrm{LikR}^+(t)$ and $\mathrm{LikR}^-(t)$, are respectively

$$\mathrm{LikR}^+(t) = \frac{P[\mathrm{HR}|D(t)=1]}{P[\mathrm{HR}|D(t)=0]} = \frac{\mathrm{Se}(t)}{1-\mathrm{Sp}(t)} \quad (5)$$

$$\mathrm{LikR}^-(t) = \frac{P[\mathrm{LR}|D(t)=1]}{P[\mathrm{LR}|D(t)=0]} = \frac{1-\mathrm{Se}(t)}{\mathrm{Sp}(t)}. \quad (6)$$

If time-dependent likelihood ratios are close to 1, then the classification rule tends not to be informative of the future event. The higher the positive likelihood ratio, the more the HR group probability is associated with the occurrence of the event before time t. Conversely, the lower the negative likelihood ratio, the more the LR group probability is associated with the absence of the event before time t. Given the thresholds used in diagnostic evaluations, positive and negative likelihood ratios close to 10 and 0.1, respectively, indicate a useful marker for a prognosis up to the time t [15]. Because time-dependent likelihood ratios are derived from sensitivity and specificity at time t, they are independent of the event probability and thus represent intrinsic characteristics of the marker's prognostic capacity.

2.5. Time-dependent posttest probability ratios

Let $\mathrm{PT}^+(t)$ and $\mathrm{PT}^-(t)$ be the positive and negative posttest probability ratios, respectively. Both quantities can be expressed according to the time-dependent likelihood ratios (demonstration provided in the Appendix at www.jclinepi.com):

$$\mathrm{PT}^+(t) = \frac{P[D(t)=1|\mathrm{HR}]}{P[D(t)=0|\mathrm{HR}]} = \frac{P[D(t)=1]}{P[D(t)=0]}\mathrm{LikR}^+(t) \quad (7)$$

$$\mathrm{PT}^-(t) = \frac{P[D(t)=1|LR]}{P[D(t)=0|LR]} = \frac{P[D(t)=1]}{P[D(t)=0]}\mathrm{LikR}^-(t) \quad (8)$$

Similarly to diagnostic context, the time-dependent positive and negative likelihood ratios appear as a multiplicative coefficient between pre- and posttest probability ratios. The corresponding interpretations are straightforward. A patient classified in the HR group has a $\mathrm{PT}^+(t)$ times greater risk of presenting the event before time t than after t. In contrast, a patient classified as LR has a $1/\mathrm{PT}^-(t)$ times greater risk of presenting the event after time t than before t. Unlike the likelihood ratio, these values must be interpreted with caution given their dependence on the population frailty.

2.6. Software

All the aforementioned methods can be performed using the online tools available at http://www.divat.fr/en/online-calculators/evalbiom.

3. Applications

To illustrate the practical utility of the proposed methods, we have selected two articles: one in kidney transplantation and one in breast cancer. Nevertheless, to confirm our approach, we have also applied it to several other articles in various fields of medicine (results not shown). In this study, we only chose articles in which (1) the markers were highly correlated with the time to event with small P-values, (2) the authors concluded on the high prognostic capacity of the marker, and (3) the NHR, NLR, SHR(t), and SLR(t) were available. This selection of articles does not achieve the standards of a meta-analysis; the purpose of this selection was simply to highlight the usefulness of the proposed approach.

3.1. Prognosis of an acute rejection episode after kidney transplantation

In the study by Hauser et al. [17], the outcome was the time between the kidney transplantation and the first acute rejection episode (ARE). The HR group was made up of 20 patients presenting an increase in urinary monokine induced by interferon-γ (MIG) above 436 pg/ml (positive MIG). The LR group comprised 49 patients presenting no increase in urinary MIG (negative MIG). The survival curves differed significantly between the two groups ($P < 0.0001$). More precisely, ARE survival probability at 40 days after transplantation was 0.29 in the HR group and 0.99 in the LR group. The overall cumulative probability of ARE beyond 40 days after transplantation was 0.21. Applying Equations 1 and 2, the sensitivity of a prognosis up to 40 days was 0.97 and the specificity was 0.89 (Table 1). For a shorter prognosis up to 10 days, the sensitivity appeared to be perfect but the specificity decreased to 0.76. Urinary MIG thus seemed to be a very sensitive and specific predictor of ARE after kidney transplantation. More precisely, a decision based on an increase in urinary MIG limited the detection of false negatives. At 40 days, less than 3% of patients were incorrectly classified as LR. Nevertheless, this was at the cost of a slightly lower specificity, that is, around 11% of HR patients were incorrectly classified for a prognosis at 40 days.

For a prognosis up to 40 days, the PPV and NPV were estimated at 0.71 and 0.99, respectively. This illustrates the high accuracy of a negative test result in practice: patients without an increase in MIG had a less than 1% risk of acute rejection within the first 40 days after transplantation. The prognostic accuracy of a positive test result was also reasonable: patients with an increase in MIG had a 71% risk of acute rejection during the first 40 days after transplantation. As previously demonstrated, the use of urinary MIG as a biomarker in other populations, that is, with different cumulative event probabilities, may lead to unexpected PPV and NPV. Based on Equations 3 and 4, Fig. 2 illustrates this variation. In France, for example, the cumulative probability of an ARE at 40 days after renal

Table 1. Description of the results obtained from two different articles based on the data available in the corresponding articles and, for this study, estimations calculated using the proposed equations

Data extracted from the article						The estimations of the novel indicators							
Prognostic time *t*	**NHR**	**SHR**	**NLR**	**SLR**	***P*-value**	**Se(*t*)**	**Sp(*t*)**	**PPV(*t*)**	**NPV(*t*)**	**LiKR$^+$(*t*)**	**LiKR$^-$(*t*)**	**PT$^+$(*t*)**	**PT$^-$(*t*)**
Hauser et al. [17][a]													
10 days	20	0.76	49	1.00	<0.01	1.00	0.76	0.24	1.00	4.22	0.00	0.32	0.00
40 days		0.29		0.99		0.97	0.89	0.71	0.99	9.05	0.04	2.45	0.01
Mook et al. [18][b]													
5 yr	439	0.80	525	0.95	<0.01	0.77	0.59	0.20	0.95	1.86	0.39	0.25	0.05
10 yr		0.72		0.87		0.64	0.59	0.28	0.87	1.57	0.60	0.39	0.15

Abbreviations: NHR, number of HR patients; SHR, survival at time *t* in the HR group; NLR, number of LR patients; SLR, survival at time *t* in the LR group; Se, sensitivity; Sp, specificity; PPV, positive predictive value; NPV, negative predictive value; LikR, likelihood ratio; PT, posttest probability ratio.

[a] End point: acute rejection after renal transplantation, biomarker: urinary monokine (MIG), HR group: MIG elevation above 436 pg/ml.

[b] End point: distant metastasis after surgery, biomarker: 70-gene MammaPrint signature, HR group: poor prognosis group using the 70-gene MammaPrint signature.

transplantation is around 9% (data from the DIVAT network, www.divat.fr). The PPV and NPV would be 0.47 and 0.99, respectively (Fig. 2). Thus, although the NPV is always expected to be almost perfect, the PPV decreases. As a result, 53% of patients who tested positive for MIG would not have an ARE.

At 40 days, the time-dependent positive and negative likelihood ratios were 9.05 and 0.04, respectively. Time-dependent likelihood ratios were thus not too different from the expected values of 10 and 0.1, respectively. The urinary MIG biomarker may therefore be considered as useful to identify patients at high and low risk of ARE. Moreover, the positive and negative posttest probability ratios at 40 days were 2.45 and 0.01, respectively. Patients who tested positive for MIG had a 2.45-fold greater risk of experiencing an ARE before the 40-day posttransplant time point than after. In contrast, patients who tested negative for MIG had a 100-fold greater risk of experiencing an ARE after the 40-day time point than before. These results also illustrate the better capacity of the MIG-based test to identify patients free of ARE during the first 40 days, but it should not be considered for populations with a different cumulative incidence of ARE.

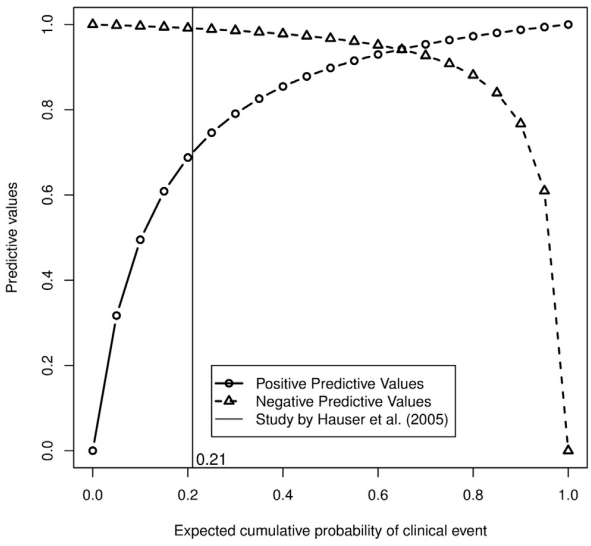


Fig. 2. Positive and negative predictive values regarding the expected cumulative probability of acute rejection up to 40 days after transplantation in kidney transplant recipients. In the study by Hauser et al. [17], the probability of acute rejection before 40 days was 21%; the positive and negative predictive values were 71% and 99%, respectively.

3.2. *Prognosis of a distant metastasis after breast cancer*

In the article by Mook et al. [18], a 70-gene MammaPrint signature (Agendia's Laboratories, Amsterdam, the Netherlands) was studied to predict distant metastasis after surgery in patients with breast cancer. The HR group was made up of 439 patients with a positive 70-gene MammaPrint signature. The LR group comprised 525 patients presenting a negative 70-gene MammaPrint signature. The survival probabilities were significantly different between the HR and LR groups (hazard ratio, 2.70; $P < 0.001$). The distant metastasis survival probability up to 5 years was 0.80 in the HR group and 0.95 in the LR group. The cumulative probability of distant metastasis was 12% at 5 years. The 70-gene MammaPrint signature thus seemed to be a relatively sensitive but not very specific predictor of distant metastasis after surgery. Indeed, the sensitivity at 0.77 may be reasonable to control the number of false negatives for a prognosis up to 5 years, but the cost is a low specificity of 0.59 (Table 1). The sensitivity was even lower for a longer prognosis up to 10 years.

For a prognosis up to 5 years, the PPV and NPV were estimated at 0.20 and 0.95, respectively. Based on the 70-gene MammaPrint signature, patients considered in the LR group had only a 5% risk of distant metastasis up to 5 years. In contrast, the accuracy of a positive test was small, with 80% of HR-classified patients without distant metastasis before 5 years. Mook et al. selected a population with a tumor size inferior to 2 cm, for which the risk of a distant metastasis at 5 years was low. In France, the overall cumulative probability

of distant metastasis at 5 years was lower (around 4.5%) in the same population (data from the BERENIS cohort, Nantes Institut de Cancérologie de l'Ouest). In this type of population, the PPV and NPV would be 8% and 98%, respectively (Fig. 3). The NPV is thus very high, but the PPV is very low. As a result, 92% of patients with a positive MammaPrint signature would not have a distant metastasis.

The positive and negative likelihood ratios, estimated at 1.86 and 0.39, respectively, were very different from the expected values. Moreover, the positive and negative posttest probability ratios at 5 years were 0.25 and 0.05, respectively. Patients with a positive 70-gene MammaPrint signature had a 0.25-fold greater risk of declaring a distant metastasis before 5 years after treatment than later. Patients with a negative 70-gene MammaPrint signature had a 20-fold greater risk of declaring a metastasis after 5 years than before. In this population, the 70-gene MammaPrint signature may help to accurately identify patients without distant metastasis at 5 years after surgery, but it would fail to identify patients with a future distant metastasis.

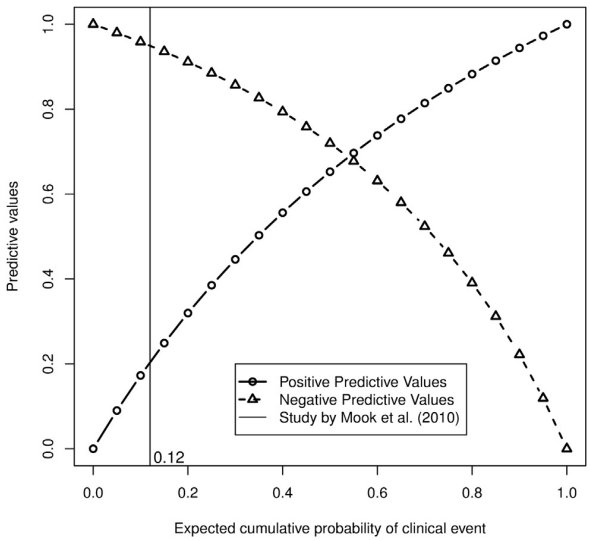


Fig. 3. Positive and negative predictive values regarding the expected cumulative probability of distant metastasis up to 5 years after surgery in women with breast cancer. In the study by Mook et al. [18], the probability of distant metastasis before 5 years was 12%, the positive and negative predictive values were 20% and 95%, respectively.

4. Discussion

Many articles evaluating the prognostic capacity of markers use the same methodology: Kaplan−Meier survival curves, Log-rank tests, and/or the Cox model, although these methods are not adequate and should not be used in routine practice because they can lead to misinterpretation of results. Here, we have proposed simple equations to calculate time-dependent sensitivity, specificity, predictive values, likelihood ratios, and posttest probability ratios from already published survival curves to confirm the prognostic capacity of a marker. We have illustrated this methodology using two different examples. This work may put published results and recommendations into perspective by providing useful additional information. Nevertheless, it is always preferable for authors to systematically use the suitable methodology in the first place, to prove the prognostic capacity of a marker, in which case our approach would not be necessary. Time-dependent receiver operating characteristic curve [9,19] can be directly estimated from individual data to estimate the prognostic capacity of a continuous marker. From the original definition of "C statistic" proposed by Harrell et al. [20], different concordance indices have been developed and presented as measures of discrimination between patients with longer event-free survival and those with shorter event-free survival [21−24]. The net reclassification improvement has been recently developed and may be used to assess the performance of new prognostic markers [25]. Our approach is useful only a posteriori when the authors do not use appropriate methodologies.

Time-dependent sensitivity or specificity can strengthen the conclusions of articles reporting on the prognostic capacity of a marker. Markers that are significantly associated with the risk of the event may not necessarily be able to discriminate LR and HR subjects: the sensitivity may be acceptable at the cost of low specificity or vice versa. Time-dependent likelihood ratios provide complementary information. One advantage of sensitivity, specificity, and likelihood ratios is their independence from the probability of the event. These indicators are therefore robust when applied to data sets from different countries and can even be used if the sample is not representative of the target population.

Moreover, time-dependent predictive values are relevant to clinical interpretations as they provide a direct probability of the event occurring or not before the prognostic time point in a specific group. Time-dependent posttest probability ratios complete the prognostic marker evaluation by quantifying how many times higher the risk that the event occurs before the prognostic time point is than that of the event occurring after the prognostic time point. In contrast to the intrinsic indicators of the marker's prognostic capacity, predictive values and posttest probability ratios are dependent on the disease in probability, as illustrated in the Applications section. Major therapeutic decisions can be wrongly made if the cumulative probability of the event differs to that on which the marker was developed. We illustrated this approach using two applications.

Hauser et al. studied the urine biomarker MIG and the prediction of an ARE early after kidney transplantation [17]. A negative MIG test indicated, with a high degree of confidence, that a patient would not suffer an ARE regardless of the cumulative probability of an ARE. Moreover, the negative posttest probability was very high: patients with a negative MIG test had a 100 times higher risk of suffering an ARE after the 40-day time point than before. In contrast, although the risk of declaring an ARE before the 40-day

time point was higher than that after the 40-day time point for patients with a positive MIG test, the positive predictive value was indicative of an unacceptably high false positive rate. If this test was to be used in a population with a more representative cumulative probability of ARE of around 9%, only 50% of MIG-positive patients would present an ARE within the 40-day period. If the strategy is to increase the immunosuppressive therapy for MIG-positive patients, this could lead to excessive treatment in a high percentage of patients. In a population with a 30% risk of ARE within the first 40 days, the expected PPV would be nearly 80% (Fig. 2). Even with this relatively small error rate (20%), this biomarker would not be sufficiently accurate to blindly treat patients without any histologic proof of the ARE. Thus, although this marker could be of interest to routinely monitor patients after kidney transplantation, to avoid unnecessary kidney biopsies in patients with a negative MIG test, the marker would be of limited use to increase immunosuppression in at-risk patients.

In the second article, Mook et al. [18] studied a population of women with a low risk of distant metastasis, that is, a breast cancer tumor inferior to 2 cm in size and mostly treated by adjuvant therapy. The 70-gene MammaPrint signature seemed to allow an acute selection of LR patients with a negative predictive value of 95% for a prognosis up to 5 years. Patients negative for the 70-gene MammaPrint signature had a 20 times greater risk of declaring a metastasis after 5 years than before. In contrast, if a supplementary chemotherapy is proposed for a positive signature, the PPV at 5 years indicates that 80% of HR patients may be overtreated. The positive posttest probability was 0.25, implying that 70-gene MammaPrint signature−positive patients had a 4 times greater risk of declaring a metastasis after 5 years than before. In normal circumstances, a marker with this type of positive posttest probability would not be considered for prognosis. Nevertheless, in the context of breast cancer, the overtreatment problem remains questionable as the main risk concerns metastatic relapses, which invariably lead to death. Actually, a woman's choice for chemotherapy is based on the hypothesis of a survival gain of several months, with a known risk of overtreatment, even those for whom medical oncologist would not propose chemotherapy [26,27]. Women generally prefer to receive unnecessary chemotherapy than not to be considered as HR, although they should be. In view of the event in this context being terminal in nature, the optimal strategy tends to favor a high NPV, that is, a small rate of women considered as LR, although they should not be considered as such. Fig. 3 shows that up to a cumulative probability of distant metastases of 22%, less than 10% of women are wrongly consider as LR. This indicates that the MammaPrint signature may not be used in populations with a higher risk of distant metastasis as the number of women incorrectly considered as LR may be considered too high (greater than 10%).

The confidence intervals of the different time-dependent indicators would have been important to complete the description of the marker's prognostic capacity. However, only the variability of time-dependent predictive values may be easily assessed from the published confidence intervals of survival probabilities. The computation of the confidence intervals of the other indicators requires information about the marker and survival variability, which are not available from published survival curves.

As with many observational cohorts, survival curves may be misleading because of confounding factors of the prognostic marker. To fill in this major drawback, several authors recommend the drawing of adjusted survival curves, obtained for instance by an inverse probability weight (IPW) approach [28−30]. Even if this type of graphic representation should be used, most of survival curves presented in clinical articles are estimated from the Kaplan−Meier unadjusted estimator. An interesting feature of our approach is that it is still available for articles presenting adjusted survival curves using the survival estimations in SHR(t) and SLR(t). Respecting the conclusions by Cole and Hernán [28], the corresponding adjusted time-dependent sensitivity will represent the sensitivity as if the patients with a positive test (classified in the HR group) have the same characteristics as all the sample. In parallel, the adjusted time-dependent specificity will represent the specificity as if the patients with a negative test (classified in the LR group) have the same characteristics as all the sample. This ensures no confounding factor between both groups. In extension, the other proposed indicators as time-dependent predictive values and posttest probability ratios can also be interpreted independently of confounding factors. We thus encourage the use of our methodology for such a type-adjusted survival curve. Nevertheless, we did not find an article based on IPW approach and devoted to the study of the capacity of a prognostic marker. In practice, for taking into account confounding factors, the authors always used Cox multivariate regression and the linear predictor as a prognostic composite score [31,32]. Again, our proposed methodology can also be used based on survival curves stratified on such scoring system.

In conclusion, the methods that we propose here can help to further interpret published results from articles reporting on raw or adjusted patient survival. Time-dependent sensitivity, specificity, and likelihood ratios evaluate a marker's intrinsic prognostic capacity, whereas time-dependent predictive values and posttest probability ratios need to be interpreted according to the targeted population. Of course, any conclusion has to be additionally considered against the pathology being studied and the medical purpose of the prognostic marker.

Appendix

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2013.10.022

References

[1] Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 2001;69(3):89−95.

[2] Buyse M, Sargent DJ, Grothey A, Matheson A, de Gramont A. Biomarkers and surrogate end points—the challenge of statistical validation. Nat Rev Clin Oncol 2010;7(6):309−17.

[3] Rector TS, Taylor BC, Wilt TJ. Systematic review of prognostic tests, in methods guide for medical test reviews. Rockville, MD: Agency for Healthcare Research and Quality (US); 2012.

[4] Lachenbruch PA, Rosenberg AS, Bonvini E, Cavaillé-Coll MW, Colvin RB. Biomarkers and surrogate endpoints in renal transplantation: present status and considerations for clinical trial design. Am J Transplant 2004;4(4):451−7.

[5] Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol 2004;159(9):882−90.

[6] Ware JH. The limitations of risk factors as prognostic tools. N Engl J Med 2006;355(25):2615−7.

[7] Spruance SL, Reid JE, Grace M, Samore M. Hazard ratio in clinical trials. Antimicrob Agents Chemother 2004;48(8):2787−92.

[8] Foucher Y, Combescure C, Ashton-Chess J, Giral M. Prognostic markers: data misinterpretation often leads to overoptimistic conclusions. Am J Transplant 2012;12(4):1060−1.

[9] Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 2000; 56(2):337−44.

[10] Carobbio A, Antonioli E, Guglielmelli P, Vannucchi AM, Delaini F, Guerini V, et al. Leukocytosis and risk stratification assessment in essential thrombocythemia. J Clin Oncol 2008;26(16):2732−6.

[11] Kaplan B, Schold J, Meier-Kriesche HU. Poor predictive value of serum creatinine for renal allograft loss. Am J Transplant 2003; 3(12):1560−5.

[12] Wang TJ, Gona P, Larson MG, Tofler GH, Levy D, Newton-Cheh C, et al. Multiple biomarkers for the prediction of first major cardiovascular events and death. N Engl J Med 2006;355(25):2631−9.

[13] Riley RD, Ridley G, Williams K, Altman DG, Hayden J, de Vet HC. Prognosis research: toward evidence-based results and a Cochrane methods group. J Clin Epidemiol 2007;60(8):863−5. author reply 865−6.

[14] Simel DL, Easter J, Tomlinson G. Likelihood ratios, sensitivity, and specificity values can be back-calculated when the odds ratios are known. J Clin Epidemiol 2013;66(4):458−60.

[15] Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ 2004;329(7458):168−9.

[16] Jaeschke R, Guyatt G, Sackett D. Users' guides to the medical literature: III. How to use an article about a diagnostic test: B. What are the results and will they help me in caring for my patients? JAMA 1994;271:703−7.

[17] Hauser IA, Spiegler S, Kiss E, Gauer S, Sichler O, Scheuermann EH, et al. Prediction of acute renal allograft rejection by urinary monokine induced by IFN-gamma (MIG). J Am Soc Nephrol 2005; 16(6):1849−58.

[18] Mook S, Knauer M, Bueno-de-Mesquita JM, Retel VP, Wesseling J, Linn SC, et al. Metastatic potential of T1 breast cancer can be predicted by the 70-gene MammaPrint signature. Ann Surg Oncol 2010;17(5):1406−13.

[19] Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics 2005;61(1):92−105.

[20] Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15(4):361−87.

[21] Chambless LE, Diao G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. Stat Med 2006;25(20): 3474−86.

[22] Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat Med 2004;23(13):2109−23.

[23] Pencina MJ, D'Agostino RB Sr, Song L. Quantifying discrimination of Framingham risk functions with different survival C statistics. Stat Med 2012;31(15):1543−53.

[24] Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 2011;30(10):1105−17.

[25] Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;27(2): 157−72. discussion 207−212.

[26] Mandelblatt JS, Sheppard VB, Hurria A, Kimmick G, Isaacs C, Taylor KL, et al. Breast cancer adjuvant chemotherapy decisions in older women: the role of patient preference and interactions with physicians. J Clin Oncol 2010;28(19):3146−53.

[27] Moumjid N, Carrère MO, Charavel M, Brémond A. Clinical issues in shared decision-making applied to breast cancer. Health Expect 2003; 6(3):222−7.

[28] Cole SR, Hernan MA. Adjusted survival curves with inverse probability weights. Comput Methods Programs Biomed 2004;75(1):45−9.

[29] Westreich D, Cole SR, Tien PC, Chmiel JS, Kingsley L, Funk MJ, et al. Time scale and adjusted survival curves for marginal structural Cox models. Am J Epidemiol 2010;171(6):691−700.

[30] Xie J, Liu C. Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. Stat Med 2005;24:3089−110.

[31] Foucher Y, Daguin P, Akl A, Kessler M, Ladrière M, Legendre C, et al. A clinical scoring system highly predictive of long-term kidney graft survival. Kidney Int 2010;78(12):1288−94.

[32] Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986;5(5):421−33.

- Chapitre 5 -

Modèle conjoint pour l'évolution des mesures de sérum de créatinine et le risque d'échec de greffe rénale

Fournier MC, Foucher Y, Blanche P, Buron F, Giral M and Dantan E. A joint model for longitudinal and time-to-event data to better assess the specific role of donor and recipient factors on long-term kidney transplantation outcomes. *Eur J Epidemiol.* 2016 May ;31(5) :469-79. Epub 2016 Feb 1.

Préambule

Les modèles conjoints pour données longitudinales et données de survie sont développés depuis plus de 15 ans [Wulfsohn and Tsiatis, 1997; Henderson et al., 2000]. De nombreux développements leur ont été apportés dans la littérature biostatistique [Rizopoulos, 2012]. Dans un article publié dans l'*International Journal of Epidemiology*, Asar et al. [2015] rappelle que les modèles conjoints devraient être utilisés pour correctement considérer les données répétées et données de survie et leur possible dépendance. Mais ces modèles restent encore peu utilisés pour de véritables applications en recherche clinique. En transplantation rénale, alors que plusieurs auteurs ont utilisé ou développé des modèles conjoints [Garre et al., 2008; Rizopoulos and Ghosh, 2011; Moranne et al., 2013; Daher Abdi et al., 2013], il n'existait pas à notre connaissance d'étude décrivant précisément les relations entre les variables explicatives et les processus d'évolution d'un marqueur quantitatif et de survie.

J'ai choisi de mettre en avant notre travail publié dans l'*European Journal of Epidemiology* [Fournier et al., 2016], puisqu'il se situe dans la continuité de mes travaux de thèse. Alors qu'au cours de celle-ci, j'ai travaillé au développement statistique d'un modèle conjoint, nous appliquons ici un modèle à effets aléatoires partagés entre le processus d'évolution et le processus de survie dans le contexte de la transplantation rénale. Plus précisément, nous modélisons simultanément l'évolution du sérum de créatinine, mesurée au cours du suivi du patient transplanté, et le risque de survenue d'un échec de greffe, défini comme un retour en dialyse ou décès avec un

greffon fonctionnel. Le risque de survenue d'un échec de greffe à un instant t semble dépendre du niveau de sérum de créatinine ainsi que de l'intensité de son augmentation à cet instant. L'approche étiologique mise en œuvre nous permet d'identifier le rôle spécifique des déterminants de santé sur l'un et/ou l'autre des deux processus.

Eur J Epidemiol (2016) 31:469–479
DOI 10.1007/s10654-016-0121-2

CLINICAL EPIDEMIOLOGY

A joint model for longitudinal and time-to-event data to better assess the specific role of donor and recipient factors on long-term kidney transplantation outcomes

Marie-Cécile Fournier[1,2] · Yohann Foucher[1] · Paul Blanche[3] · Fanny Buron[4] · Magali Giral[2,5] · Etienne Dantan[1]

Received: 26 August 2015 / Accepted: 16 January 2016 / Published online: 1 February 2016
© Springer Science+Business Media Dordrecht 2016

Abstract In renal transplantation, serum creatinine (SCr) is the main biomarker routinely measured to assess patient's health, with chronic increases being strongly associated with long-term graft failure risk (death with a functioning graft or return to dialysis). Joint modeling may be useful to identify the specific role of risk factors on chronic evolution of kidney transplant recipients: some can be related to the SCr evolution, finally leading to graft failure, whereas others can be associated with graft failure without any modification of SCr. Sample data for 2749 patients transplanted between 2000 and 2013 with a functioning kidney at 1-year post-transplantation were obtained from the DIVAT cohort. A shared random effect joint model for longitudinal SCr values and time to graft failure was performed. We show that graft failure risk depended on both the current value and slope of the SCr. Deceased donor graft patient seemed to have a higher SCr increase, similar to patient with diabetes history, while no significant association of these two features with graft failure risk was found. Patient with a second graft was at higher risk of graft failure, independent of changes in SCr values. Anti-HLA immunization was associated with both processes simultaneously. Joint models for repeated and time-to-event data bring new opportunities to improve the epidemiological knowledge of chronic diseases. For instance in renal transplantation, several features should receive additional attention as we demonstrated their correlation with graft failure risk was independent of the SCr evolution.

Keywords Joint modeling · Time-to-event data · Repeated measurements · Serum creatinine · Graft failure · Kidney transplantation

Introduction

Recently, Asar et al. [1] pointed out that joint models for repeated and time-to-event data should be used to correctly consider longitudinal and survival processes and their relationship when they are strongly dependent [2, 3]. Joint models could allow identification of specific associations for each risk factor: (1) association with an event risk, (2) association with an event risk throughout a previous marker evolution modification, and (3) associations on both longitudinal and survival processes. With this type of suitable model, the precise description of specific associations could provide new insights in the knowledge of a disease pathway mechanism.

Indeed, from a methodological standpoint, longitudinal measurements and time-to-event data are typically analyzed separately, using a mixed model and survival model

Electronic supplementary material The online version of this article (doi:10.1007/s10654-016-0121-2) contains supplementary material, which is available to authorized users.

✉ Etienne Dantan
Etienne.Dantan@univ-nantes.fr

[1] EA4275 SPHERE - bioStatistics, Pharmacoepidemiology and Human sciEnces REsearch, Nantes University, 1 rue Gaston Veil, 44035 Nantes, France

[2] Labex Transplantex, Inserm U1064, Institut de Transplantation Urologie Néphrologie (ITUN), Nantes University Hospital, Nantes, France

[3] Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark

[4] Service de Néphrologie, Transplantation et Immunologie Clinique, Hôpital Edouard Herriot, Lyon, France

[5] Centre d'investigation clinique biothérapie, Nantes, France

respectively, without considering their possible relationship [2]. However, it is known that these two processes are often mutually dependent in a chronic disease context. Inferences from mixed models may be biased in cases of an informative censoring process [4–6]. Similarly, the time-dependent Cox model fails to correctly handle a time-dependent endogenous variable, which is a variable generated by the patient themselves (e.g. creatinine) in contrast to an exogenous variable which is not (e.g. air pollution level) [7–9]. It also often fails to correctly account for measurement error.

In many chronic diseases, the occurrence of major events and the assessment of corresponding risk factors guide physicians in implementing the most beneficial care for patients. In renal transplantation, serum creatinine (SCr) is a well-known longitudinal marker used to assess the health of kidney transplant recipients especially after the first year post-transplantation, i.e. in the chronic phase of the disease evolution [10]. In the chronic phase, graft failure, defined as return to dialysis or death with a functioning graft, is often preceded by a continuous deterioration in renal function and is associated with an irreversible increase in SCr levels. In contrast, during the first year post-transplantation, patients are submitted to a risk of early graft failure due to acute clinical events such as delayed graft function, acute rejection episode, and infections or complications. Therefore the disease evolution mechanism is very different between the acute and chronic phases [11, 12]. Few authors have studied the association of risk factors on renal function evolution [13, 14] while the risk factors associated with long-term graft failure have been well described [12, 15, 16]. Renal function has been shown to be one of the most important risk factors [17, 18]. However, the precise mechanism is not well known: the majority of risk factors leading to graft failure may be associated with chronic SCr changes, finally leading to graft failure, but one can also hypothesize that some features may be related to graft failure risk independently from their association on SCr evolution.

Whilst several authors have used joint models for longitudinal and time-to-event data in renal transplantation [19–22], none have precisely studied the specific association of each baseline explicative variable. The precise study of such associations would be of primary importance for physicians in order to improve their appraisal of kidney transplant recipients' health. Therefore, we present for the first time a shared random effect multivariable joint model to study the baseline characteristics that could be related to long-term kidney graft outcomes. By studying SCr evolution and graft failure from one year post-transplantation, it brings an epidemiological approach to understand the risk factors associated with the disease evolution in its chronic phase.

Materials and methods

Study population

Data were extracted from the French observational and prospective DIVAT cohort (www.divat.fr) of kidney transplant recipients from 6 University hospitals (French Research Ministry: RC12_0452, last agreement No 13 334, No CNIL for the cohort: 891735, No CNIL for the study: 914226). According to the following inclusion criteria, 2749 patients were studied: adult recipients who received a first or second renal transplant between January 2000 and August 2013 from a living or heart beating deceased donor, alive with a functioning graft at 1-year post-transplantation and maintained under Tacrolimus and Mycofenolic acid. All study participants gave informed consent.

Available data

Most classical risk factors susceptible to influence SCr evolution and/or graft failure risk were extracted from the database. Donor features included: age, gender, last SCr level, deceased (from cardiovascular cause vs. other)/living donation. Recipient characteristics were: age, gender, body mass index (BMI), history of comorbidities (diabetes, hypertension, dyslipidemia, neoplasia, cardiovascular), duration of dialysis before transplantation, preemptive graft, hemodialysis or peritoneal dialysis, pre-transplant anti-class I or class II Human Leucocyte Antigen (HLA) immunization and cause of initial renal disease (recurrent nephropathy or not). Transplantation parameters were: cold ischemia time and number of HLA-A-B-DR incompatibilities. The following variables were collected within the first year post-transplantation: occurrence of delayed graft function (defined as the need for dialysis after transplantation), occurrence of acute rejection episodes and SCr levels at 3 and 6 months post-transplantation.

Outcomes

The baseline was the 1-year post-transplantation anniversary. The best marker of renal function should be the measured GFR (mGFR) [23]. Unfortunately, this measurement is costly and is not performed in practice for routine patient follow-up. Different equations to estimate GFR (eGFR) have been proposed [24–27], and are principally based on SCr values adjusted on recipient age, gender and ethnicity. A major limitation lies in the fact that these eGFR formulae have been developed from general population data or from patients with chronic kidney disease and thus may not be applicable to kidney transplant patients [28–30]. Despite this, SCr or eGFR are equivalent

in terms of relative evolution, the second resulting from a transformation of the first parameter. We decided to study the SCr (μmol/L), which was the longitudinal marker routinely recorded yearly until patient death with a functioning graft or return to dialysis. The time-to-event (graft failure) was defined as the delay between 1-year post-transplantation and the first event between return to dialysis or death with a functioning graft.

Statistical analysis

We used a shared random effect model. It combines a mixed model for the longitudinal process and a parametric survival model for the time-to-event process, for which underlying hypotheses were checked in an independent manner. For the longitudinal assessment, we used a logarithmic transformation of SCr values in order to respect both assumptions related to residual's homoscedasticity and linear relationship over time. Two random effects were considered for the baseline value and the slope. For survival, hazard proportionality and log-linearity were assessed.

In a joint shared random effect framework, longitudinal and survival processes are linked through common random effects. Rizopoulos has previously described the possible parameterizations to model this dependence [2]. For instance, the survival process can be modeled as dependent on the current level of the marker, on the intensity of marker evolution (i.e. the slope), on both current level and slope, on cumulative effects or on lagged effects.

In the first step of model building, we defined a baseline risk function and the dependence between the two processes from a joint model without baseline explicative variables. We graphically retained a Weibull distribution for the baseline risk function. According to the Bayesian Information Criteria (BIC), the dependence between the two processes was characterized by the instantaneous hazard of graft failure depending on both the level and the slope of the longitudinal marker at the current time. As recommended by Rizopoulos to solve optimization difficulties, all quantitative variables were standardized in order to scale the coefficients [2].

In the second step, we performed the selection of baseline explicative variables. Univariable models were composed using three effects of each variable: on baseline value, on the slope (interaction with time) and on the graft failure risk. Among these parameters, those which were not significant ($p > 0.05$) were removed in a hierarchical manner: if the association on the slope was significant, the corresponding association on baseline value was also considered. Finally, a multivariable joint model was generated by including effects retained in the univariable models, and a forward stepwise selection was performed (always using a 5 % type-I error rate).

In order to study the relevance of the joint modeling, we also performed separate analyses: (1) a linear mixed model to study the SCr evolution and (2) a time-dependent Cox model to study the graft failure risk. We used the same variables selection procedure.

As sensitivity analyses, we performed two joint models in a cause specific approach: (1) time-to-return to dialysis by censoring death, and in contrast (2) time-to-death with a functioning graft by censoring return to dialysis.

Joint model parameters were estimated by likelihood maximization. The complete mathematical formulation of the joint model is shown in "Appendix 1". Due to the logarithmic transformation of SCr, coefficients for the longitudinal process have an interpretation as relative change rather than absolute change. Details related to interpretations are presented in "Appendix 2". Confidence intervals for relative change were obtained using parametric simulations (5000 iterations) [31]. All analyses were performed using the 3.0.1 version of the R software [32] with the 1.3-0 version of the JM package [33].

Results

Baseline characteristics

Baseline characteristics are presented in Table 1. Sixty percent of the recipients were male, with a mean age of 49.7 ± 13.6 years. Histories of cardiovascular disease or dyslipidemia were observed in one third of recipients, 11.6 % had history of diabetes, 82.6 % had hypertension, and 8.3 % had a cancer before the transplantation. Second transplantations were realized in 17.2 % of studied patients. Immunologic characteristics included: 12.8 % of patients presented more than 4 HLA–A-B-DR incompatibilities, and around one third were immunized against class I or class II HLA prior to transplantation. Donors were mainly deceased (84.8 %) with a mean age of 50.7 ± 15.5 years and 56.4 % were male. Delayed graft function occurred for 714 patients (26.1 %). SCr at 3 and 6 months were on average 138.3 ± 53.4 and 136.6 ± 53.2 μmol/L respectively. Finally, 21.5 % of recipients presented at least one episode of acute rejection before the first anniversary of the graft.

Follow-up description

During follow-up, 278 patients returned to dialysis and 203 died with a functioning graft. The median event-free follow-up time was 3.99 years. The patient-graft survival curve and its corresponding 95 % confidence interval (95 % CI) are presented in Fig. 1. Patient-graft survival rates at 10 years after the first anniversary of the graft was

Table 1 Description of recipients, donors, and transplantation characteristics of the studied population (n = 2749)

	Missing	Estimations
Quantitative characteristics: mean ± SD		
Recipient age (years)	0	49.7 ± 13.6
Recipient BMI (kg/m^2)	10	24.0 ± 4.2
Donor age (years)	1	50.7 ± 15.5
Last donor SCr (μmol/L)	25	89.9 ± 52.8
Cold ischemia time (h)	10	17.8 ± 9.8
3-month SCr (μmol/L)	38	138.3 ± 53.4
6-month SCr (μmol/L)	75	136.6 ± 53.2
Categorical characteristics: N (%)		
Recipient men	0	1674 (60.9)
Transplanted before 2008	0	1369 (49.8)
Second transplantation	0	474 (17.2)
Dialysis technique	3	
Pre-emptive transplantation		342 (12.5)
Hemodialysis		2192 (79.8)
Peritoneal dialysis		212 (7.7)
Time to dialysis prior to transplantation (at least 1 year)	36	1948 (71.8)
Recurrent causal nephropathy	0	899 (29.1)
Diabetes history	0	319 (11.6)
Hypertension history	0	2272 (82.6)
Cardiovascular history	0	933 (33.9)
History of dyslipidemia	0	860 (31.3)
History of cancer	0	228 (8.3)
HLA A-B-DR incompatibilities (>4)	7	350 (12.8)
Positive anti-class I immunisation	66	876 (32.6)
Positive anti-class II immunisation	87	792 (29.8)
Donor men	8	1545 (56.4)
Status	6	
Living donor		418 (15.2)
Cerebrovascular donor death		1309 (47.7)
Non cerebrovascular donor death		1016 (37.1)
Delayed graft function	15	714 (26.1)
Acute rejection episode during the first year	0	591 (21.5)

SD standard deviation, *BMI* body mass index, *SCr* serum creatinine, *HLA* human leukocyte antigen

58.4 % [95 % CI 54.8 %; 62.3 %]. Additionally, 12 843 SCr measurements were collected, with a median of 4 measurements per patient (ranging from 1 to 14). The median time between two measurements was 11.7 months (interquartile range: 9.2, 12.5).

Joint modeling

Table 2 presents the estimations related to the final multivariable joint model.

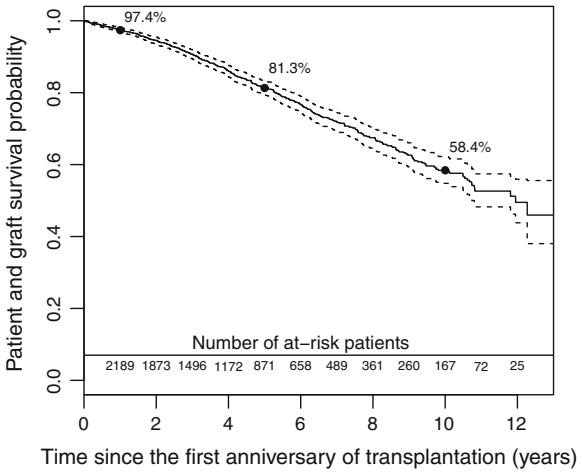


Fig. 1 Patient and graft survival according to the time since the first anniversary of the transplantation (n = 2749) from Kaplan–Meier estimator and their corresponding 95 % CI

Dependence between SCr dynamic and graft failure risk

For any time 1 year after transplantation (t > 1), the graft failure risk depended on both the current value and the current slope of the SCr. If a patient had a 25 % higher SCr, graft failure risk was twice as high (HR = 1.92, 95 % CI [1.75; 2.11]). Moreover, for a given SCr value, where a patient had a steeper increase in SCr graft failure risk was significantly worse (HR = 1.89, 95 % CI [1.17; 3.06] for an increase of 25 % in SCr value in 1 year).

Factors associated with 1-year post-transplantation SCr

Several factors appeared significantly correlated with a higher 1-year SCr without significant association with the SCr evolution or with graft failure risk. An increase of 50 μmol/L in the 6-month SCr level was associated with a 1-year SCr increase of 17.99 % (95 % CI [16.62 %; 19.34 %]). Patient with a graft provided from a donor 10 years older compared to other donor had a 3.68 % higher SCr at 1 year (95 % CI [2.97 %; 4.39 %]).

Factors associated with SCr evolution during follow-up

A history of diabetes was associated with a higher SCr increase. After 5 years, the presence of this comorbidity for a patient was associated with a SCr 14.45 % higher (95 % CI [7.76 %; 21.46 %]) compared to its absence, while we did not observe any significant difference at 1 year. For a given patient, a graft from a deceased donor due to a cerebrovascular cause was associated with a relative increase of 12.52 % in expected 5-year SCr (95 % CI [6.50 %; 18.89 %]), compared to a graft from a living

Table 2 Multivariable joint model for longitudinal evolution of logarithmic transformation of serum creatinine (SCr) and risk of graft failure (return to dialysis or death with a functioning graft) in kidney transplant patients (n = 2584 patients, 165 patients excluded due to missing data)

	Longitudinal process						Survival process		
	Association with the log(1-year SCr) (baseline effect)			Association with the log(SCr evolution) (slope effect)					
	Coef.	95 % CI	*p* value	Coef.	95 % CI	*p* value	HR	95 % CI	*p* value
Current SCr (μmol/L), for an increase of 25 %							1.92	[1.75; 2.11]	<0.0001
Current SCr increase (μmol/L), for a growth of 25 % in 1 year							1.89	[1.17; 3.06]	0.0097
Recipient age (years, standardized)	−0.028	[−0.038; −0.018]	<0.0001	−0.010	[−0.014; −0.006]	<0.0001	1.51	[1.35; 1.68]	<0.0001
Recipient gender: male versus female	0.074	[0.057; 0.091]	<0.0001	−0.007	[−0.014; 0.000]	0.0392			
Diabetes history: yes versus no	0.000	[−0.026; 0.025]	0.9866	0.027	[0.016; 0.039]	<0.0001			
Cardiovascular history: yes versus no	0.000	[−0.017; 0.017]	0.9812	0.008	[0.000; 0.015]	0.0371	1.39	[1.14; 1.69]	0.0011
3-month SCr (μmol/L, standardized)	0.083	[0.071; 0.096]	<0.0001				0.84	[0.74; 0.95]	0.0062
6-month SCr (μmol/L, standardized)	0.176	[0.164; 0.189]	<0.0001						
Acute rejection episode during the first year: yes versus no	0.055	[0.036; 0.074]	<0.0001				1.46	[1.17; 1.83]	0.0010
Anti-class I immunization: positive versus negative	0.010	[−0.008; 0.027]	0.2707	0.011	[0.004; 0.019]	0.0036	1.50	[1.19; 1.90]	0.0006
Rank of graft: second versus first							1.32	[1.02; 1.73]	0.0381
Donor type (ref: living donor)			0.0773			0.0022			
Cerebrovascular death	0.028	[0.004; 0.052]		0.018	[0.007; 0.028]				
Non cerebrovascular death	0.019	[−0.005; 0.043]		0.010	[−0.001; 0.020]				
Donor gender: male versus female							0.83	[0.69; 1.01]	0.0589
Donor age (years, standardized)	0.056	[0.045; 0.066]	<0.0001						

Referential value for 1-year SCr was 4.024, 95 % CI [3.982; 4.065]. Referential value for SCr evolution was 0.034 95 % CI [0.018; 0.050]. This model is adjusted on a time effect with a threshold at 2008 (before 2008 vs. after): coefficient for the relation to the SCr at 1-year: 0.018 95 % CI [0.002; 0.034] and to the SCr evolution: 0.013 95 % CI [0.005; 0.020] and HR = 0.73 [0.57; 0.94]. Parameters of the Weibull baseline risk function were: intercept −20.247 ± 0.982; log(shape): 0.337 ± 0.046. $\alpha = 2.93$; $\alpha_2 = 3.29$

Coef coefficient, *HR* hazard ratio, *CI* confidence interval

donor, but only a 7.14 % increase when compared to a deceased donor due to another cause (95 % CI [1.48 %; 13.00 %]).

Factors associated with both 1-year post-transplantation SCr and SCr evolution during the follow-up

Male recipient had a 7.68 % higher 1-year SCr (95 % CI [5.83 %; 9.51 %]). Nevertheless, he was likely to have a slower increase: after 5 years follow-up the gap reduced to 3.98 % (95 % CI [0.18 %; 7.81 %]).

Factors associated with the graft failure risk

Two factors were significantly associated with graft failure risk. Without significant correlation with the SCr, a patient transplanted for the second time had a higher graft failure risk compared to his first transplantation (HR = 1.32, 95 % CI [1.02; 1.73]). On the contrary, for a given patient, graft from male tended to be associated with a smaller risk of graft failure compared to graft from female donor (HR = 0.83, 95 % CI [0.69; 1.01]).

Factors associated with both SCr and the graft failure processes

A 10-year older patient was associated with a 2.04 % lower 1-year SCr (95 % CI [1.31 %; 2.77 %]). Moreover, this difference increased during the follow-up, i.e. 5.57 % lower at 5-years post-transplantation (95 % CI [4.20 %; 6.95 %]). This could be explained by lower creatinine production in the elderly population due to a smaller muscle mass. However, for a given SCr evolution during the follow-up, the situation where the recipient is older was associated with a higher graft failure risk (HR = 1.35 for a 10-year older patient, 95 % CI [1.25; 1.46]).

For a given patient, higher 3-month SCr was associated with a significantly higher 1-year SCr level. For instance, for a 50 µmol/L difference at 3-months, the 1-year SCr level increased by 8.08 % (95 % CI [6.83; 9.32]). Additionally, for a given evolution of SCr from 1-year post-transplantation, a higher 3-month SCr was associated with a lower graft failure risk (HR = 0.85 for an increase of 50 µmol/L in 3-month SCr level, 95 % CI [0.75; 0.95]).

Major risk factors included acute rejection in the first year post-transplantation, immunization, and cardiovascular history. When a patient had an acute rejection episode during the first year post-transplantation, a 5.65 % higher 1-year SCr was observed compared to cases where no acute rejection occurred (95 % CI [3.65; 7.71]). Nevertheless, independently of the current value and the slope of SCr, the situation where acute rejection has occurred appeared with a higher risk of graft failure (HR = 1.46, 95 % CI [1.17; 1.83]). A patient with cardiovascular history was more likely to have an increased SCr compared to the same patient without this history, and a higher graft failure risk independently from this increase (HR = 1.39, 95 % CI [1.14; 1.69]). Similarly, a significant SCr increase was demonstrated for pre-transplant immunized patient, with an additional graft failure risk not related to this SCr increase, compared to the same patient non-immunized.

Separate models

The linear mixed model estimations and those of the time-dependent Cox model are presented in Tables 3 and 4 respectively. One can note differences in the retained variables. An acute rejection episode in the first year post-transplantation was significantly associated with the SCr evolution by using a linear mixed model, in opposition to the results obtained by using joint models. No relationship between cardiovascular history or donor type with SCr evolution were retained by using the linear mixed model, while we concluded from the final joint model that patients with cardiovascular history may have a significantly higher SCr increase during the follow-up ($p = 0.0371$) and SCr evolution could be different given the donor type status ($p = 0.0022$). Slight underestimations of hazard ratios were obtained from the time-dependent Cox model compared to the joint model. For example, the hazard ratio related to the recipient age was 1.25 (95 % CI [1.19–1.38]) for 10 years older by using the time-dependent Cox model against 1.35 (95 % CI [1.25; 1.46]) by using the joint model. Additionally, diabetes was retained as a risk factor for graft failure by using the Cox model, while it does not by using the joint model. In contrast, acute rejection episode was not retained as a risk factor for graft failure by using the time-dependent Cox model, while it was by using the joint model.

Cause specific approach

Using a cause specific approach (Tables S1 and S2 in supplementary materials), we observed that current SCr level was more importantly associated with return to dialysis (HR = 2.51, 95 % CI [2.22; 2.84]) compared to death with a functioning graft (HR = 1.47, 95 % CI [1.24; 1.74]). As expected, this higher association was also observed for acute rejection episode (HR = 1.63, 95 % CI [1.20; 2.20] vs. HR = 1.24, 95 % CI [0.86; 1.80]). In contrast, cardiovascular history (HR = 1.07, 95 % CI [0.81; 1.41] vs. HR = 2.01, 95 % CI [1.49; 2.70]) and recipient age (HR = 1.20, 95 % CI [1.05; 1.39] vs. HR = 2.36, 95 % CI [1.95; 2.86]) were strongly associated with the time-to-death.

Table 3 Multivariable analysis for the logarithmic transformation of SCr

	Longitudinal process					
	Association with the log(1-year SCr) (baseline effect)			Association with the log(SCr evolution) (slope effect)		
	Coef.	95 % CI	*p* value	Coef.	95 % CI	*p* value
Recipient age (years, standardized)	−0.025	[−0.035; −0.015]	<0.0001	−0.010	[−0.013; −0.006]	<0.0001
Recipient gender: male versus female	0.073	[0.056; 0.090]	<0.0001	−0.007	[−0.013; 0.001]	0.0441
History of diabetes: yes versus no	0.003	[−0.022; 0.028]	0.7987	0.026	[0.015; 0.038]	<0.0001
3 month SCr (μmol/L, standardized)	0.085	[0.072; 0.097]	<0.0001			
6 month SCr (μmol/L, standardized)	0.176	[0.164; 0.189]	<0.0001			
Acute rejection episode <1 year: yes versus no	0.055	[0.036; 0.075]	<0.0001	−0.009	[−0.017; −0.001]	0.0233
Anti class I immunization: positive versus negative	0.010	[−0.007; 0.028]	0.2344	0.010	[0.002; 0.017]	0.0111
Donor age (years, standardized)	0.057	[0.046; 0.067]	<0.0001			
Cold ischemia time (h)—standardized	0.002	[−0.006; 0.010]	0.6717	0.004	[0.001; 0.007]	0.0253

12,241 observations from 2583 patients

Referential value for 1-year SCr was 4.024, 95 % CI [3.985, 4.064]. Referential value for SCr evolution was 0.040 95 % CI [0.025; 0.055]. This model is adjusted on a time effect with a threshold at 2008 (before 2008 vs. after): coefficient for relation to the 1-year SCr: 0.020 95 % CI [0.004; 0.037] and to the SCr evolution: 0.011 95 % CI [0.003; 0.018]

Coef coefficient, *CI* confidence interval, *SCr* serum creatinine

Table 4 Multivariable time-dependent Cox model for patient-graft failure risk (2604 patients, 455 events observed)

	Survival process		
	HR	95 % CI	*p* value
Last observation of SCr carried forward (LOCF), (μmol/L), for an increase of 25 %	2.07	[1.97; 2.18]	<0.0001
Recipient age (years, standardized)	1.40	[1.26; 1.55]	<0.0001
Diabetes history: yes versus no	1.50	[1.15; 1.95]	0.0031
Cardiovascular history: yes versus no	1.36	[1.12; 1.65]	0.0021
3-month SCr (μmol/L, standardized)	0.80	[0.73; 0.89]	<0.0001
Anti class I immunization: positive versus negative	1.44	[1.15; 1.81]	0.0015
Graft rank: second versus first	1.37	[1.05; 1.80]	0.0221
Donor gender: male versus female	0.82	[0.68; 1.00]	0.0456
Time to dialysis prior transplantation (≥1 year vs. <1)	1.27	[1.00; 1.60]	0.0500

HR hazard ratio, *CI* confidence interval, *SCr* serum creatinine

A period effect is included with a HR = 0.71, 95 % CI [0.55; 0.92]

Discussion

Our results show that during the chronic phase of renal transplantation, elevated SCr levels as well as the magnitude of SCr increases are associated with a higher risk of graft failure. Accordingly, physicians routinely supervise both the current SCr level and its increase. The large majority of baseline explicative variables are firstly associated with the baseline SCr level or its evolution, finally leading to graft failure. Interestingly, we demonstrated that besides the association of cardiovascular history with increased SCr, this risk factor was additionally associated with an increase in the risk of graft failure. Therefore, at a given time for a given SCr level and slope, the presence of cardiovascular history should be considered as a risk factor for graft failure. Similarly, patient transplanted for a second time seemed at higher risk of graft failure, regardless of the SCr level or its slope, compared to its first graft. Other factors independent of SCr leading to increased graft failure risk may result from stronger immunosuppression or undetected immunization against donor specific antigens. In addition to retransplantation or the presence of cardiovascular history, particular attention should also be paid to patients with a high 3-month SCr level, transplantation in older patients or when an acute rejection episode during the first year has occurred. These patients may be more

susceptible to graft failure without having previously displayed aberrant SCr levels. On the contrary, if a patient received a deceased donor graft but had a SCr evolution analogous to those which would be observed if the graft had come from a living donor, the monitoring of this patient should be the same regardless of the donor status.

In renal transplantation, numerous studies have focused on only one or two measurements of renal function to study their association with graft failure [34–36]. However, the joint modeling approach allows the whole trajectory of longitudinal SCr measurements to be taken into account. In this paper, we used for the first time a shared random effect joint modeling to more precisely specify the association between chronic SCr evolution and graft failure risk. Different types of dependence can be considered such as the current marker level, the evolution intensity during the follow-up, cumulative effects or lagged effects [2].

More generally, our approach illustrates that joint modeling constitutes a powerful approach for time-to-event analysis with endogenous time-dependent variable [1], which supports a real mechanistic evolution for many chronic diseases. However, their use in observational studies is still uncommon. As previously acknowledged by Asar et al. [1], differing results and interpretations between the joint modeling and the separate approaches reinforce the necessity to use joint modeling in the presence of endogenous variable. We also highlighted the differences in our application in kidney transplantation. In other diseases, the informative censoring or the endogenous nature for the longitudinal variable can result in even higher differences.

One limitation in our study may be the graft failure definition: the first event between the return to dialysis and death with a functioning graft. Because it is very difficult to distinguish the cause of death related or not to the disease, we performed a sensitivity analyses. The results illustrated the overall robustness of the results but with a probable underestimation of the association between the SCr and the acute rejection episode on the risk of graft failure.

In conclusion, our results illustrate the importance of joint models and their potential usefulness in improving chronic disease research. It brings a more complete epidemiological view of the risk factors and the related natural disease history mechanisms. The use of this novel statistical model on a large cohort of kidney transplant recipients highlights that several risk factors were associated with SCr evolution while others were associated with graft failure risk independently of the initial SCr value or its subsequent evolution. These included elderly or immunized recipients, second transplantations, grafts coming from female donors, patients experiencing an acute rejection episode in the first year post-transplantation, patients with cardiovascular history or with a high gap between 3 and 12 month SCr measurements, features that should receive additional attention.

Acknowledgments We wish to thank the DIVAT scientific board (C. Legendre, H. Kreis, L. Rostaing, N. Kamar, E. Morelon, G. Mourad, V. Garrigue, M. Kessler and M. Ladrière) as well as members of the clinical research assistant team (S. Le Floch, A. Petit, J. Posson, C. Scellier, V. Eschbach, K. Zurbonsen, C. Dagot, F. M'Raiagh, V. Godel, X. Longy and P. Przednowed). The DIVAT cohort is partially supported by Roche Laboratory since 1994.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Appendix 1: Mathematical formulation of the shared random effect joint model

Let Y be the longitudinal marker and t_{ij} the time of measurement of the jth ($j = 1, \ldots, n_i$) measure for the patient i ($i = 1, \ldots, N$). Let $h(\cdot)$ denotes the instantaneous risk function of graft failure. The joint model combines a linear mixed model (Eq. 1) with a parametric regression model (Eq. 2). They share the random effects (b_{0i}; b_{1i}).

$$Y_{ij}(t_{ij}) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \beta_2 X_{1i} + \beta_3 X_{2i} t_{ij} + \varepsilon_{ij} \quad (1)$$

$$h_i(t) = h_0(t)\, \exp(\gamma^T X_{3i} + g(m_i(t); \alpha)) \quad (2)$$

with $(b_{0i}; b_{1i})^T \sim \mathrm{MVN}(0,B)$, B an unstructured variance–covariance matrix, X_{1i} a vector of baseline covariates influencing the baseline value of longitudinal marker, X_{2i} another vector of baseline covariates that may change marker evolution over time and β_0, β_1 two scalars defining the referential value of the baseline level and the slope of the longitudinal biomarker $Y(\cdot)$ respectively, and β_2, β_3 two p-vectors of the same dimension as X_1 and X_2 respectively. The evolution of the measurements $Y_{ij}(t_{ij})$ are defined by the sum of a subject specific trend $m_i(t_{ij})$ plus an error term $\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$. For the instantaneous risk function of graft failure, $h_0(t)$ denotes the baseline risk function, and X_{3i} is a vector of baseline covariates that could influence the graft failure risk, with a corresponding vector of fixed regression coefficients γ. g is a function of the true level of the marker m_i, which specifies the type of dependence between the longitudinal and the survival processes. Classically, it may be the current level of the marker ($g(m_i(t)) = \alpha m_i(t)$), the intensity of marker deterioration during the follow-up i.e. the slope ($g(m_i(t)) = \alpha_2 m_i'(t)$), or both ($g(m_i(t)) = \alpha_1 m_i(t) + \alpha_2 m_i'(t)$) [2]. This latter is the retained association of the model presented in Table 2.

Appendix 2: Clinical interpretations of the joint model parameters

Parameters of the longitudinal process

Due to the log transformation of the longitudinal marker SCr, the parameters in the linear mixed submodel should be interpreted as the log of relative change. The longitudinal equation can be written as follows:

$$\log\left(\mathrm{SCr}(t_{ij})\right) = \beta_0 + b_{0i} + (\beta_1 + b_{1i})t_{ij} + \beta_2 X_{1i} + \beta_3 X_{2i} t_{ij}$$

and the SCr evolution can be re-written as:

$$\mathrm{SCr}(t_{ij}) = \exp(\beta_0 + b_{0i})\ \exp\left((\beta_1 + b_{1i})t_{ij}\right) \exp(\beta_2 X_{1i})\exp\left(\beta_3 X_{2i} t_{ij}\right)$$

Qualitative variables

Let Z be a qualitative variable associated with:

- the 1-year SCr only ($Z \subseteq X_1;\ Z \nsubseteq X_2$). The excess of SCr for a patient with $Z = 1$ as compared to the case where $Z = 0$ for the same patient is:

$$\frac{\mathrm{SCr}(t_{ij})_{[Z_i=1 \text{ vs } Z_i=0]}}{\mathrm{SCr}(t_{ij})_{[Z_i=1 \text{ vs } Z_i=0]}} = \frac{\exp(\beta_0 + b_{0i})\exp\left((\beta_1 + b_{1i})t_{ij}\right)\exp(\beta_2)}{\exp(\beta_0 + b_{0i})\exp\left((\beta_1 + b_{1i})t_{ij}\right)}$$

This gap of SCr is constant beyond 1-year post-transplantation.

- Both the 1-year SCr and the SCr increase during the follow-up ($Z \subseteq X_1; Z \subseteq X_2$)

$$\mathrm{SCr}(t_{ij})_{[Z_i=1 \text{ vs } Z_i=0]} = \exp\left(\beta_2 + \beta_3 t_{ij}\right)$$

This gap of SCr value is increasing or decreasing during the follow-up according to the sign of β_3. For clinical purposes, in the interpretations, we used the time $t = 5$ to quantify a relative change at 5 years after the first year post-transplantation.

Quantitative variables

Let W_1 be a quantitative variable with sd_{W1} its standard deviation, w a value of W_1 and Δ a relevant clinical increase.

- Let X_1 be the standardized version of W_1, X_2 be null (W_1 was associated with the 1-year SCr only).

$$\begin{aligned}\mathrm{SCr}(t_{ij})\left[W_{1i} = w + \Delta \text{ vs } W'_{1i} = w\right] &= \frac{\exp(\beta_0 + b_{0i})\exp\left((\beta_1 + b_{1i})t_{ij}\right)\exp\left(\beta_2\left(\frac{(w+\Delta)}{sd_{W1}}\right)\right)}{\exp(\beta_0 + b_{0i})\exp\left((\beta_1 + b_{1i})t_{ij}\right)\exp(\beta_2(w/sd_{W1}))} \\ &= \frac{\exp(\beta_2 w/sd_{W1} + \beta_2\Delta/sd_{W1})}{\exp(\beta_2(w/sd_{W1}))} \\ &= \exp(\beta_2\Delta/sd_{W1})\end{aligned}$$

- Now, let $X_1 = X_2$ be the standardized version of W_1 (W_1 was associated with both the 1-year SCr and the SCr evolution).

$$\begin{aligned}\mathrm{SCr}(t_{ij})\left[W_{1i} = w + \Delta \text{ vs } W'_{1i} = w\right] &= \frac{\exp(\beta_0 + b_{0i})\exp\left((\beta_1 + b_{1i})t_{ij}\right)\exp\left(\beta_2\left(\frac{(w+\Delta)}{sd_{W1}}\right)\right)\exp\left(\beta_3 t_{ij}\left(\frac{(w+\Delta)}{sd_{W1}}\right)\right)}{\exp(\beta_0 + b_{0i})\exp\left((\beta_1 + b_{1i})t_{ij}\right)\exp(\beta_2(w/sd_{W1}))\exp\left(\beta_3 t_{ij}(w/sd_{W1})\right)} \\ &= \exp(\beta_2\Delta/sd_{W1})\exp\left(\beta_3 t_{ij}\Delta/sd_{W1}\right))\end{aligned}$$

Hazard ratio for the longitudinal marker

As we have seen in "Appendix 1", the instantaneous risk function is written as follows:

$$\begin{aligned}h_i(t) &= h_0(t)\ \exp\left(\gamma^T X_{3i} + \alpha_1 m_i(t) + \alpha_2 m'_i(t)\right) \\ &= h_0(t)\ \exp\left(\gamma^T X_{3i} + \alpha_1 m_i(t)\ + \alpha_2 \frac{\delta m_i(t)}{\delta t}\right)\end{aligned}$$

With $m_i(t) = \beta_{0i} + \beta_{1i}t$ and $\frac{\delta m_i(t)}{\delta t} = \beta_{1i}$

As we use a log transformation of SCr measurement ($Y(t) = \log(\mathrm{SCr}(t))$), the hazard ratio which quantifies the association between the longitudinal marker and the risk of event was expressed for a clinically relevant difference.

- For the current level of the marker, we can rewrite the HR for a difference of 25 % in SCr values at the same time for the same patient and the same slope:

$$\begin{aligned}\mathrm{HR}_{1.25\mathrm{SCr}(t)\mathrm{vsSCr}(t)} &= \frac{h_0(t)\exp\left(\gamma^T X_{3i} + \alpha_1 \log(1.25\mathrm{SCr}(t)) + \alpha_2 m'_i(t)\right)}{h_0(t)\exp\left(\gamma^T X_{3i} + \alpha_1 \log(\mathrm{SCr}(t)) + \alpha_2 m'_i(t)\right)} \\ &= \exp(\alpha_1(\log(1.25\mathrm{SCr}(t)) - \log(\mathrm{SCr}(t)))) \\ &= 1.25^{\alpha_1}\end{aligned}$$

- For the intensity of the marker, the HR which compares the situation in which $\frac{\delta}{\delta t}\log(SCr_i(t)) = s_1$ to another in which $\frac{\delta}{\delta t}\log(SCr_i(t)) = s_2$, for same covariates X_{3i} and level of SCr at time t is equal. $HR = \exp(\alpha_2(s_2 - s_1))$. Besides, because we assume a linear model, $\frac{\delta}{\delta t}\log(SCr_i(t))$ is constant, that is $\forall t$, $\frac{\delta}{\delta t}\log(SCr_i(t)) =$ s for some $s \in \mathbb{R}$. This implies $\forall t' > t$:

  $SCr_i(t') = SCr_i(t)\exp(s(t' - t))$.

  If the SCr increases by x % between t-1 and t, then $s = \log(1 + x/100)$ because $s = \log\left(\frac{SCr_i(t)}{SCr_i(t-1)}\right)$

  This leads to: $HR = \exp(\alpha_2(\log(1 + x/100) - \log(1 + y/100)))$ which is the HR which compares an increase of x % between t-1 and t to an increase of y %. In our paper, we choose to compare an increase of 25 % compare to the mean evolution (a growth of 3 % each year).

Hazard ratio for the quantitative variables

Because the quantitative variables have been standardized, the HR for these factors were expressed for an increase of one standard deviation. In order to calculate them for an increase of relevant threshold in the variable unit, we can proceed as follows:

Let X_1 be the standardization of W_1 with sd_1 its standard deviation. HR_X is the HR obtained for the standardized variable and HR_W is the one for an increase of Δ unit of W_1.

$$HR_W = HR_X^{\left(\frac{\Delta}{sd_1}\right)}$$

References

1. Asar Ö, Ritchie J, Kalra PA, Diggle PJ. Joint modelling of repeated measurement and time-to-event data: an introductory tutorial. Int J Epidemiol. 2015;44(1):334–44.
2. Rizopoulos D. Joint models for longitudinal and time-to-event data: with applications in R. Boca Raton: CRC Press; 2012. p. 279.
3. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. Biometrics. 1997;53(1):330–9.
4. Leffondre K, Boucquemont J, Tripepi G, Stel VS, Heinze G, Dunkler D. Analysis of risk factors associated with renal function trajectory over time: a comparison of different statistical approaches. Nephrol Dial Transplant. 2015;30(8):1237–43.
5. Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. Test Madr Spain. 2009;18(1):1–43.
6. Tsiatis AA, Davidian M. joint modeling of longitudinal and time-to-event data: an overview. Stat Sin. 2004;14:809–34.
7. Rizopoulos D, Takkenberg JJM. Tools & techniques–statistics: dealing with time-varying covariates in survival analysis–joint models versus Cox models. EuroIntervention. 2014;10(2):285–8.
8. Andrinopoulou E-R, Rizopoulos D, Jin R, Bogers AJJC, Lesaffre E, Takkenberg JJM. An introduction to mixed models and joint modeling: analysis of valve function over time. Ann Thorac Surg. 2012;93(6):1765–72.
9. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. New York: Wiley; 2011. p. 464.
10. Levey AS, Coresh J, Balk E, Kausz AT, Levin A, Steffes MW, et al. National Kidney Foundation practice guidelines for chronic kidney disease: evaluation, classification, and stratification. Ann Intern Med. 2003;139(2):137–47.
11. Galichon P, Xu-Dubois Y-C, Finianos S, Hertig A, Rondeau E. Clinical and histological predictors of long-term kidney graft survival. Nephrol Dial Transplant. 2013;28(6):1362–70.
12. Foucher Y, Daguin P, Akl A, Kessler M, Ladriere M, Legendre C, et al. A clinical scoring system highly predictive of long-term kidney graft survival. Kidney Int. 2010;78(12):1288–94.
13. Ferro CJ, Hodson J, Moore J, McClure M, Tomson CRV, Nightingale P, et al. Bayesian analysis of glomerular filtration rate trajectories in kidney transplant recipients: a pilot study. Transplantation. 2015;99(3):533–9.
14. Marcén R, Morales JM, Fernández-Rodriguez A, Capdevila L, Pallardó L, Plaza JJ, et al. Long-term graft function changes in kidney transplant recipients. NDT Plus. 2010;3(Suppl_2):ii2–8.
15. Pascual M, Theruvath T, Kawai T, Tolkoff-Rubin N, Cosimi AB. Strategies to improve long-term outcomes after renal transplantation. N Engl J Med. 2002;346(8):580–90.
16. Debout A, Foucher Y, Trébern-Launay K, Legendre C, Kreis H, Mourad G, et al. Each additional hour of cold ischemia time significantly increases the risk of graft failure and mortality following renal transplantation. Kidney Int. 2015;87(2):343–9.
17. de Bruijne MHJ, Sijpkens YWJ, Paul LC, Westendorp RGJ, van Houwelingen HC, Zwinderman AH. Predicting kidney graft failure using time-dependent renal function covariates. J Clin Epidemiol. 2003;56(5):448–55.
18. Kasiske BL, Andany MA, Danielson B. A thirty percent chronic decline in inverse serum creatinine is an excellent predictor of late renal allograft failure. Am J Kidney Dis. 2002;39(4):762–8.
19. Daher Abdi Z, Essig M, Rizopoulos D, Le Meur Y, Prémaud A, Woillard JB, et al. Impact of longitudinal exposure to mycophenolic acid on acute rejection in renal-transplant recipients using a joint modeling approach. Pharmacol Res. 2013;72:52–60.
20. Moranne O, Maillard N, Fafin C, Thibaudin L, Alamartine E, Mariat C. Rate of renal graft function decline after one year is a strong predictor of all-cause mortality. Am J Transplant. 2013;13(3):695–706.
21. Rizopoulos D, Ghosh P. A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. Stat Med. 2011;30(12):1366–80.
22. Garre FG, Zwinderman AH, Geskus RB, Sijpkens YWJ. A joint latent class changepoint model to improve the prediction of time to graft failure. J R Stat Soc Ser A Stat Soc. 2008;171(1):299–308.
23. Stevens LA, Coresh J, Greene T, Levey AS. Assessing kidney function-measured and estimated glomerular filtration rate. N Engl J Med. 2006;354(23):2473–83.
24. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF III, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009;150(9):604–12.
25. Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, Roth D. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of diet in Renal Disease Study Group. Ann Intern Med. 1999;130(6):461–70.

26. Nankivell BJ, Gruenewald SM, Allen RD, Chapman JR. Predicting glomerular filtration rate after kidney transplantation. Transplantation. 1995;59(12):1683–9.
27. Cockcroft DW, Gault MH. Prediction of creatinine clearance from serum creatinine. Nephron. 1976;16(1):31–41.
28. Buron F, Hadj-Aissa A, Dubourg L, Morelon E, Steghens J-P, Ducher M, et al. Estimating glomerular filtration rate in kidney transplant recipients: performance over time of four creatinine-based formulas. Transplantation. 2011;92(9):1005–11.
29. White CA, Akbari A, Doucette S, Fergusson D, Knoll GA. Estimating glomerular filtration rate in kidney transplantation: Is the new chronic kidney disease epidemiology collaboration equation any better? Clin Chem. 2010;56(3):474–7.
30. Gaspari F, Ferrari S, Stucchi N, Centemeri E, Carrara F, Pellegrino M, et al. Performance of different prediction equations for estimating renal function in kidney transplantation. Am J Transplant. 2004;4(11):1826–35.
31. Aalen OO, Farewell VT, De Angelis D, Day NE, Gill ON. A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: application to AIDS prediction in England and Wales. Stat Med. 1997;16(19):2191–210.
32. R Development Core Team. R: a language and environment for statistical computing [Internet]. Computing RF for S, éditeur. Vienna, Austria; 2010. Disponible sur: http://www.R-project.org/.
33. Rizopoulos D. JM: an R package for the joint modelling of longitudinal and time-to-event data. J Stat Softw. 2010;35(9):1–33.
34. Kasiske BL, Israni AK, Snyder JJ, Skeans MA, Peng Y, Weinhandl ED. A simple tool to predict outcomes after kidney transplant. Am J Kidney Dis. 2010;56(5):947–60.
35. Lenihan CR, O'Kelly P, Mohan P, Little D, Walshe JJ, Kieran NE, et al. MDRD-estimated GFR at one year post-renal transplant is a predictor of long-term graft function. Ren Fail. 2008;30(4):345–52.
36. Hariharan S, McBride MA, Cherikh WS, Tolleris CB, Bresnahan BA, Johnson CP. Post-transplant renal function in the first year predicts long-term kidney transplant survival. Kidney Int. 2002;62(1):311–8.

- Chapitre 6 -

Evaluation des capacités prédictives d'un score pronostique dynamique

Fournier MC, Dantan E and Blanche P. An R2 -curve for evaluating the accuracy of dynamic predictions. *Stat Med.* 2018 Mar 30 ;37(7) :1125-1133. Epub 2017 Dec 4.

Préambule

A l'ère d'une médecine de précision, les outils pronostiques d'événements cliniques sont de plus en plus populaires. Les récentes avancées technologiques, en particulier dans le domaine des données génomiques, ont transformé la recherche biomédicale et offrent de nombreuses opportunités pour le développement d'une médecine personnalisée [Ioannidis, 2010]. Pour autant, une très faible proportion des marqueurs identifiés ou des outils de prédictions proposés ont été transférés dans la pratique clinique courante [Joyner and Paneth, 2015]. Cette désillusion s'explique en partie par certains écueils méthodologiques rencontrés, conduisant à un manque de crédibilité de ces outils [Ioannidis, 2006; Iglesias et al., 2014]. Alors que la littérature scientifique dans le domaine biomédical est foisonnante, celle-ci pourrait être de meilleure qualité [Chalmers and Glasziou, 2009; Chalmers et al., 2014]. En pronostic, les récentes recommandations TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) ont vu le jour afin d'améliorer la qualité du report des informations relatives à un outil de prédiction [Collins et al., 2015], l'évaluation des performances prédictives en est un élément important.

Un outil pronostique a pour objectif d'estimer la probabilité de survenue d'un événement dans une fenêtre de temps donné. Un outil pronostique statique est défini comme un outil pronostique utilisant exclusivement les informations disponibles au début de l'étude. Par exemple, le KTFS est un score calculable à un an post-transplantation à partir d'informations disponibles au moment de la transplantation et durant la première année de greffe afin de pronostiquer

l'échec de greffe rénale sur une fenêtre de 7 ans [Foucher et al., 2010]. Dans le contexte de maladie chronique, l'évolution de l'état de santé de patients est souvent évaluée à l'aide de marqueurs mesurés de manière répétée au cours du suivi des patients, comme le sérum de créatinine en transplantation rénale. L'information apportée par les mesures répétées au cours du suivi pourrait contribuer à améliorer la prédiction d'un événement clinique. Par extension, un outil pronostique dynamique est donc un outil utilisant des informations disponibles initialement mais intégrant également des informations accumulées au cours du suivi du patient (Figure 6.1). Un score pronostique dynamique peut se mettre à jour afin de recalculer la probabilité de survenue de l'événement en fonction d'informations longitudinales collectées. Plusieurs approches méthodologiques, telles que l'approche par landmarking [Houwelingen, 2007] ou les approches par modélisation conjointe de données longitudinales et données de survie [Rizopoulos, 2012], ont été envisagées pour estimer ces prédictions dynamiques. Une prédiction dynamique est définie comme l'estimation de la probabilité de survenue d'un événement entre les temps s et $s+t$ conditionnellement aux caractéristiques $\chi^{\pi}(s)$ disponibles jusqu'au temps de landmark s et au fait de ne pas avoir subi l'événement avant s, i.e. $\pi(s,t) = \mathbb{P}(s < T \leq s+t | \chi^{\pi}(s), T > s)$.

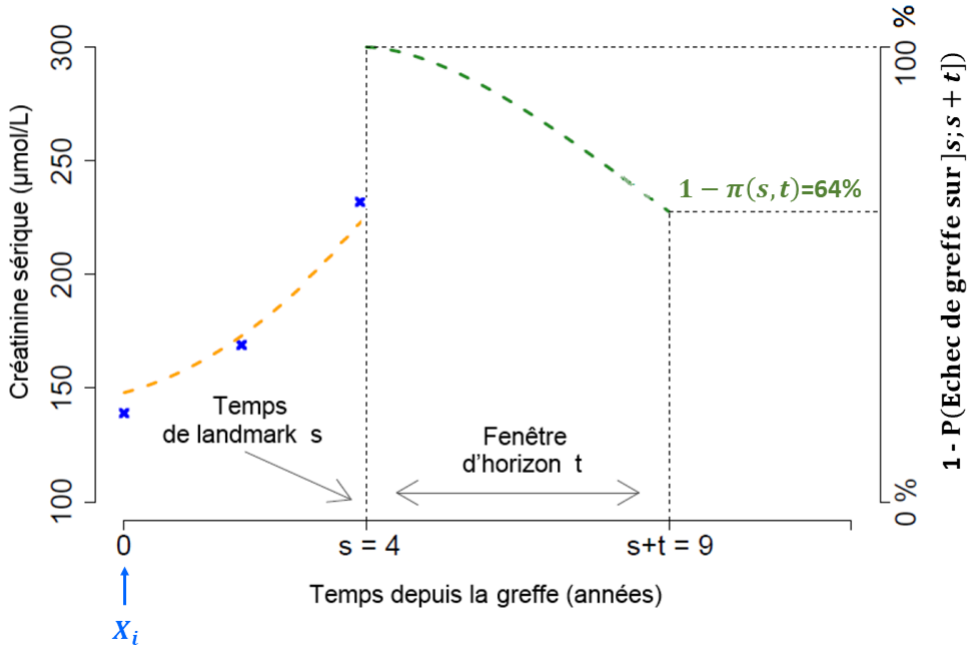


Figure 6.1: *Illustration d'une prédiction de l'échec de greffe estimée au temps de landmark $s = 4$ pour une fenêtre d'horizon t de 5 ans, et calculée en fonction de l'évolution du sérum de créatinine jusqu'en s et de covariables X_i connues à l'inclusion.*

Qu'il soit statique ou dynamique, un outil pronostique doit avoir de bonnes propriétés de calibration et de discrimination [Graf et al., 1999; Steyerberg et al., 2010]. L'indicateur le plus connu évaluant les capacités discriminantes est l'aire sous la courbe ROC évaluant à quel point les prédictions distinguent les individus qui vont connaître l'événement de ceux qui ne vont pas le

connaître. La calibration est la capacité d'un outil à faire coïncider prédictions et observations. Le Brier Score est un indicateur évaluant globalement les performances de calibration et de discrimination et se définit comme une erreur quadratique entre l'observation et la prédiction [Gerds and Schumacher, 2006]. Dans le travail de Fournier et al. [2018], nous reprenons les écritures de ces indicateurs dans le contexte dynamique et commentons les limites du Brier Score. Nous proposons un critère alternatif de type courbe de R^2 afin d'évaluer les performances globales d'un score pronostique dynamique. Pour cela, nous avons étendu une définition proposée par Graf et al. [1999] au contexte d'un score dynamique en standardisant le Brier Score $BS_\pi(s,t)$ par le Brier Score d'un modèle de référence.

Plus généralement, le travail de thèse de Marie-Cécile Fournier me semble très bien résumer notre travail pluridisciplinaire. En effet, Marie-Cécile aura eu l'occasion au cours de son travail de thèse de produire 3 travaux complémentaires. Le premier publié dans une revue d'Epidémiologie est présenté en section 5. Le deuxième, réalisé en collaboration avec Paul Blanche, est publié dans une revue de Biostatistique et est présenté dans cette section. Le troisième porte sur le développement d'un score pronostique de l'échec de greffe rénale et vient d'être accepté dans une revue de transplantation rénale [Fournier et al., 2019]. Il sera développé en section 9.2.

Received: 17 November 2016 | Revised: 4 November 2017 | Accepted: 6 November 2017

DOI: 10.1002/sim.7571

RESEARCH ARTICLE

WILEY Statistics in Medicine

An R^2-curve for evaluating the accuracy of dynamic predictions

Marie-Cécile Fournier[1,2] | Etienne Dantan[1] | Paul Blanche[3]

[1]INSERM UMR 1246–SPHERE, Nantes University, Tours University, Nantes, France

[2]ITUN Institut de Transplantation Urologie Néphrologie INSERM UMR 1064, Nantes, France

[3]LMBA, Université de Bretagne Sud, Vannes, Brittany, France

Correspondence
Paul Blanche, LMBA, Campus de Tohannic, Université de Bretagne Sud, 56017 Vannes, France.
Email: paul.blanche@univ-ubs.fr

In the context of chronic diseases, patient's health evolution is often evaluated through the study of longitudinal markers and major clinical events such as relapses or death. Dynamic predictions of such types of events may be useful to improve patients management all along their follow-up. Dynamic predictions consist of predictions that are based on information repeatedly collected over time, such as measurements of a biomarker, and that can be updated as soon as new information becomes available. Several techniques to derive dynamic predictions have already been suggested, and computation of dynamic predictions is becoming increasingly popular.

In this work, we focus on assessing predictive accuracy of dynamic predictions and suggest that using an R^2-curve may help. It facilitates the evaluation of the predictive accuracy gain obtained when accumulating information on a patient's health profile over time. A nonparametric inverse probability of censoring weighted estimator is suggested to deal with censoring. Large sample results are provided, and methods to compute confidence intervals and bands are derived. A simulation study assesses the finite sample size behavior of the inference procedures and illustrates the shape of some R^2-curves which can be expected in common settings. A detailed application to kidney transplant data is also presented.

KEYWORDS

joint models, landmarking, longitudinal data, prediction modeling, predictive accuracy, survival analysis

1 | INTRODUCTION

In the field of personalized medicine, prediction scores of major clinical events are increasingly popular and may help physicians for medical decision making. In the context of chronic diseases, longitudinal markers are often routinely measured to assess the patient's health evolution. This may bring valuable information to regularly update prediction scores all along the patient follow-up. For instance, serum creatinine (SCr) is repeatedly measured among kidney transplant recipients as a marker of graft filtration rate.[1] Prediction scores have already been proposed in kidney transplantation to predict return to dialysis[2,3] or death,[4,5] but they do not use any longitudinal measurements of SCr. Meanwhile, there is a growing interest on methods to compute dynamic predictions in the statistical literature.[6] Among them, the so-called landmarking[7-9] and joint modeling approaches[10-12] are very popular. In this context, dynamic predictions are defined as predictions of a clinical event which can be dynamically updated when the time of making predictions changes.

 Copyright © 2017 John Wiley & Sons, Ltd.

Interestingly, such dynamic predictions enable to take into account the valuable information consisting of the entire marker trajectory known at the time of making predictions.

As pointed out by Graf et al,[13] the growing number of prognostic models proposed in the medical literature emphasizes the need to evaluate their accuracy. Because the clinical utility of predictions (dynamic or not) strongly relies on their discrimination and calibration properties, it is widely recommended to evaluate them carefully.[14] The concept of discrimination relates to the fact that different subjects which are not all as likely to experience the event should have different predicted values. The concept of calibration relates to the fact, for any percentage x, we expect that x subjects out of 100 experience the event among all subjects that have a predictions of experiencing the event equal to x %. Graf et al[13] and Gerds and Schumacher[15] first used inverse probability of censoring weighting estimators to estimate mean squared errors for predictions, that is, the so-called Brier Score. More recently, methods about Brier Scores and ROC curves estimation in the specific dynamic context have also been suggested.[16] Dynamic ROC curve enables to evaluate discrimination performances and relates to popular and easy understandable concepts, as it is a graphical representation of sensitivity and specificity concepts. However, the ROC curve cannot assess calibration while the Brier Score is an interesting tool which quantifies both discrimination and calibration performances simultaneously. Schoop et al[17] first extended the work of Graf et al[13] to deal with longitudinal patients measurements. In a related context, Parast et al[18] also recently suggested to use the Brier Score to evaluate the prediction accuracy of long-term predictions which incorporate short-term event time information. While Schoop et al[17] mainly focused on consistent estimation of the Brier Score, Blanche et al[16] have further suggested inference procedures based on large sample properties to compute confidence regions and tests.

Although the Brier Score can be interpreted as an expected squared distance between observations and predictions, this mathematical interpretation is somehow difficult to directly translate into a clinical interpretation. Especially, the interpretation of the Brier Score highly depends on the marginal probability of event.[19] This is an important issue in the dynamic prediction context as this probability may vary substantially from one time of making prediction to another. This further complicates the interpretation of the trend of Brier Score estimates over times of making prediction. However, the Brier Score of predictions obtained from any prediction model can be systematically compared to the one of a "null" prediction model, which does not use any subject specific information and defines a benchmark value. As pointed out by Graf et al[13] and Parast et al,[18] this approach provides an interesting R^2-type criterion, closely related to the popular concept of "explained variation." Several explained variation criteria for predictive models have long been investigated in many contexts including survival analysis.[20-22] Nevertheless, to our knowledge, there is currently very limited literature for the specific context of dynamic predictions.

The main objective of the manuscript is to present an R^2-curve for dynamic predictions for evaluating predictive accuracy. This R^2-curve has so far been only briefly introduced in section 3.4 of van Houwelingen and Putter[8] and deserves more attention. We present related inference procedures and new insights about the interpretation of the R^2-curve. Section 2 presents the Brier Score and the R^2-curve for dynamic predictions and their main characteristics and Section 3 presents an application to the kidney transplant data which have motivated this work. The details of the inference procedures and the results of a simulation study are provided in the Supporting Information. A web link to R codes and reproducible examples are also provided in the Supporting Information, to make the methodology easily usable by others. A discussion concludes the manuscript in Section 4.

2 | AN R^2-CURVE FOR DYNAMIC PREDICTIONS

2.1 | Notations

Let T denote a time-to-event, $S(u) = \mathbb{P}(T > u)$ the survival function of T evaluated at time u, and $S(u|s) = \mathbb{P}(T > u|T > s)$ the survival function of T at time u given that the patient is event-free at landmark time s. Because of censoring, we observe $\tilde{T} = \min(T, C)$ and $\Delta = \mathbb{1}\{T \leq C\}$, where $\mathbb{1}\{\cdot\}$ denotes the indicator function. Let $\pi(s, t)$ denote a subject-specific dynamic prediction, which can be any *estimate* of $\mathbb{P}(s < T \leq s + t \,|\, \mathcal{X}^{\pi}(s), T > s)$, obtained from any prediction model, for a given time horizon t and subject-specific characteristics $\mathcal{X}^{\pi}(s)$ available at landmark time s. We would like to emphasize that in this manuscript, we do not make any assumption about the dynamic prediction $\pi(s, t)$. It can be a prediction obtained from any model, well specified or misspecified, and it can be biased.

2.2 | Brier Score: definition, decomposition, and interpretation

Following Schoop et al,[23] the Brier Score of the dynamic prediction $\pi(s, t)$ is defined as

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

with $D(s,t) = \mathbb{1}\{s < T \leq s+t\}$ the indicator function of experiencing the event within the interval $(s; s+t]$. As any error-based criterion, the interpretation is the lower, the better. Following ideas going back to Murphy,[24] it is shown in the Supporting Information that the Brier Score can be decomposed as

$$\begin{aligned}\mathrm{BS}_\pi(s,t) = \mathbb{E}\left[\{\pi(s,t) - \mathbb{E}\left[D(s,t)|\pi(s,t), T > s\right]\}^2 | T > s\right] \\ + \mathrm{Var}\left[D(s,t) | T > s\right] - \mathrm{Var}\left[\mathbb{E}\left\{D(s,t)|\pi(s,t), T > s\right\} | T > s\right].\end{aligned} \tag{1}$$

The first term in Equation 1 measures the *calibration* of the predictions. It measures how close are the dynamic predictions $\pi(s,t)$ to the "true underlying" risk of event in $(s; s+t]$ given $\{\pi(s,t), T > s\}$, which is $\mathbb{E}[D(s,t)|\pi(s,t), T > s] = \mathbb{P}(s < T \leq s+t|\pi(s,t), T > s)$. Note that if the dynamic predictions are *weakly calibrated* at (s,t) (as defined by Pepe and Janes[25]), that is, if $\pi(s,t) = \mathbb{E}[D(s,t)|\pi(s,t), T > s]$, then this first term is equal to zero. The second term in Equation 1 is sometimes named the *uncertainty term*. It is equal to $S(s+t|s)\{1 - S(s+t|s)\}$, and it only depends on the marginal risk of experiencing the event within $(s; s+t]$ given that $T > s$. The third term in Equation 1 measures to what extent the dynamic predictions $\pi(s,t)$ *discriminate* subjects that will experience an event within $(s; s+t]$ from the others, among those at risk at s. Indeed, the more different the risks of event within $(s; s+t]$ given information $\{\pi(s,t), T > s\}$, for different values of the prediction $\pi(s,t)$, which is $\mathbb{E}[D(s,t)|\pi(s,t), T > s]$, the bigger the variance and the smaller the Brier Score.

2.3 | R^2: definition and interpretation

Unfortunately, the curve of $BS_\pi(s,t)$ over s is not easy to interpret because the range of values that $BS_\pi(s,t)$ can take strongly depends on the landmark time s, when t is fixed. Indeed, because of the second term in Equation 1, the value of $BS_\pi(s,t)$ strongly depends on the survival probability $S(s+t|s)$, which typically changes with the landmark time s. Specifically, the smaller $|S(s+t|s) - 0.5|$ the higher the Brier Score of the predictions. This is unfortunate because (1) this has nothing to do with the inherent prediction accuracy of the predictions and (2) this *uncertainty term* may be a leading term. For instance, in the application presented in Section 3, the ratio between $BS_\pi(s,t)$ and this term is estimated to be 91.5% at $s = 0$ with $t = 5$. Consequently, a naive interpretation of the increasing (or decreasing) trend of the Brier Score curve can be misleading as it can reflect a change of scale only. This is unfortunate because, in the context of dynamic prediction, it is natural to be interested in knowing whether prediction accuracy increases with increasing landmark time s, ie, with increasing information.

To circumvent this issue, one solution is to standardize the Brier Score using a benchmark value.[13,18,26] This leads to the R^2-curve which has already been briefly introduced in section 3.4 of van Houwelingen and Putter.[8] For any s and t, let us define the "null" prediction $\pi_0(s,t)$ which is estimated from the best "null" model that does not use any subject-specific information, that is $\pi_0(s,t) = \mathbb{P}(s < T \leq s+t|T > s) = 1 - S(s+t|s)$. Since $\pi_0(s,t) = \mathbb{E}[D(s,t)|T > s]$, the Brier Score associated to $\pi_0(s,t)$, denoted by $BS_0(s,t)$, is simply

$$\mathrm{BS}_0(s,t) = \mathbb{E}\left[(D(s,t) - \pi_0(s,t))^2 | T > s\right] = \mathrm{Var}\left\{D(s,t)|T > s\right\}.$$

Following Graf et al,[13] a natural definition of an R^2-type criterion is therefore

$$\mathrm{R}^2_\pi(s,t) = 1 - \frac{\mathrm{BS}_\pi(s,t)}{\mathrm{BS}_0(s,t)}.$$

The scale for the values of $\mathrm{R}^2_\pi(s,t)$ is easily understandable. $\mathrm{R}^2_\pi(s,t)$ is equal to one when $\pi(s,t) = D(s,t)$, which corresponds to the case in which the predictions are able to perfectly distinguish the subjects that will experience an event in $(s; s+t]$ from those who will not. However, it is the most often unrealistic to think that any kind of information available at time s, ie, any $\mathcal{X}^\pi(s)$, could lead to such predictions. By contrast, $\mathrm{R}^2_\pi(s,t)$ will be close to 0 when $\pi(s,t) = \pi_0(s,t)$, which corresponds to the case in which whatever the subjects' characteristics all predictions are the same, and are equal to the marginal t-year risk of event. In extreme cases, if the subject-specific information is so wrongly used by a working prediction model that it would be better to not use it at all (which unfortunately happens when using overfitted prediction models) then $\mathrm{R}^2_\pi(s,t)$ can be negative. However, most of the time $\mathrm{R}^2_\pi(s,t)$ is expected to take values within $(0,1)$, with higher values indicating better predictive accuracy.

Assuming that the dynamic predictions $\pi(s,t)$ are *weakly* calibrated, meaning that $\pi(s,t) = \mathbb{E}[D(s,t)|\pi(s,t), T > s]$, this $R^2_\pi(s,t)$ also has several interesting interpretations. The first one is that it can be interpreted as an "explained variation" criterion. Indeed, under the weak calibration assumption the first term in Equation 1 is equal to zero and it follows that:

$$R^2_\pi(s,t) = \frac{\text{Var}\{\mathbb{E}[D(s,t)|\pi(s,t), T > s]|T > s\}}{\text{Var}\{D(s,t)|T > s\}} = \frac{\text{Var}\{\pi(s,t)|T > s\}}{\text{Var}\{D(s,t)|T > s\}}.$$

Therefore, under this calibration assumption, $R^2_\pi(s,t)$ can indeed be interpreted as a ratio of the variance of the prediction $\pi(s,t)$ over the variance of the outcome $D(s,t)$. Two other interesting interpretations can be pointed out. As shown in the Supporting Information, a little algebra also leads to a correlation interpretation,

$$R^2_\pi(s,t) = \text{Corr}^2\{D(s,t), \pi(s,t)|T > s\},$$

and also to a mean risk difference interpretation,

$$R^2_\pi(s,t) = \mathbb{E}\{\pi(s,t)|D(s,t) = 1, T > s\} - \mathbb{E}\{\pi(s,t)|D(s,t) = 0, T > s\}.$$

Similar interpretations have already been pointed out in discussions about R^2-criteria for logistic regression.[27-30] The correlation interpretation is similar to a well-known interpretation which holds with univariate linear models. The mean risk difference interpretation, which is less intuitive, is especially useful. First, as pointed out by Pepe et al,[28] the mean risk difference interpretation helps to understand the connection between the change in explained variation between two rival prediction models and the IDI reclassification index. Second, it facilitates the understanding of why this R^2-criterion will typically take lower values than those we may sometimes observe with the usual R^2 coefficient of a linear model. Whereas the R^2 coefficient of a linear model may be observed large, close to one, when the variance of the random noise is small, the mean risk difference will rarely be expected to be very close to one in practice. Indeed, as already mentioned above, it is often unrealistic to assume that it is possible to give a prediction close to one to all subjects who will experience the event and a prediction close to zero to all those who will not.

2.4 | Inference

In the Supporting Information, we provide details about the computation of consistent pointwise estimates and of pointwise confidence intervals and simultaneous confidence bands. They are based on large sample results, and it is assumed that we observe a sample of n independent and identically distributed subjects $\{(\pi_i(s,t), \tilde{T}_i, \Delta_i), i = 1, \ldots, n\}$. Let us point out that this implies that the dynamic model used to compute $\pi_i(s,t)$ from the observed subject-i-specific characteristics $\mathcal{X}_i^\pi(s)$, for all $i = 1, \ldots, n$, has been estimated from an independent learning data set. Following previous works,[16,18,23] inference is based on the Inverse Probability of Censoring Weighting method. It is a versatile approach as it does not depend on how the predictions were computed and because it is based on the common independent censoring assumption only.

The results of a simulation study are provided in the Supporting Information. Overall, they suggest that the inference procedures perform well. The simulation study also provides us with different examples of R^2-curves and illustrates different interpretations.

3 | APPLICATION TO KIDNEY TRANSPLANT DATA

The objective was to evaluate the accuracy of dynamic predictions of graft failure among kidney transplant recipients. We focus on the chronic phase of the disease starting 1 year after transplantation as the graft failure risk should be modeled differently before and after 1 year.[2,31] For the sake of simplicity, the dynamic predictions were computed from a simplified version of a previously published joint model for longitudinal and time-to-event data.[32] The dynamic predictions depend on repeatedly collected SCr values over time and on two binary and one continuous baseline covariates: recipient gender (SEX), cardiovascular history (CARDV), and age (AGE). Following discussions with clinicians, we chose $t = 5$ years as time horizon and landmark times s ranging from $s = 0$, corresponding to 1-year post transplantation, to $s = 5$, equally spaced every 0.5 year.

3.1 | Available data

Data come from the French multicentric observational and prospective DIVAT cohort (www.divat.fr). Adult recipients transplanted for the first or second time between 2000 and 2013 from living or heart beating donor, and alive with a

functioning graft at 1-year post transplantation have been selected in our sample. Time-to-event was defined as the delay from the 1-year visit to either return to dialysis or death with a functioning graft. Serum creatinine (μmol/L) was routinely recorded yearly until the graft failure event.

Our initial data sample was comprised of 4121 kidney recipients, which has been randomly divided in two samples for learning and validation. The joint model which is required to compute dynamic predictions has been fitted using two third of the subjects (n = 2749). The remaining one third of the subjects (n = 1372) have been used to estimate the accuracy of the dynamic predictions computed from independent subjects.

3.2 | Dynamic prediction modeling

The shared random effect joint model combines a linear mixed model (2a) with a parametric proportional hazards model (2b):

$$\begin{cases} Y_i(t_{ij}) = (\beta_0 + b_{0i} + \beta_{0,\text{age}}\text{AGE}_i + \beta_{0,\text{sex}}\text{SEX}_i) + (\beta_1 + b_{1i} + \beta_{1,\text{age}}\text{AGE}_i) \times t_{ij} + \varepsilon(t_{ij}) & \text{(2a)} \\ \qquad = m_i(t_{ij}) + \varepsilon(t_{ij}), & \\ h_i(t) = h_0(t) \exp\left\{ \gamma_{\text{age}}\text{AGE}_i + \gamma_{\text{cardV}}\text{CARDV}_i + \alpha_1 m_i(t) + \alpha_2 \dfrac{dm_i(t)}{dt} \right\}, & \text{(2b)} \end{cases}$$

where $Y_i(\cdot)$ denotes the longitudinal biomarker trajectory of subject i (here the log-transformation of SCr), and t_{ij} the time of the j^{th} ($j = 1, \ldots, n_i$) measurement of patient i ($i = 1, \ldots, n$). The evolution of the measurements was modeled by the sum of an expected subject-specific trend $m_i(\cdot)$ plus an error term $\varepsilon(t_{ij}) \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, where $\varepsilon(t) \perp\!\!\!\perp \varepsilon(t')$ for $t \neq t'$. Shared random effects were assumed to be normally distributed, $(b_{0i}, b_{1i})^\top \sim \mathcal{N}(\mathbf{0}, \Sigma)$ where Σ is an unstructured variance-covariance matrix. The baseline hazard $h_0(t)$ was modeled from a Weibull distribution. The hazard function for the time to graft failure was modeled as depending on both the current value of the marker $m_i(t)$ and the current intensity of the marker deterioration $dm_i(t)/dt$ through parameters α_1 and α_2.

From this joint shared random effect model fitted using two-thirds of the subjects (n = 2749), we computed the subject-specific predictions at each landmark times s for the remaining one-third of the subjects (n = 1372). By exploiting the fact that the joint model is fully parametric, dynamic predictions were computed as estimates of $\mathbb{P}(s < T_i \leq s + t | \mathcal{X}_i^\pi(s), T_i > s)$ by plugging-in fixed parameters estimates and using a first-order expansion and empirical Bayes estimates for (b_{0i}, b_{1i}), as detailed by Rizopoulos.[33,34]

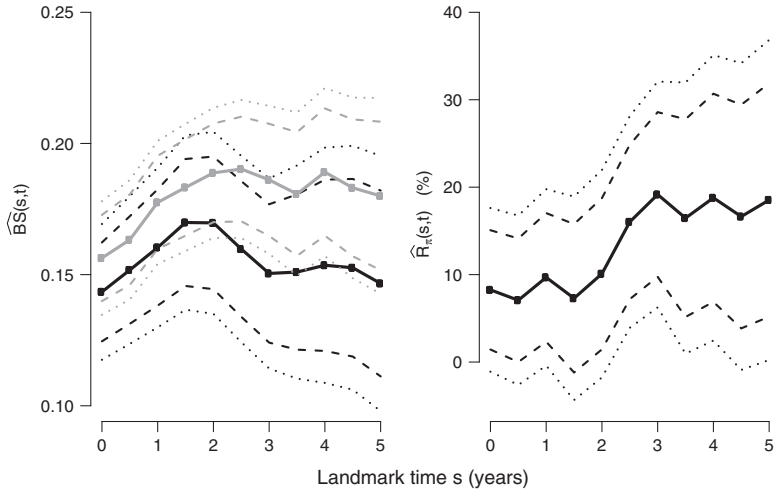


FIGURE 1 Brier Score curves and R^2-curve. $\widehat{R}_\pi(s,t)$ and $\widehat{BS}_\pi(s,t)$, in black, and $\widehat{BS}_0(s,t)$, in grey, over s with $t = 5$ years. Subject specific dynamic predictions $\pi(s,t)$ depend on sex, age, cardiovascular history and repeatedly collected measurements of serum creatinine until time s. Plain lines display pointwise estimates, dashed lines display 95% pointwise confidence intervals and dotted lines display 95% simultaneous confidence bands. DIVAT data, n = 1372

3.3 | Dynamic prediction accuracy

Figure 1 displays the curves of $\widehat{\mathrm{BS}}_\pi(s,t)$ over s and of $\widehat{\mathrm{BS}}_0(s,t)$ over s. As expected, the curve of $\widehat{\mathrm{BS}}_\pi(s,t)$ over s is below the curve of $\widehat{\mathrm{BS}}_0(s,t)$ over s. Interestingly, we can notice that the difference between $\widehat{\mathrm{BS}}_\pi(s,t)$ and $\widehat{\mathrm{BS}}_0(s,t)$ strongly changes with the landmark time s. As already mentioned in Section 2.3, this illustrates that the value of $\widehat{\mathrm{BS}}_\pi(s,t)$ is typically difficult to interpret across different landmark times s. For instance, it would be misleading to interpret the increasing trend from $s = 0$ to $s = 2$ as suggesting a decrease in predictive accuracy and the decrease from $s = 2$ to $s = 5$ as suggesting an increase in predictive accuracy. Only the trend of the curve of the ratio of $\widehat{\mathrm{BS}}_\pi(s,t)$ and $\widehat{\mathrm{BS}}_0(s,t)$ over s can reasonably be interpreted, which is precisely the motivation of displaying the estimate of the R^2-curve.

The estimated R^2-curve (Figure 1) enables to appraise how well the subject-specific dynamic predictions computed from the joint model perform as compared to the marginal prediction. As the trend is increasing, it suggests that the more available information to make predictions, the better the accuracy in terms of quadratic error, which quantify both discrimination and calibration. By contrast to the conclusions of the misleading interpretation mentioned above, it seems,

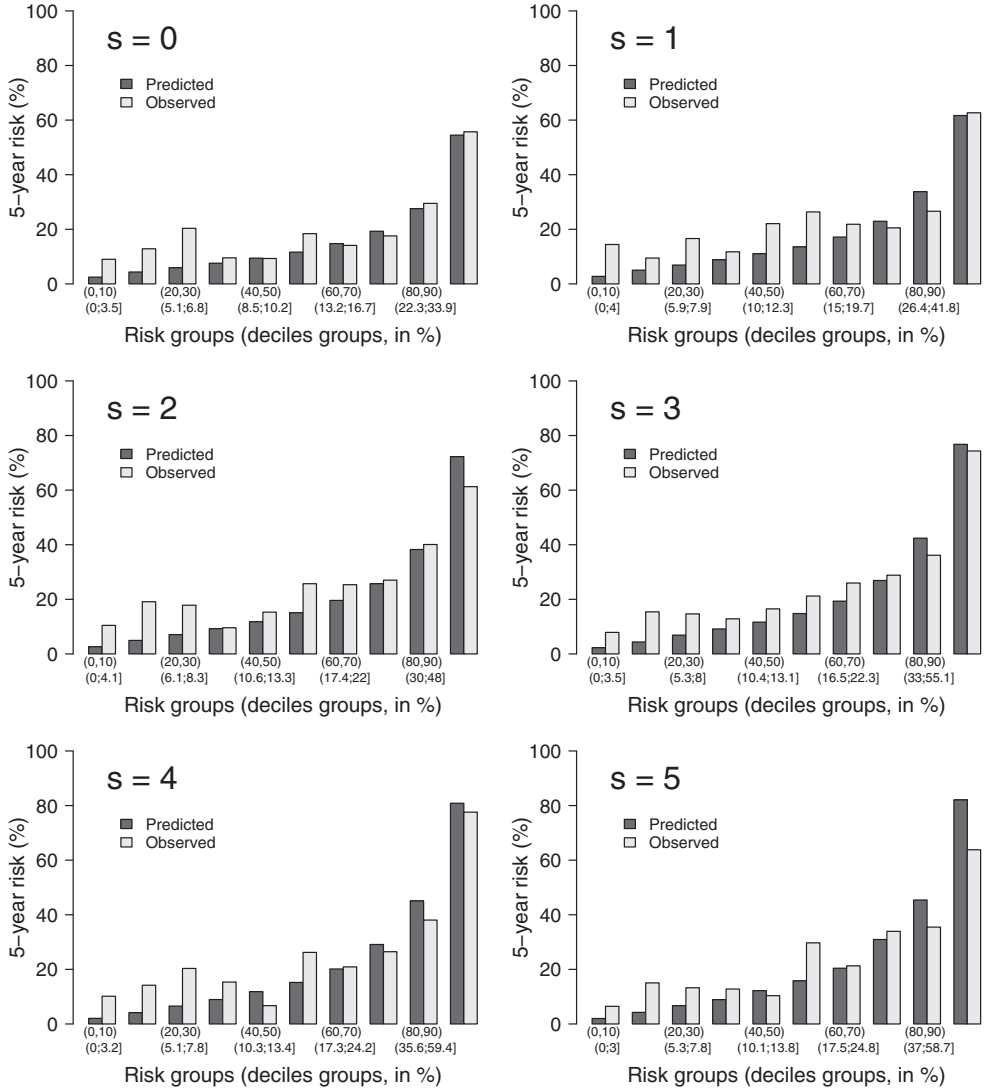


FIGURE 2 Calibration plot for landmark times $s = 0, 1, 2, 3, 4, 5$ years. Mean predicted risks and observed risks (Kaplan-Meier) are displayed for each subgroup. The x-axis displays the ranges of the values of the predictions which are included in some of the decile groups. For instance, at $s = 5$ years, the decile group which includes subjects having predicted risk values within the second and the third deciles (20% and 30%) includes predicted risks ranging from 5.3% to 7.8%. DIVAT data, n = 1372

here, rather clear that in terms of quadratic error the results tend to suggest that the more available information to compute predictions, the better the predictive accuracy of the predictions computed from the joint model.

Because of the rather large confidence intervals and bands, the previous comments about the pointwise estimates should nevertheless be moderated. The fact that the confidence regions are wide illustrate the well-known bias-variance dilemma: because the estimates are nonparametric (unbiased without requiring any assumption) their variance can be large. It is expected that the confidence regions would be narrower if the sample size was larger and/or if the proportions of censored data at each landmark time were lower. In our sample, the proportions of censored observations within $(s; s + t]$ among subjects at risk at s, when s varies, are rather large, ranging from 55% to 65%.

To evaluate prediction accuracy, it is also often interesting to look at calibration plots which enable to compare observed and predicted risks within groups of subjects having similar predictions. They are typically used to assess the weak calibration assumption.[25] Figure 2 displays such calibration plots. Deciles of the dynamic predictions distribution were used to define the groups of patients having similar predictions at each landmark time. For instance, by comparing the grey bars to the black, it seems that, for all landmark times, the risk of event is under estimated in groups of patients corresponding to the first deciles of the dynamic predictions. Interestingly, this type of calibration plots also helps to better understand each term in the Brier Score decomposition of Equation 1, and so to better understand the R^2-curve. The first term in (1) can be interpreted as the mean squared difference between the height of the grey bars (observed risks) and of the black bars (predicted risks). The second term can be interpreted as $h(1-h)$ where h denotes the average height of the grey bars (average observed risks). The third term can be interpreted as the variance of the heights of the grey bars, from one group to another (variance of the observed risks).

The discrimination performance at each landmark time s can also be measured by the dynamic area under the ROC curve (AUC),[16] that is, $\text{AUC}_\pi(s,t) = \mathbb{P}(\pi_i(s,t) > \pi_j(s,t) | D_i(s,t) = 1, D_j(s,t) = 0, T_i > s, T_j > s)$, where i and j denote the indices of two independent subjects. Interestingly, the value of the AUC at each landmark time can also be gauged graphically from a calibration plot,[25] and this helps to compare its interpretation to that of the R^2-curve. Briefly, because Figure 2 shows that the observed risks (grey bars) seem more and more increasing from the left to the right when the landmark time increases, it shows an increasing AUC trend with increasing landmark times. Figure 3 displays the AUC-curve, which is estimated higher than 65% for all landmark times. This can be interpreted as follows. For all landmark times, the probability that a subject experiencing a graft failure in a 5-year prediction window has a higher dynamic prediction than that of a subject event-free at the end of the prediction window is greater than 65%. Overall, the trend of the AUC-curve tends to suggest that the discrimination abilities increase with the landmark time and so with increasing available information about SCr subject-specific trajectory.

The interpretation of the R^2-curve is therefore here concordant with the interpretation of the AUC-curve. However, let us point out that, as they do not measure the same aspect of predictive accuracy, this concordance is not systematic. This is illustrated by the simulation scenario (5) presented in the Supporting Information, where the assumption of the weak calibration of the predictions does not hold.

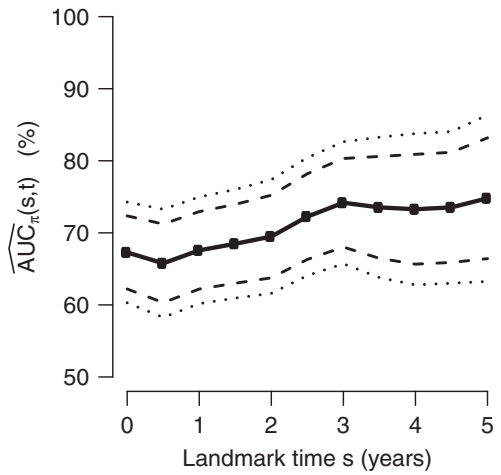


FIGURE 3 AUC-curve. $\widehat{\text{AUC}}_\pi(s,t)$ over s is displayed by the plain line, 95% pointwise confidence intervals are displayed by the dashed lines and 95% simultaneous confidence bands by the dotted lines. DIVAT data, n = 1372

4 | DISCUSSION

Under the calibration assumption, the R^2-curve presented in this manuscript has several interesting interpretations. Without assuming any assumption, it still has a natural interpretation and follows the recommendations about the correct choice of an R^2 statistic introduced by Kvålseth.[35] In particular, it is model-free and is applicable to evaluate the accuracy of any type of dynamic predictions, whatever the kind of working model from which they have been computed. Asymptotic results for the proposed estimator were established and inference procedures to build pointwise and simultaneous confidence regions were derived. Our simulation results suggest that they perform well.

Tests for comparing several R^2-curves from different dynamic predictions could also be derived. Such tests would, however, be equivalent to previously proposed tests for comparing Brier Score curves,[16] as they would have the same null hypothesis.

One limitation of the proposed inference procedures, which is exemplified by the application and the simulations, is the need of rather large sample sizes and long follow-ups to enable the estimation of narrow confidence intervals and bands. This is unfortunately the drawback of using a versatile nonparametric approach.

Other R^2-curves could be considered by using other benchmark Brier score than the Brier Score of the best "null" predictions that we have considered. The Brier Score of predictions from simpler models than the one from which the predictions under study are computed could be used (eg, simpler models including less covariates as in Parast et al[18]). However, that would violate Kvålseth's list of good properties that a "good" R^2 statistic should have.[35]

In this paper, we show that it is easier to interpret the trend of R^2-curves over landmark time as compared to the trend of the Brier Score curves. In our opinion, this is an important point as, within the dynamic prediction context, we often wish to know whether predictions are more accurate when we accumulate more information over time, ie, when the landmark time s increases. The R^2-curve that we have suggested does not only enable us to assess whether such an increase exists but also enables us to quantify it.

ACKNOWLEDGEMENTS

We thank the DIVAT consortium for providing the data of the DIVAT cohort (www.divat.fr). PB is also grateful to the Bettencourt Schueller foundation for its support (www.fondationbs.org). We also thank the editor and the associate editor and two anonymous referees for their constructive comments that significantly improved this article.

ORCID

Paul Blanche http://orcid.org/0000-0003-1415-7976

REFERENCES

1. Luis-Lima S, Marrero-Miranda D, González-Rinne A, et al. Estimated glomerular filtration rate in renal transplantation: the nephrologist in the mist. *Transplantation*. 2015;99(12):2625-2637.
2. Foucher Y, Daguin P, Akl A, et al. A clinical scoring system highly predictive of long-term kidney graft survival. *Kidney Int*. 2010;78(12):1288-94.
3. Foucher Y, Meurette A, Daguin P, et al. A personalized follow-up of kidney transplant recipients using video conferencing based on a 1-year scoring system predictive of long term graft failure (TELEGRAFT study): Protocol for a randomized controlled trial. *BMC Nephrology*. 2015;16(1):1-6.
4. Baskin-Bey ES, Kremers W, Nyberg SL. A recipient risk score for deceased donor renal allocation. *Am J Kidney Dis*. 2007;49(2):284-293.
5. Lorent M, Giral M, Pascual M, et al. Mortality prediction after the first year of kidney transplantation: an observational study on two European cohorts. *PloS One*. 2016;11(5):1-15 e0155278.
6. Proust-Lima C, Blanche P. Dynamic predictions. *Wiley StatsRef: Statistics Reference Online*. 2016:1-6.
7. van Houwelingen HC. Dynamic prediction by landmarking in event history analysis. *Scand J Stat*. 2007;34(1):70-85.
8. van Houwelingen HC, Putter H. *Dynamic Prediction in Clinical Survival Analysis*. Boca Raton: CRC Press; 2012.
9. Nicolaie MA, van Houwelingen HC, de Witte TM, Putter H. Dynamic pseudo-observations: a robust approach to dynamic prediction in competing risks. *Biometrics*. 2013;69(4):1043-1052.
10. Rizopoulos D. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Boca Raton: CRC Press; 2012.
11. Proust-Lima C, Taylor JMG. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. *Biostatistics*. 2009;10(3):535-549.
12. Taylor JMG, Park Y, Ankerst DP, et al. Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics*. 2013;69(1):206-213.

13. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med*. 1999;18:2529-2545.
14. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*. 2010;21(1):128-138.
15. Gerds TA, Schumacher M. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical J*. 2006;48(6):1-5.
16. Blanche P, Proust-Lima C, Loubère L, Berr C, Dartigues J-F, Jacqmin-Gadda H. Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics*. 2015;71(1):102-113.
17. Schoop R, Graf E, Schumacher M. Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics*. 2008;64(2):603-610.
18. Parast L, Cheng S-C, Cai T. Landmark prediction of long-term survival incorporing short-term event time information. *J Am Stat Assoc*. 2012;107(500):1492-1501.
19. Steyerberg E. *Clinical Prediction Models*. New York: Springer; 2009.
20. van Houwelingen HC, Le Cessie S. Predictive value of statistical models. *Stat Med*. 1990;9(11):1303-1325.
21. Stare J, Pohar Perme M, Henderson R. A measure of explained variation for event history data. *Biometrics*. 2011;67(3):750-759.
22. Choodari-Oskooei B, Royston P, Parmar M. A simulation study of predictive ability measures in a survival model I: explained variation measures. *Stat Med*. 2012;31(23):2627-2643.
23. Schoop R, Schumacher M, Graf E. Measures of prediction error for survival data with longitudinal covariates. *Biometrical J*. 2011;53(2):275-293.
24. Murphy AH. A new vector partition of the probability score. *J Appl Meteorol*. 1973;12(4):595-600.
25. Pepe M, Janes H. Methods for Evaluating Prediction Performance of Biomarkers and Tests. In: Lee M-L, Gail G, Cai T, Pfeiffer R, Gandy A, eds. *Risk Assessment and Evaluation of Predictions*. New York: Springer; 2013:107-142.
26. Lawless JF, Yuan Y. Estimation of prediction error for survival models. *Stat Med*. 2010;29(2):262-274.
27. Hu B, Palta M, Shao J. Properties of R^2 statistics for logistic regression. *Stat Med*. 2006;25(8):1383-1395.
28. Pepe M, Feng Z, Gu J. Comments on 'evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond' by mj pencina et al., Statistics in Medicine (doi: 10.1002/sim. 2929). *Stat Med*. 2008;27(2):173-181.
29. Gu W, Pepe M. Measures to summarize and compare the predictive capacity of markers. *The Int J Biostat*. 2009;5(1):1-47.
30. Tjur T. Coefficients of determination in logistic regression models—a new proposal: the coefficient of discrimination. *The American Statistician*. 2009;63(4):366-372.
31. Galichon P, Xu-Dubois Y-C, Finianos S, Hertig A, Rondeau E. Clinical and histological predictors of long-term kidney graft survival. *Nephrology, Dialysis, Transplantation*. 2013;28(6):1362-1370.
32. Fournier M-C, Foucher Y, Blanche P, Buron F, Giral M, Dantan E. A joint model for longitudinal and time-to-event data to better assess the specific role of donor and recipient factors on long-term kidney transplantation outcomes. *Eur J Epidemiol*. 2016;31(5):469-479.
33. Rizopoulos D. JM: An R package for the joint modelling of longitudinal and time-to-event data. *J Stat Software*. 2010;35(9):1-33.
34. Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*. 2011;67(3):819-829.
35. Kvålseth TO. Cautionary note about R^2. *The American Statistician*. 1985;39(4):279-285.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Fournier MC, Dantan E, Blanche P. An R^2-curve for evaluating the accuracy of dynamic predictions. *Statistics in Medicine*. 2017;1–9. https://doi.org/10.1002/sim.7571

- Chapitre 7 -

Stratification du risque de survenue d'un événement : une prise de décision médicale centrée sur le patient

Dantan E, Foucher Y, Lorent M, Giral M and Tessier P. Optimal threshold estimator of a prognostic marker by maximizing a time-dependent expected utility function for a patient-centered stratified medicine. *Stat Methods Med Res*. 2018 Jun ;27(6) :1847-1859. Epub 2016 Oct 20.

Préambule

Le projet d'équipe INSERM UMR 1246 - SPHERE porte sur le développement de méthodes dans le domaine de la recherche en santé centrée sur le patient. Au sein de notre groupe de travail, nous avons initié plusieurs travaux de recherche autour de ces questions (voir la section 3.3). J'ai choisi de mettre en avant notre travail publié dans *Statistical Methods and Medical Research* car il s'agit de l'une de nos premières contributions scientifiques dans le domaine de la médecine de précision en lien avec le projet d'équipe SPHERE. Nous y abordons la question de la mise en place d'une médecine stratifiée et notamment de l'identification d'un seuil de décision médicale à partir d'un marqueur quantitatif [Dantan et al., 2018].

A partir d'un marqueur quantitatif, nous cherchons à estimer un seuil pour définir deux groupes de sujets pour lesquels deux prises en charge alternatives peuvent être envisagées. Traditionnellement, l'identification d'un tel seuil peut se faire à partir des aptitudes discriminantes d'un marqueur. L'index de Youden, par exemple, correspond à la valeur du marqueur qui permet de maximiser la somme de la sensibilité et de la spécificité [Youden, 1950]. Mais, Subtil and Rabilloud [2015] rappelle que cet indicateur purement statistique ignore les préférences des patients face aux conséquences des thérapies envisagées. Ceci apparaît clairement non recevable dans de nombreux contextes pour lesquels la balance bénéfice/risque est à discuter. Par exemple, en cancérologie, l'intérêt d'une chimiothérapie se pose en fonction de la balance entre gain d'espérance de vie et perte de qualité de vie du fait du traitement, en particulier chez les

personnes les plus âgées.

Alors que le développement d'outils de prédiction ne cesse de croître, un très faible pourcentage de ces outils est finalement utilisé en pratique clinique [Joyner and Paneth, 2015]. Nous évoquions dans la section 6 que cela puisse venir de certains problèmes méthodologiques. J'ajouterai que cela peut également être lié au manque de considération du point de vue des patients dans la prise de décision médicale, c'est-à-dire dans l'arbitrage entre quantité et qualité de vie. Les recherches centrées sur le patient ont pour objectif de permettre aux patients d'exprimer leurs opinions face aux différentes options thérapeutiques [Ziegelstein, 2015]. Dans la question de l'identification d'un seuil de décision médicale, nous proposons d'intégrer les conséquences des bras de traitement sur la qualité de vie en considérant les préférences des patients pour les différents états de santé. Dans notre travail, nous définissons une fonction d'utilité dépendante du temps considérant les différents états de santé pondérés par la qualité de vie perçue des patients pour ces différents états. Maximiser cette fonction d'utilité revient à maximiser le nombre de QALYs dans chacun des bras de traitement et permet d'identifier un éventuel seuil de décision. Initialement proposés en évaluation médico-économique, les QALYs sont une mesure combinant quantité et qualité de vie. Ils se calculent en pondérant chaque période de temps par un score d'utilité compris entre 0 et 1 reflétant les préférences des patients pour les différents états de santé [Weinstein et al., 2009], 0 qualifiant le décès et 1 la parfaite santé.

Dans la littérature, l'utilité d'un marqueur pronostique est souvent jugée à partir d'indicateurs purement statistiques telle que l'aire sous la courbe ROC. La validation des aptitudes pronostiques d'un marqueur me semble effectivement être un préalable important à son utilisation en pratique clinique. De bonnes propriétés de discrimination et de calibration permettent d'évaluer si le marqueur à l'étude réussit correctement à séparer les sujets à faible et à fort risque de connaître l'événement (i.e. discrimination) et s'il réussit à bien estimer le risque de survenue d'un événement (i.e. calibration). Pour autant, il me paraît difficilement raisonnable de s'arrêter à cette simple *validation* pour juger de son *utilité*. La question de la discrimination d'une population pour mettre en place une médecine stratifiée à partir d'un marqueur quantitatif ne peut se poser que s'il existe une interaction entre le traitement d'intérêt et le marqueur à l'étude. Sans interaction entre ces deux variables, l'utilisation d'un marqueur pronostique pour une prise en charge stratifiée apparaît d'emblée inutile puisque l'ensemble des sujets auront intérêt à recevoir l'un ou l'autre des traitements indépendamment de la valeur du marqueur. Par ailleurs, nous nous intéressons à la notion d'utilité clinique d'une décision médicale en terme de QALYs espérés. Même si notre approche présente des limites, comme le fait de ne considérer que des niveaux d'utilités aggrégés par bras de prise en charge et non des utilités individuelles, je crois que ce travail contribue à une meilleure considération de l'utilité clinique d'un marqueur pronostique en intégrant les conséquences d'une prise en charge stratifiée en terme de gain d'espérance de vie ainsi que les conséquences en terme de qualité de vie.

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–13
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280216671161
smm.sagepub.com
SAGE

Optimal threshold estimator of a prognostic marker by maximizing a time-dependent expected utility function for a patient-centered stratified medicine

Etienne Dantan,[1] Yohann Foucher,[1,2] Marine Lorent,[1] Magali Giral[3,4] and Philippe Tessier[1]

Abstract

Defining thresholds of prognostic markers is essential for stratified medicine. Such thresholds are mostly estimated from purely statistical measures regardless of patient preferences potentially leading to unacceptable medical decisions. Quality-Adjusted Life-Years are a widely used preferences-based measure of health outcomes. We develop a time-dependent Quality-Adjusted Life-Years-based expected utility function for censored data that should be maximized to estimate an optimal threshold. We performed a simulation study to compare estimated thresholds when using the proposed expected utility approach and purely statistical estimators. Two applications illustrate the usefulness of the proposed methodology which was implemented in the R package *ROCt* (www.divat.fr). First, by reanalysing data of a randomized clinical trial comparing the efficacy of prednisone vs. placebo in patients with chronic liver cirrhosis, we demonstrate the utility of treating patients with a prothrombin level higher than 89%. Second, we reanalyze the data of an observational cohort of kidney transplant recipients: we conclude to the uselessness of the Kidney Transplant Failure Score to adapt the frequency of clinical visits. Applying such a patient-centered methodology may improve future transfer of novel prognostic scoring systems or markers in clinical practice.

Keywords

Prognostic marker, threshold definition, QALY, censored data, stratified medicine, patient preferences

1 Introduction

Stratified medicine currently relies on the use of biological markers capable of discriminating patients into sub-groups receiving the most appropriate treatment.[1,2] The recent opportunities toward this personalized patient care, especially offered by omics technologies, have transformed biomedical research.[3] Despite this, a very low proportion of discovered biomarkers or scoring systems has been transferred to clinical practice. This disillusion is partially explained by the numerous methodological pitfalls, leading to low credibility of the results,[4,5] and by the non-consideration of the patients themselves: studies are often isolated from their social or environmental context,[6] and independent of their preferences over the possible outcomes of individual-tailored clinical choices. In the literature, the potential usefulness of prognostic markers is often judged regarding purely statistical measures such as the highest area under the ROC curve (AUC). This indicator is widely associated with the following normative premise: the higher the AUC, the lower the error rates of a decision based on the marker,

[1]EA 4275 SPHERE – methodS in Patient-centered outcomes & HEalth ResEarch, Nantes University, France
[2]Nantes University Hospital, Nantes, France
[3]Transplantation, Urology and Nephrology Institute (ITUN), CHU Nantes, RTRS Centaure, Nantes and Inserm U1064, Labex Transplantex, Nantes University, France
[4]Centre d'Investigation Clinique biothérapie, Nantes, France

Corresponding author:
Etienne Dantan, EA 4275 SPHERE – methodS in Patient-centered outcomes & HEalth ResEarch, Nantes University, 22 boulevard Bénoni Goullin, 44200 Nantes, France.
Email: Etienne.Dantan@univ-nantes.fr

Downloaded from smm.sagepub.com at INSERM on October 26, 2016

and the higher the corresponding usefulness in future medical decision-making. Several common approaches[7,8] can be envisaged to estimate a marker threshold in order to discriminate between low and high risk patients, as the Youden index corresponding to the marker value that maximizes the sum of sensitivity and specificity.

However, as recently recalled by Subtil et al.,[9] these purely statistical approaches had already been clearly criticized in the literature since they ignored patient preferences of the health outcome consequences of the stratified medical decision-making, possibly leading to unacceptable trade-offs between the harms and the benefits of a clinical decision.[10,11] From a decision analysis framework, several theoretical-based papers have shown how a clinically useful threshold for a diagnostic test can be determined by maximizing an expected utility function, utility referring to the intensity of individual preferences over health outcomes.[9,12–15] Even if the expected utility maximization in order to estimate optimal threshold on a continuous marker appears relatively well known in diagnostic context, these decision analytic approaches remain difficult to extrapolate to time-dependent prognostic frameworks that imply dealing with right censoring data and time-dependent utility measures for health outcomes. Fortunately, the former point can be treated by considering survival models and the latter point can be handled using Quality-Adjusted Life Years (QALYs). QALYs are a widely used composite measure of the health care consequences that combines in a single number information about the quantity and the health-related quality of life (QoL).[16] Although QALYs have been primarily designed for the purpose of economic evaluation, their potential usefulness for clinical decision-making has been acknowledged and warrants further consideration.[17–19] More precisely, QALYs represent any sequence of health states over time as an equivalent number of years lived in perfect health. Each life year is weighted by preference-based measures, called utility scores. The utility scores usually range from 0 (dead) to 1 (perfect health), so that a higher score indicates a more preferred health state. For instance, a patient alive 10 years with a utility of 0.8 will have 8 QALYs (10×0.8), a value higher than the one for a patient alive 12 years with a utility of 0.6, this last one having 7.2 QALYs (12×0.6) due to a more efficient intervention but with more side effects. The utility scores can be assessed directly by asking patients. For this purpose, different methods exist, such as the standard gamble or the time trade-off methods for instance.[17] The utility scores can also be obtained indirectly from pre-scored generic health-states description systems, such as the Euroqol EQ-5D[18] or the SF-6D,[19] that represent the preferences of the general population. Another possibility is to gather utility scores through literature review.[20] Given the widespread use of QALYs, published scores are available for a wide variety of health conditions.

The aim of this paper is to extend expected utility maximization to prognostic frameworks in order to estimate prognostic marker thresholds to achieve patient-centered stratified medicine. We performed a simulation study to highlight the interest of considering the individual preferences comparatively to purely statistical approaches. We illustrate the usefulness of the methodology by assessing two different clinical decision contexts. The first concerns reanalysis of an old clinical trial aimed at assessing prednisone prescription in patients suffering from chronic liver cirrhosis based on the prothrombin level.[21] The second relies on an observational cohort aiming to use the Kidney Transplant Failure Score (KTFS) to adapt the follow-up of kidney transplant recipients after one year post-transplantation.[22]

2 Methods

2.1 Definition of the time-dependent expected utility function for a stratified medical decision

Let a sample be constituted by n patients. Let T, C, Z, and X be, respectively, the time to the time-to-failure, the time-to-censoring, the treatment (A or B) and the baseline prognostic marker under investigation to potentially drive the treatment allocation. For each subject i ($i = 1, \ldots, n$), we observe $\{y_i, z_i, x_i\}$ with $Y = \min(T, C)$. Let $D(\tau)$ be the indicator of failure such that $D(\tau) = 1$ if it occurs before the forecast horizon time τ, and $D(\tau) = 0$ otherwise. By convention, assume that the incremental benefit of the treatment A compared to the treatment B in terms of delayed failure increases with X at the price of possible side effects deteriorating QoL (again compared to the treatment B). The aim is to define a threshold κ for the following decision: proposition of the treatment A for a patient i when $\{x_i > \kappa\}$ and B otherwise. Patient profiles may therefore be distinguished by combining their two possible initial strata ($X > \kappa$ or $X \leq \kappa$) and their two possible outcomes ($D(\tau) = 1$ or $D(\tau) = 0$). The stratified medical decision leads to four possible health outcomes that each correspond to specific flows of time-dependent health states or health-related QoL levels over the prognostic period. Adapting the initial proposition of Metz,[13] the expected utility function corresponds to a sum over all four health outcomes weighted by their corresponding utility. In a prognostic framework, taking into account the patients' preferences over health states, the optimal threshold κ^* can be estimated by maximizing the following

Downloaded from smm.sagepub.com at INSERM on October 26, 2016

time-dependent expected utility function

$$\psi_\tau(\kappa) = \sum_{g\in\{X>\kappa, X\le\kappa\}} \sum_{j\in\{0,1\}} P(g, D(\tau)=j) Q(\tau|\, g, D(\tau)=j) \tag{1}$$

where $Q(\tau|\, g, D(\tau)=j)$ represents the numbers of QALYs up to the prognostic time τ corresponding to each possible profile and $P(g, D(\tau)=j)$ are the associated probabilities. The number of QALYs is defined as the sum of life years weighted by the instantaneous utility ranging from 0 (being dead) to 1 (being in perfect health) of the corresponding health states,[23] i.e. a mean of the health-state utilities weighted by the corresponding probabilities up to time τ

$$Q(\tau|\, g, D(\tau)=j) = \sum_{l\in\{0,1\}} \int_0^\tau u(t|\, g, D(t)=l) P(D(t)=l\,|\, g, D(\tau)=j)\, \mathrm{d}t \tag{2}$$

where $u(t|\, g, D(t)=l)$ is the instantaneous utility defined on the scale 0–1 (death-perfect health) of the health state at time t. We assumed, as the QALY model does, that the instantaneous utility scores at time t for health states are constant over time, i.e. $u(t|\, g, D(t)=l) = u_{g,l}$ for all t. If $g=\{X>\kappa\}$, the utility of receiving treatment A is considered ($u_{X>\kappa,l} = u_{A,l}$); otherwise, the utility of receiving B ($u_{X\le\kappa,l} = u_{B,l}$). It can be shown that the time-dependent expected utility function (equation (1)) may then be simplified as follows (see Appendix 1)

$$\psi_\tau(\kappa) = \sum_{g\in\{X>\kappa, X\le\kappa\}} P(g)\left[u_{g,0} E(\min(T,\tau)|\, g) + u_{g,1}(\tau - E(\min(T,\tau)|\, g))\right] \tag{3}$$

where $E(\min(T,\tau)|\, g)$ is the restricted mean survival time (RMST) up to time τ in the group g, i.e. the average survival time when patients are followed up to τ.[24] Because we assumed that, compared to the treatment B, the treatment A is associated to a better efficacy in delaying the failure for high values of the marker X and a lower QoL due to side-effects, we also have to respect the condition $\left\{u_{X>\kappa,l} \le u_{X\le\kappa,l}\right\}$ for all $l\in\{0,1\}$. Because we considered a time-to-failure, we also have $\left\{u_{g,0} > u_{g,1}\right\}$ for all $g\in\{X>\kappa, X\le\kappa\}$.

For a patient-centered medical decision, the time τ should be determined as the most important value with sufficient at-risk patients to allow a reliable statistical inference. Patient expectations are both to delay the event as late as possible and to maintain their QoL as long as possible. Note also that the value of τ enters in equation (3), therefore the choice of τ affects the corresponding number of QALYs, resulting in the non-consideration of the decision consequences beyond τ. We, therefore, recommend to define τ as the highest possible value.

Note that the maximization of the time-dependent expected utility function (equation (3)) may result in an estimation of κ^* equal to an extremum of the marker. This implies treating all patients with A if $\kappa^* = \min(X)$, or with B if $\kappa^* = \max(X)$. These situations indicate the futility of the prognostic marker for patient-centered stratified medical decision-making, even if its prognostic capacities are important.

2.2 Non-parametric estimation of the time-dependent expected utility function

We used a non-parametric approach in order to maximize the expected utility function leading to the optimal threshold κ^*. As demonstrated, the expected utility function can be decomposed into three terms (equation (3)). First, the probabilities $P(g)$ for $g\in\{X>\kappa, X\le\kappa\}$ can be estimated by the empirical cumulative distribution function of X in g, i.e. $P(g) = n^{-1}\sum_{i=1}^n 1(X_i\in g)$, with $1(\,)$ the indicator function equal to 1 if the condition within the brackets is verified and 0 otherwise. Second, the utility scores $u_{g,l}$, for $g\in\{X>\kappa, X\le\kappa\}$ and $l\in\{0,1\}$, can be obtained from individual preference scores already published in the literature. Third, $E(\min(T,\tau)|\, g)$ can be estimated by the area under the Kaplan–Meier survival curve in patients of the strata g with the adequate treatment, noticed $\hat{S}(\cdot|g)$, up to time τ. More precisely, among patients of the strata g, let $t_{1,g} < \ldots < t_{q,g} < \ldots < t_{p,g}$ be the different event time, $d_{q,g}$ the number of subjects experiencing the event at time $t_{q,g}$ and $R_{q,g}$ the number of at-risk patients at the same time. Corresponding to the average survival time when patients are followed up to τ, the RMST can therefore be written as

$$E(\min(T,\tau)|g) = \int_0^\tau S(t|g)\mathrm{d}t = \sum_{m:t_{m,g}\le\tau} \left(t_{m,g} - t_{m-1,g}\right) \times \hat{S}\left(t_{m,g}|g\right) = \sum_{m:t_{m,g}\le\tau} \left[\left(t_{m,g} - t_{m-1,g}\right) \times \prod_{q:t_{q,g}\le t_{m,g}} \left(1 - \frac{d_{q,g}}{R_{q,g}}\right)\right]$$

Downloaded from smm.sagepub.com at INSERM on October 26, 2016

Depending on the study design, the RMST may be directly estimated in each strata g. For instance, in a clinical trial, the two treatments are observed: $E(\min(T,\tau)|X>\kappa)$ is then estimated in patients receiving treatment A, i.e. $E(\min(T,\tau)\,|X>\kappa, Z=A)$, while $E(\min(T,\tau)|X\leq\kappa)$ is estimated in patients receiving treatment B, $E(\min(T,\tau)|X\leq\kappa, Z=B)$. However, in some other cases, specific assumptions have to be formulated. For instance, from an observational cohort in which only treatment B is observed, $E(\min(T,\tau)|X>\kappa)$ cannot be directly estimated since treatment A is not observed. In this situation, we have to assume a relative increase ($\boldsymbol{\Delta}$) in the RMST of patients treated by A compared to B, i.e. $E(\min(T,\tau)|X>\kappa) = \min(E(\min(T,\tau)|X>\kappa, Z=B)\times(1+\Delta);\tau)$.

The 95% confidence interval (CI) of the optimal threshold κ^* can be estimated by using non-parametric bootstrap resampling.[25] From the observational data, 2000 independent bootstrap samples are generated and the expected utility function is maximized in order to estimate an optimal threshold for each sample. The corresponding 2.5th and the 97.5th percentiles represent the 95% CI. All analyses were performed using the 3.0.2. version of the R software.[26] This method has been implemented in the package *ROCt* available at www.divat.fr.

3 Simulation study

3.1 Design

Data were simulated according a 1:1 randomized clinical trial design. For each subject i ($i=1,\ldots,1000$), the prognostic marker X_i was obtained from a Gaussian distribution truncated between 20 and 600 of mean 140 and standard deviation 55. The binary variable Z_i related to the treatment A ($Z=1$) or B ($Z=0$) was simulated according to a Bernouilli distribution with probability at 0.5. The time-to-censoring C_i was generated from a Weibull distribution in order to obtain around 10% of censoring at 6 years. The time-to-event T_i was simulated from a proportional hazard model: $\lambda_0(t_i)\exp(\beta_X X_i+\beta_Z Z_i+\beta_{XZ}X_iZ_i)$, with a Weibull distribution for the baseline risk $\lambda_0(t)$. Weibull parameters were chosen in order to have roughly a 30% survival probability at six years of follow-up. The prognostic marker X_i was assumed to be a risk factor with $\beta_X=0.002$.

We considered different scenarios for which we simulated 1000 samples of patients. More precisely, we investigated three treatment effects $\beta_Z=(0,-0.5,-1.25)$ and three interaction levels with the marker $\beta_{XZ}=(0,-0.004,-0.008)$. Moreover, different impacts on QoL were investigated: a similar utility between the two treatments ($u_{A,0}=u_{B,0}=0.7$), a 15% decrease due to treatment A compared to B ($u_{A,0}=0.6; u_{B,0}=0.7$), and a 30% decrease due to treatment A compared to B ($u_{A,0}=0.5; u_{B,0}=0.7$). We considered the death as the event of interest, and we therefore have $u_{A,1}=u_{B,1}=0$. The prognostic time τ was fixed at six years.

We compared three estimators: our proposed approach based on the expected utility maximization, the Youden index, and the profile likelihood maximization of a Cox model including the treatment, the marker, and the corresponding interaction.

3.2 Results

Simulation results of the different scenarios are presented in Table 1. The two purely statistical approaches approximately estimated a threshold round the value 140, i.e. the mean of prognostic marker. As expected, the Youden index and the profile likelihood maximization lead to thresholds insensitive to the assumptions made on the utilities. It may result in unacceptable stratified medical decision-making. For instance, when considering a higher efficacy of treatment A compared to treatment B $(\beta_Z=-1.25)$ with no interaction with the marker ($\beta_{XZ}=0$) and similar utilities ($u_{A,0}=u_{B,0}=0.7$), the Youden index led to a threshold at 142.91 and the profile likelihood maximization at 141.88. It leads to recommend treatment B for about one half of the population, a nonsense regarding such an obvious balance benefit/risk in favour of treatment A. In this situation, maximizing the time-dependent expected utility leads to a threshold at 25.92 for which more than 99% of patients would receive treatment A. This threshold increased to 91.61 when utility for treatment A is 0.6 and utility for treatment B is 0.7, corresponding to treat 20% of patients with treatment A. In contrast, when side effects due to treatment A were too strong compared to treatment B $(u_{A,0}=0.5; u_{B,0}=0.7)$, the threshold was 281.49, corresponding maintain all patients with treatment B.

When the interaction coefficient is negative, the effect of treatment A increases with the marker and should lead to recommend treatment A to more patients. This trend was observed from our QALYs-based expected utility maximization. For example, for a 15% utility decrease $(u_{A,0}=0.6; u_{B,0}=0.7)$ and an intermediate treatment

Downloaded from smm.sagepub.com at INSERM on October 26, 2016

Table 1. Results of the simulation study performed on 1000 samples of 1000 patients.

			$u_A = u_B = 0.7$		$u_A = 0.6$ $u_B = 0.7$		$u_A = 0.5$ $u_B = 0.7$	
Interaction between marker and treatment β_{XZ}	Treatment effect β_Z	Estimator	Mean of estimated κ	SE	Mean of estimated κ	SE	Mean of estimated κ	SE
0	0	QALYs-based expected utility	144.10	75.56	282.13	34.07	304.15	27.31
		Youden index	141.67	25.06	141.67	25.06	141.67	25.06
		Profile likelihood maximization	137.35	36.10	137.35	36.10	137.35	36.10
	−0.5	QALYs-based expected utility	36.71	18.10	243.88	45.73	297.61	29.92
		Youden index	142.39	25.99	142.39	25.99	142.39	25.99
		Profile likelihood maximization	137.78	36.79	137.78	36.79	137.78	36.79
	−1.25	QALYs-based expected utility	25.92	7.76	91.61	45.58	281.49	34.32
		Youden index	142.91	28.06	142.91	28.06	142.91	28.06
		Profile likelihood maximization	141.88	38.97	141.88	38.97	141.88	38.97
−0.004	−0.5	QALYs-based expected utility	31.49	12.51	126.48	39.31	276.78	34.54
		Youden index	141.11	33.98	141.11	33.98	141.11	33.98
		Profile likelihood maximization	145.61	43.40	145.61	43.40	145.61	43.40
	−1.25	QALYs-based expected utility	25.36	7.11	62.67	30.32	264.40	36.90
		Youden index	142.69	33.30	142.69	33.30	142.69	33.30
		Profile likelihood maximization	144.19	45.17	144.19	45.17	144.19	45.17
−0.008	−0.5	QALYs-based expected utility	29.18	10.19	85.57	31.53	260.56	36.86
		Youden index	141.04	34.53	141.04	34.53	141.04	34.53
		Profile likelihood maximization	141.12	38.93	141.12	38.93	141.12	38.93
	−1.25	QALYs-based expected utility	24.89	6.37	52.59	24.44	252.98	38.37
		Youden index	144.02	34.31	144.02	34.31	144.02	34.31
		Profile likelihood maximization	140.93	43.13	140.93	43.13	140.93	43.13

effect $(\beta_Z = -0.5)$, the threshold was 126.48 for a $\beta_{XZ} = -0.004$ and 85.57 for a $\beta_{XZ} = -0.008$. In contrast, purely statistical approaches were not sensitive.

Considering obvious balances benefit–risk, such as the better efficacy of treatment A without more side effects compared to treatment B, or the similar efficacy of treatment A but with more side effects compared to the treatment B. In these situations, our proposed approach gives standard errors (SE) lower than the ones obtained from purely statistical approaches. In contrast, when the balances benefit–risk require an arbitration, SE obtained from our approach are higher than the SE obtained from purely statistical approaches. Therefore, it appears that the SE obtained from our approach better reflects the uncertainty which surround the therapeutic decision, while variability from purely statistical approaches will mainly depend on the sample size.

4 Applications

4.1 Patients with chronic liver cirrhosis

In an randomized clinical trial studying the efficacy of prednisone (treatment A, $n = 226$) to increase the survival of patients with chronic liver cirrhosis,[21] no significant difference was reported compared to patients receiving placebo (treatment B, $n = 220$). Nevertheless, an interaction with the initial prothrombin level, an essential blood clotting glycoprotein, suggested the better efficacy of prednisone in patients with a high prothrombin level.[27] Patients with a poorer liver function, i.e. a low prothrombin level, did not benefit from prednisone as their livers were unable to metabolize the high steroid dose, and therefore may be steroid-poisoned. For the purpose of illustration, we aimed to reanalyze the data of this historic clinical trial to define the prothrombin threshold above which patients may receive prednisone, by taking into account that prednisone is a glucocorticosteroid that may be beneficial to liver function[28] whilst causing side effects such as osteopenia, diabetes mellitus, elevation of arterial blood pressure, psychiatric disorders, glaucoma or serious infections.[29] We chose the prognostic time τ at 8 years, as the maximal follow-up time with sufficient at-risk patients for inference. More precisely, 36 prednisone-treated patients and 32 placebo-treated patients were still at risk.

Downloaded from smm.sagepub.com at INSERM on October 26, 2016

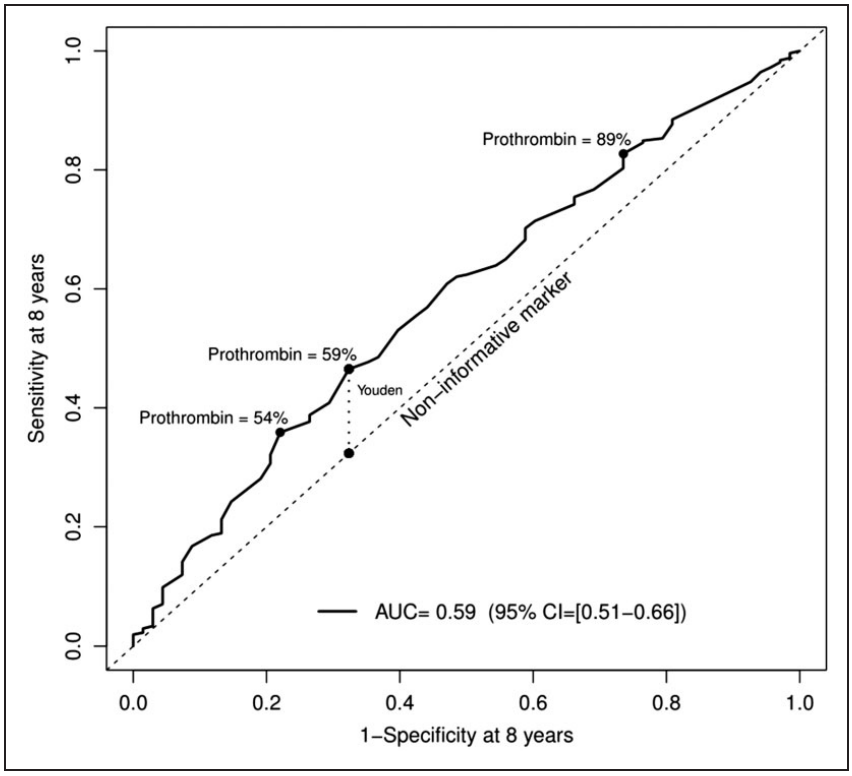


Figure 1. Receiver-Operating Characteristic (ROC) curves for eight-year predictions to evaluate the prognostic capacity of the prothrombin marker among patients with chronic liver cirrhosis ($n = 446$).

The patient survival probabilities at eight years were 37% (95% CI from 30% to 45%) for the prednisone-treated patients and 28% (95% CI from 22% to 36%) for the placebo-treated patients. The median prothrombin level was 68% (range from 12% to 134%). The area under the ROC curve for a prognostic up to eight years (Figure 1) was 0.59 (95% CI from 0.51 to 0.66) reflecting moderate prognostic capacities of prothrombin measurement.

Utilities of the various health states were defined according to the mean Euroqol EQ-5D values published in a systematic review of health-state utilities in liver disease performed by Mc Lernon et al.[30] The EQ-5D is a widely used five-item health-related QoL descriptive system that provides utility scores for 243 possible health states estimated by applying the time-trade-off method over samples of the general population.[31,32] The assessed utility score was 0.75 for patients with compensated cirrhosis. We used this value for patients receiving the placebo, i.e. $u_{B,0} = 0.75$. Compared to placebo, we may reasonably assume that the impact of prednisone on the QoL ranges between a 5% decrease ($u_{A,0} = 0.95 \times u_{B,0} = 0.71$) and a 10% decrease ($u_{A,0} = 0.90 \times u_{B,0} = 0.67$). Because the treatment allocation was performed independently of the prothrombin level, the RMSTs used in equation (3) can be estimated assuming no confounding factors.

The results are presented in Table 2. With a 5% utility decrease due to prednisone, the optimal prothrombin threshold was estimated at 89% (95% CI from 16% to 91%). The corresponding expected utility was 3.54 years in perfect health (QALYs). In comparison, this mean value should have been 3.41 QALYs by treating all the patients with prednisone and 3.32 QALYs by treating all the patients with placebo. In others words, for 100 patients treated by prednisone if their prothrombin level is higher than 89% and by placebo otherwise, the expected gain was 13 years in perfect health ($100 \times (3.54 - 3.41)$) compared to treating the 100 patients by prednisone. Similarly, the expected gain of the stratified medical decision for a prothrombin at 89% was 22 years in perfect health ($100 \times (3.54 - 3.32)$) compared to treating the 100 patients by placebo. As also indicated in the Table 2, one can note that stratifying the medical decision at 89% results in a comparable RMST than treating systematically by prednisone: 4.80 vs. 4.81 years, respectively. Similarly, the difference of RMST when all the patients received the placebo is small: 4.80 vs. 4.42 years. This illustrates the importance to not only consider the decision in terms of

Downloaded from smm.sagepub.com at INSERM on October 26, 2016

Table 2. Results of the application on patients suffering chronic liver cirrhosis for scenarios with 5% and 10% utility decrease due to prednisone treatment compared to placebo.

Utility decrease	Threshold κ^*	Percentage with $X > \kappa^*$	RMST change (years) over misclassified treated patients		QALY change (years) over misclassified treated patients		Mean RMST (in years)	Expected utility (in QALYs)
			$X > \kappa^*$	$X \leq \kappa^*$	$X > \kappa^*$	$X \leq \kappa^*$		
5%	Maximum: 134	0%	–	–	–	–	4.42	3.32
	QALYs-based expected utility: 89[16–91][a]	22%	1.67	−0.01	0.99	0.17	4.80	3.54
	Youden index: 59 43–92][a]	62%	0.53	−0.23	0.18	−0.02	4.74	3.42
	Profile likelihood maximisation: 54 [36–92][a]	72%	0.36	−0.47	0.06	−0.21	4.69	3.37
	Minimum: 12	100%	–	–	–	–	4.81	3.41
10%	Maximum: 134	0%	–	–	–	–	4.42	3.32
	QALYs-based expected utility: 89 [13–99][a]	22%	1.67	−0.01	0.74	0.34	4.80	3.48
	Youden index: 59 [43–92][a]	62%	0.53	−0.23	−0.04	0.13	4.74	3.29
	Profile likelihood maximisation: 54 [36–92][a]	72%	0.36	−0.47	−0.16	−0.07	4.69	3.21
	Minimum: 12	100%	–	–	–	–	4.81	3.22

[a]CI 95% calculated from 2000 bootstrap samples.

efficacy, which may not be so pertinent than using QALYs. Twenty-two percent of patients had a prothrombin higher than 89%. In this subgroup and for patients followed up to eight years, we estimated at 1.7 years the mean increase of time-to-death attributable to prednisone (vs. placebo). This was equivalent to a one-year survival gain in perfect health. In patients with a prothrombin level lower than 89%, the placebo resulted in a similar RMST compared to prednisone. Nevertheless, this equivalent RMST corresponded to an increase of 0.17 QALY due to the side effects of prednisone. One can note that the results were similar with a 10% decrease in the utility due to prednisone treatment. The prothrombin threshold is still estimated at 89% (95% CI from 36% to 99%). Additionally, by exploring others' prognostic times τ from three years to eight years, the optimal threshold appeared stable, ranging from 87% to 89%.

From purely statistical approaches, we retained different thresholds (Table 2 and Figure 1). Using the Youden index for instance, the retained threshold of the prognostic marker was 59% (95% CI from 43% to 92%), corresponding to 62% of patients that should be treated by prednisone in a future-stratified medical decision-making situation. The corresponding expected utility was 3.42 QALYs for a 5% utility decrease. In other words, for 100 patients treated by prednisone, if their prothrombin level is higher than 59% and by placebo otherwise, the expected loss was 12 years in perfect health ($100 \times (3.42 - 3.54)$) compared to the same decision rules with the threshold at 89% we previously reported by using our proposed approach. By a profile likelihood maximization of a Cox model including the treatment, the prothrombin level (1 if the level is higher than κ and 0 otherwise) and the corresponding interaction, we estimated a prothrombin threshold at 54% (95% CI from 36% to 90%): 72% of patients who should be treated by prednisone. The corresponding expected utility was 3.37 QALYs for a 5% utility decrease, which represents an expected loss for 100 patients equals to 17 years in perfect health ($100 \times (3.37 - 3.54)$) compared to the same decision rules with the threshold at 89%.

4.2 Kidney transplant recipients

Obtained from the observational DIVAT cohort, the Kidney Transplantation Failure Score (KTFS) is a score calculated at one year post-transplantation and composed of eight clinical parameters aiming to predict returns to dialysis within the eight years post-transplantation.[22] The authors proposed a threshold at 4.17 by maximizing both the time-dependent sensitivity and specificity,[33] a method close to the Youden index maximization. A randomized clinical trial is currently in progress in order to determine whether the efficiency of the transplantation could be improved by adapting the recipient follow-up, in particular with a higher frequency of visits for patients stratified at high-risk, i.e. with KTFS higher than 4.17.[34]

Downloaded from smm.sagepub.com at INSERM on October 26, 2016

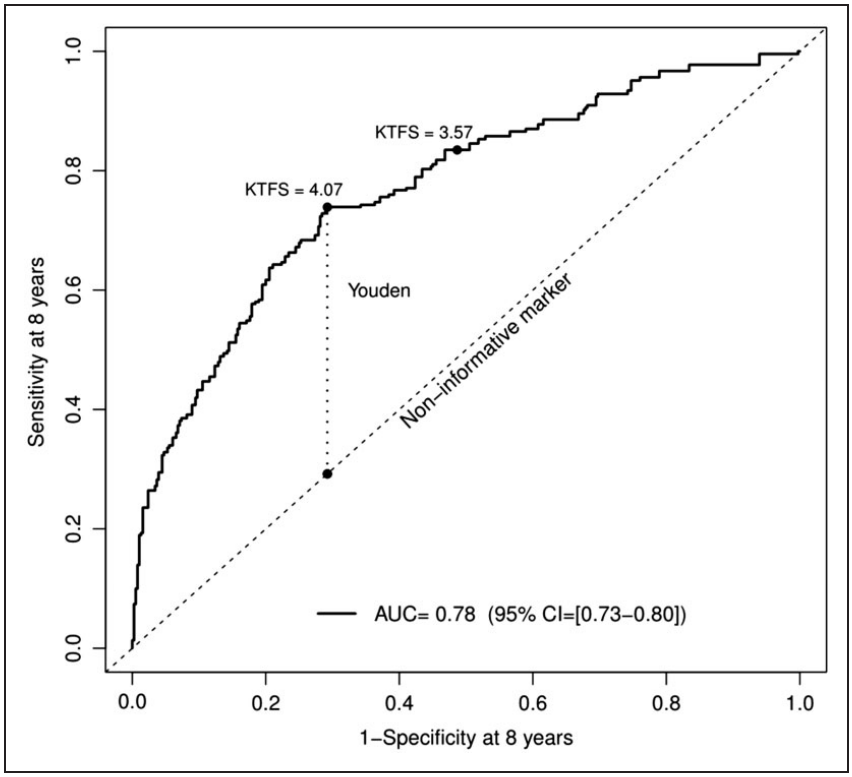


Figure 2. Receiver-Operating Characteristic (ROC) curves for eight-year predictions to evaluate the prognostic capacity of the KTFS among kidney transplant recipients ($n = 2169$).

We reanalyzed the training sample used in the initial paper[22] with the same prognostic time τ at eight years. In this cohort, all patients received standard care in terms of visit frequency (treatment B). We wondered whether some patients could benefit from a more intensive follow-up (treatment A) regarding their higher risk of return to dialysis, even if one can also expect a decrease of recipients' QoL due to repeated in-hospital visits with negative consequences such as anxiety, depression and distress that have been documented in some diseases.[35] This sample was composed of 2169 kidney transplant patients, 182 had returned to dialysis at the end of the follow-up. The graft survival probabilities at four and eight years post-transplantation were, respectively, 95% (95% CI from 93% to 96%) and 85% (95% CI from 83% to 88%). The median KTFS was equal to 3.73 (range from 1.23 to 15.33). Prognostic capacities at eight years appeared good with an area under the ROC curve (Figure 2) estimated at 0.78 (95% CI from 0.73 to 0.80).

The mean EQ-5D utility score for a functional kidney transplant was estimated at 0.81 in a meta-analysis by Liem et al.[36] We found only one study reporting the utility after a return in dialysis,[37] which was estimated at 65% of the utility of having a functional transplant. More formally, we have: $u_{B,0} = 0.81$ and $u_{A,1} = u_{B,1} = 0.65 \times u_{B,0} = 0.53$. We defined several expected QoL decreases related to a higher frequency of visits compared to standard follow-up: (i) 10% ($u_{A,0} = 0.90 \times u_{B,0} = 0.73$), (ii) 5% ($u_{A,0} = 0.95 \times u_{B,0} = 0.77$), and (iii) 1% ($u_{A,0} = 0.99 \times u_{B,0} = 0.80$). We found no published data supporting the gain in RMST related to the follow-up frequency increase, and we therefore hypothesized two possible scenarios: a 5% and a 10% RMST increase compared to standard follow-up. Note that the gain was assumed identical whatever the KTFS threshold.

The results are presented in Table 3. For a 10% decrease in the utility of a transplanted patient, the maximization of the time-dependent expected utility led to maintaining all patients on a standard follow-up regardless the RMST gain ($\kappa^* = 15.33$, the maximum observed value of KTFS). This means that the expected gain of RMST was insufficient to outweigh the consequences due to the increased follow-up frequency in terms of QoL. The corresponding mean RMST and the expected utility were, respectively, 7.54 years and 6.35 QALYs. Similarly, with a 5% decrease in the utility of a transplanted patient and a 10% RMST increase, the optimal

Downloaded from smm.sagepub.com at INSERM on October 26, 2016

Table 3. Results of the application on kidney transplant recipients for scenarios with 1%, 5%, and 10% utility decrease due to prednisone treatment associated with 5% or 10% RMST increase due to a higher consultation frequency.

Utility decrease	RMST increase (%)	Threshold κ^* [Bootstrap 95% CI][a]	Percentage with $X > \kappa^*$	RMST change (years) over observed patients $X > \kappa^*$	QALY change (years) over observed patients $X \leq \kappa^*$	Mean RMST (in years)	Expected utility (in QALYs)
10%	10	15.33 [10.85–15.33]	0%	–	–	7.54	6.35
	5	15.33 [11.59–15.33]	0%	–	–	7.54	6.35
5%	10	9.34 [7.45–15.33]	1%	0.39	−0.06	7.54	6.35
	5	10.31 [8.36–15.33]	1%	0.18	−0.10	7.56	6.35
1%	10	3.57 [2.96–3.80]	57%	0.73	0.14	7.95	6.43
	5	1.23 [1.23–2.03]	100%	0.38	0.04	7.91	6.39
10%	10	Youden index: 4.07 [4.05–4.43]	68%	0.70	−0.43	7.80	6.19
	5			0.35	−0.50	7.67	6.16
5%	10			0.70	−0.12	7.80	6.30
	5			0.35	−0.20	7.67	6.27
1%	10			0.70	0.13	7.80	6.40
	5			0.35	0.04	7.67	6.36

[a]CI 95% calculated from 2000 bootstrap samples.

threshold was estimated at 9.34 (95% CI from 7.45 to 15.33) with only 1% of patients having a higher KTFS. Only a very small (but unrealistic) 1% decrease in the utility due to a higher visit frequency could lead to intermediate optimal thresholds. The corresponding mean RMST and the expected utility were, respectively, 7.95 years and 6.43 QALYs. In addition, we also explored optimal threshold estimation for different prognostic times τ from two years to eight years post-transplantation. Our conclusions were similar: only a 1% utility decrease allows to estimate discriminating thresholds (data not shown). Note that using the Youden index, the threshold was 4.07 (95% CI from 4.05 to 4.43) (Figure 2), close to the threshold 4.17 initially proposed by Foucher et al.,[22] leading to 68% of patients with a higher KTFS. But the Youden index does not lead to the threshold that best represent the patients' preferences. For instance, assuming a 10% utility decrease and a 5% RMST increase, for 100 kidney recipients for which a stratified medical decision is made given a KTFS threshold at 4.07, the expected loss was 19 years in perfect health ($100 \times (6.16 - 6.35)$) compared to propose systematically the standard follow-up.

5 Discussion

The use of prognostic markers for the development of stratified medicine is conditioned upon the determination of clinically relevant thresholds allowing to discriminate patients according to whether they could benefit or not from a treatment. In agreement with recent commentaries on personalized medicine,[6] we believe that the clinical utility of such prognostic markers should be evaluated in a patient-centered framework,[38] taking into consideration the ultimate objective of biomedical studies: improving patient well-being.

We proposed a threshold estimator by maximizing a time-dependent QALYs-based expected utility function. This approach has several interesting features. First, it takes into account both a marker's prognostic capacities in the presence of censored data and individual preferences over health outcomes. Second, our proposal may prove simple and directly applicable to various medical contexts, as illustrated by the applications we provided. It avoids having to directly assess patients' preferences, which is often considered a difficult and costly task,[39] since QALYs can be estimated from indirect information about patient preferences, available in the literature for a wide variety of health states. Third, as showed in the applications as well as in the simulation study, the maximization of the proposed time-dependent expected utility function lead to estimating thresholds that are potentially different from those estimated using purely statistical methods, which questions their clinical relevance. Lastly, unlike statistical approaches, our method allows to take into account the consequences of therapeutic uncertainty regarding the possible outcomes of a stratified medicine program.

Subtil et al.[9] recalled that expected utility function for stratified decision making are related to decision curves developed by Vickers et al.[39,40] and initially proposed to compare predictive models. Decision Curves Analysis (DCA) does not directly compare to our approach since it does not aim at determining an optimal threshold for

Downloaded from smm.sagepub.com at INSERM on October 26, 2016

the marker. DCA does not require explicitly patients' preferences elicitation since it assumes that physicians have knowledge of threshold probability of a disease at which a patient would opt for treatment instead of doing nothing. This threshold probability is associated to the ratio of net benefit of treating a diseased patient to net cost of treating a non-diseased one. In this spirit, Foucher et al.[41] asked physicians for their estimation of the probability threshold for an intensive follow-up program of renal transplant patients. However, physicians may find it difficult to assess this relative weight and their assessments may not be in accordance with patients' preferences.[42,43] By contrast, our approach needs neither to assume a prior knowledge of patients' preferences (patient's threshold probability) nor to ask physicians about these preferences but uses a well-established and widely used measure of health outcomes.

Besides, the two applications highlight that high prognostic capacities of a marker, summarized for instance through a high area under the ROC curve, is not sufficient to demonstrate its clinical utility for patient-centered stratified medicine. Actually, the KTFS appeared useless at driving the frequency of recipient follow-up, although the corresponding AUC at 0.78 (95% CI from 0.73 to 0.80) can be considered as elevated. Conversely, in the chronic liver cirrhosis application, the ROC curve related to the prothrombin level leads to a moderate AUC at 0.59 (95% CI from 0.51 to 0.66) for a prognostic at eight years. Nevertheless, a stratified medical program based on the prothrombin threshold may increase patient well-being. Note that the interaction between the marker of interest and the treatment appears as important as the prognostic capacities of the marker and may influence the threshold estimation.

Our proposal has nevertheless some limitations that are worth mentioning. First, it has been argued that regret minimization offers a more accurate description of physicians' behavior than expected utility.[44,45] However, we are concerned here by prescriptive medical decision-making and expected utility maximization is still thought of as the best theory to recommend optimal decisions.[46] Second, implementing our approach requires making assumptions about the potential consequences of stratified medicine. In some ways, this is similar to the exercise made in randomized controlled trial planning to estimate a sample size. Interestingly, Royston et al. recently proposed the use of the difference in RMST to design and analyze clinical trials.[47] The anticipation of potential future health outcomes is intrinsic to the exercise of stratified medicine. Third, our time-dependent expected utility (equation (3)) does not discount future QALYs, in contrast to the general formulation (equation (1)). In a normative perspective, the choice of the discount rate and the legitimacy of discounting future health effects are controversial.[29] Adding a discount factor would be straightforward and would not change the central message of the paper. Finally, when the two treatments are observed, we did not develop the issue related to potential confounders. Based on the recent development we published on adjusted survival curves,[48] we are currently developing a confounder-adjusted time-dependent expected utility.

To conclude, we have proposed a decision analytic method to determine optimal thresholds of prognostic markers based on the maximization of a time-dependent expected utility function, allowing to overcome the limitation of purely statistical approaches. The package *ROCt* in R has been updated to make this methodology simple for users and available for future applications. Applying such a patient-centered methodology may improve future transfer of novel prognostic scoring systems or markers in clinical practice.

Acknowledgements

We thank the DIVAT scientific council and the members of the clinical research assistant team (M. Kessler, M. Ladrière, JP. Soulillou, C. Legendre, H. Kreis, G. Mourad, V. Garrigue, L. Rostaing, N. Kamar, E. Morelon, F. Buron, S. Le Floch, C. Scellier, V. Eschbach, P. Przednowed, K. Zurbonsen, V. Godel, X. Longy, C. Dagot, and F. M'Raiagh) for the collection of the data in the DIVAT cohort.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the French National Research Agency (grant no. ANR-11-JSV1-0008-01, 2011) and the French National Cancer Institute (grant no. INCA 2013-137, 2013).

Downloaded from smm.sagepub.com at INSERM on October 26, 2016

References

1. Baker SG and Sargent DJ. Designing a randomized clinical trial to evaluate personalized medicine: a new approach based on risk prediction. *J Natl Cancer Inst* 2010; **102**: 1756–1759.
2. Steffen J and Steffen J. Driving forces behind the past and future emergence of personalized medicine. *J Pers Med* 2013; **3**: 14–22.
3. Ioannidis JPA. Expectations, validity, and reality in omics. *J Clin Epidemiol* 2010; **63**: 945–949.
4. Ioannidis JPA. Evolution and translation of research findings: from bench to where? *PLoS Clin Trials* [Internet]. 17 November 2006 [cited 3 December 2015]; **1**. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1851723/.
5. Iglesias AI, Mihaescu R, Ioannidis JPA, et al. Scientific reporting is suboptimal for aspects that characterize genetic risk prediction studies: a review of published articles based on the Genetic RIsk Prediction Studies statement. *J Clin Epidemiol* 2014; **67**: 487–499.
6. Joyner MJ and Paneth N. SEven questions for personalized medicine. *JAMA* 2015; **314**: 999–1000.
7. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; **3**: 32–35.
8. Böhning D, Holling H and Patilea V. A limitation of the diagnostic-odds ratio in determining an optimal cut-off value for a continuous diagnostic test. *Stat Meth Med Res* 2011; **20**: 541–550.
9. Subtil F and Rabilloud M. An enhancement of ROC curves made them clinically relevant for diagnostic-test comparison and optimal-threshold determination. *J Clin Epidemiol* 2015; **68**: 752–759.
10. Baker SG, Cook NR, Vickers A, et al. Using relative utility curves to evaluate risk prediction. *J R Stat Soc Ser A Stat Soc* 2009; **172**: 729–748.
11. Baker SG and Kramer BS. Peirce, Youden, and receiver operating characteristic curves. *Am Stat* 2007; **61**: 343–346.
12. Pauker SG and Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med* 1980; **302**: 1109–1117.
13. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978; **8**: 283–298.
14. Irwin RJ and Irwin TC. A principled approach to setting optimal diagnostic thresholds: where ROC and indifference curves meet. *Eur J Intern Med* 2011; **22**: 230–234.
15. Cantor SB, Sun CC, Tortolero-Luna G, et al. A comparison of C/B ratios from studies using receiver operating characteristic curve analysis. *J Clin Epidemiol* 1999; **52**: 885–892.
16. Weinstein MC, Torrance G and McGuire A. QALYs: the basics. *Value Health J Int Soc Pharmacoeconomics Outcomes Res* 2009; **12**: S5–S9.
17. Kind P, Lafata JE, Matuszewski K, et al. The use of QALYs in clinical and patient decision-making: issues and prospects. *Value Health J Int Soc Pharmacoecon Outcome Res* 2009; **12**: S27–S30.
18. Towse A. Net clinical benefit: the art and science of jointly estimating benefits and risks of medical treatment. *Value Health J Int Soc Pharmacoecon Outcome Res* 2010; **13**: S30–S32.
19. Ferguson ND, Scales DC, Pinto R, et al. Integrating mortality and morbidity outcomes. *Am J Respir Crit Care Med* 2013; **187**: 256–261.
20. Torrance GW. Measurement of health state utilities for economic appraisal. *J Health Econ* 1986; **5**: 1–30.
21. EuroQol Group. EuroQol – a new facility for the measurement of health-related quality of life. *Health Policy Amst Neth* 1990; **16**: 199–208.
22. Brazier J, Roberts J and Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ* 2002; **21**: 271–292.
23. Ara R and Wailoo A. Using health state utility values in models exploring the cost-effectiveness of health technologies. *Value Health* 2012; **15**: 971–974.
24. Schlichting P, Christensen E, Andersen PK, et al. Prognostic factors in cirrhosis identified by Cox's regression model. *Hepatol Baltim Md* 1983; **3**: 889–895.
25. Foucher Y, Daguin P, Akl A, et al. A clinical scoring system highly predictive of long-term kidney graft survival. *Kidney Int* 2010; **78**: 1288–1294.
26. Weinstein MC, Torrance G and McGuire A. QALYs: the basics. *Value Health J Int Soc Pharmacoecon Outcome Res* 2009; **12**: S5–S9.
27. Andersen PK and Perme MP. Pseudo-observations in survival analysis. *Stat Meth Med Res* 2010; **19**: 71–99.
28. Efron B. Bootstrap methods: another look at the Jacknife. *Ann Stat* 1979; **7**: 1–26.
29. R Foundation for Statistical Computing. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria, 2011. ISBN 3-900051-07-0. Available at: http://www.R-project.org/.
30. Christensen E, Schlichting P, Andersen PK, et al. Updating prognosis and therapeutic effect evaluation in cirrhosis with Cox's multiple regression model for time-dependent variables. *Scand J Gastroenterol* 1986; **21**: 163–174.
31. Mitchison HC, Palmer JM, Bassendine MF, et al. A controlled trial of prednisolone treatment in primary biliary cirrhosis. Three-year results. *J Hepatol* 1992; **15**: 336–344.
32. Swartz SL and Dluhy RG. Corticosteroids: clinical pharmacology and therapeutic use. *Drugs* 1978; **16**: 238–255.
33. McLernon DJ, Dillon J and Donnan PT. Health-state utilities in liver disease: a systematic review. *Med Decis Mak Int J Soc Med Decis Mak* 2008; **28**: 582–592.

Downloaded from smm.sagepub.com at INSERM on October 26, 2016

34. Dolan P, Gudex C, Kind P, et al. The time trade-off method: results from a general population study. *Health Econ* 1996; **5**: 141–154.
35. Chevalier J and de Pouvourville G. Valuing EQ-5D using time trade-off in France. *Eur J Health Econ HEPAC Health Econ Prev Care* 2013; **14**: 57–66.
36. Heagerty PJ, Lumley T and Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000; **56**: 337–344.
37. Foucher Y, Meurette A, Daguin P, et al. A personalized follow-up of kidney transplant recipients using video conferencing based on a 1-year scoring system predictive of long term graft failure (TELEGRAFT study): protocol for a randomized controlled trial. *BMC Nephrol* 2015; **16**: 6.
38. Kew FM, Galaal K and Manderville H. Patients' views of follow-up after treatment for gynaecological cancer. *J Obstet Gynaecol* 2009; **29**: 135–142.
39. Liem YS, Bosch JL and Hunink MGM. Preference-based quality of life of patients on renal replacement therapy: a systematic review and meta-analysis. *Value Health J Int Soc Pharmacoecon Outcome Res* 2008; **11**: 733–741.
40. Girardi V, Schaedeli F, Marti H-P, et al. The willingness of patients to accept an additional mortality risk in order to improve renal graft survival. *Kidney Int* 2004; **66**: 375–382.
41. Walter SD, Turner R, Macaskill P, et al. Beyond the treatment effect: Evaluating the effects of patient preferences in randomised trials. *Stat Meth Med Res* 2014; **0** 1–19.
42. Vickers AJ and Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006; **26**: 565–574.
43. Steyerberg EW and Vickers AJ. Decision curve analysis: a discussion. *Med Decis Making* 2008; **28**: 146–149.
44. Foucher Y, Giral M, Soulillou J-P, et al. Cut-off estimation and medical decision making based on a continuous prognostic factor: the prediction of kidney graft failure. *Int J Biostat* 2012; **8**: 1–13.
45. Sonnenberg A. Patient–physician discordance about benefits and risks in gastroenterology decision-making. *Aliment Pharmacol Ther* 2004; **19**: 1247–1253.
46. Yuan Z, Levitan B, Burton P, et al. Relative importance of benefits and risks associated with antithrombotic therapies for acute coronary syndrome: patient and physician perspectives. *Curr Med Res Opin* 2014; **30**: 1733–1741.
47. Djulbegovic B, Hozo I, Schwartz A, et al. Acceptable regret in medical decision making. *Med Hypotheses* 1999; **53**: 253–259.
48. Tsalatsanis A, Barnes LE, Hozo I, et al. Extensions to regret-based decision curve analysis: an application to hospice referral for terminal patients. *BMC Med Inform Decis Making* 2011; **11**: 77.
49. Wakker PP. Lessons learned by (from?) an economist working in medical decision making. *Med Decis Making* 2008; **28**: 690–698.
50. Royston P and Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol* 2013; **13**: 152.
51. Le Borgne F, Giraudeau B, Querard AH, et al. Comparisons of the performance of different statistical tests for time-to-event analysis with confounding factors: practical illustrations in kidney transplantation. *Stat Med* 2016; **35**: 1103–1116.

Appendix 1

The expected utility $\psi_\tau(\kappa)$, as defined by the equation (1), may be developed as follows by considering the number of QALYs (equation (2))

$$
\begin{aligned}
\psi_\tau(\kappa) &= \sum_{g\in\{X>\kappa, X\le\kappa\}} \sum_{j\in\{0,1\}} P(g, D(\tau)=j) Q(\tau|\, g, D(\tau)=j) \\
&= \sum_{g\in\{X>\kappa, X\le\kappa\}} \left\{ P(g, D(\tau)=0) Q(\tau|\, g, D(\tau)=0) + P(g, D(\tau)=1) Q(\tau|\, g, D(\tau)=1) \right\} \\
&= \sum_{g\in\{X>\kappa, X\le\kappa\}} \left\{ P(g, D(\tau)=0) \left[\int_0^\tau u(t|\, g, D(t)=0) P(D(t)=0|\, g, D(\tau)=0) + u(t|\, g, D(t)=1) P(D(t)=1|\, g, D(\tau)=0) \mathrm{d}t \right] \right. \\
&\quad \left. + P(g, D(\tau)=1) \left[\int_0^\tau u(t|\, g, D(t)=0) P(D(t)=0|\, g, D(\tau)=1) + u(t|\, g, D(t)=1) P(D(t)=1|\, g, D(\tau)=1) \mathrm{d}t \right] \right\} \\
&= \sum_{g\in\{X>\kappa, X\le\kappa\}} \left\{ P(g, D(\tau)=0) \left[\int_0^\tau u(t|\, g, D(t)=0) \mathrm{d}t \right] \right. \\
&\quad \left. + P(g, D(\tau)=1) \left[\int_0^\tau u(t|\, g, D(t)=0) P(D(t)=0|\, g, D(\tau)=1) + u(t|\, g, D(t)=1) P(D(t)=1|\, g, D(\tau)=1) \mathrm{d}t \right] \right\}
\end{aligned}
$$

We assumed a constant utility over time, i.e. $u(t|\, g, D(t)=l) = u_{g,l}$ for all t. Thus, the expected utility $\psi_\tau(\kappa)$ can be written as

Downloaded from smm.sagepub.com at INSERM on October 26, 2016

$$
\begin{aligned}
\psi_\tau(\kappa) &= \sum_{g\in\{X>\kappa, X\le\kappa\}} \left\{ P(g, D(\tau)=0)\left[\int_0^\tau u_{g,0}\mathrm{d}t\right] + P(g, D(\tau)=1)\left[\int_0^\tau u_{g,0}P(D(t)=0|\, g, D(\tau)=1) + u_{g,1}P(D(t)=1|\, g, D(\tau)=1)\mathrm{d}t\right]\right\} \\
&= \sum_{g\in\{X>\kappa, X\le\kappa\}} \left\{ \int_0^\tau u_{g,0}(P(g, D(\tau)=0) + P(g, D(\tau)=1)P(D(t)=0|\, g, D(\tau)=1))\mathrm{d}t \right. \\
&\quad \left. + \int_0^\tau u_{g,1}P(g, D(\tau)=1)P(D(t)=1|\, g, D(\tau)=1)\mathrm{d}t\right\} \\
&= \sum_{g\in\{X>\kappa, X\le\kappa\}} \left\{ u_{g,0}\int_0^\tau (P(g, D(\tau)=0) + P(D(t)=0, g, D(\tau)=1))\mathrm{d}t + u_{g,1}\int_0^\tau P(D(t)=1, g, D(\tau)=1)\mathrm{d}t\right\} \\
&= \sum_{g\in\{X>\kappa, X\le\kappa\}} \left\{ u_{g,0}\int_0^\tau P(D(t)=0, g)\mathrm{d}t + u_{g,1}\int_0^\tau P(D(t)=1, g)\mathrm{d}t\right\} \\
&= \sum_{g\in\{X>\kappa, X\le\kappa\}} \left\{ P(g)\times\left[u_{g,0}\times\int_0^\tau P(D(t)=0|\, g)\mathrm{d}t + u_{g,1}\times\int_0^\tau P(D(t)=1|\, g)\mathrm{d}t\right]\right\} \\
&= \sum_{g\in\{X>\kappa, X\le\kappa\}} \left\{ P(g)\times\left[u_{g,0}\times\int_0^\tau S(t|\, g)\mathrm{d}t + u_{g,1}\times\int_0^\tau 1 - S(t|\, g)\mathrm{d}t\right]\right\} \\
&= \sum_{g\in\{X>\kappa, X\le\kappa\}} \left\{ P(g)\times\left[u_{g,0}\times E(\min(T,\tau)|\, g) + u_{g,1}\times(\tau - E(\min(T,\tau)|\, g))\right]\right\}
\end{aligned}
$$

Downloaded from smm.sagepub.com at INSERM on October 26, 2016

- Chapitre 8 -

Comparaison du risque de survenue d'un échec de greffe entre les receveurs de reins marginaux et standards : une approche par score de propension

Quérard AH, Le Borgne F, Dion A, Giral M, Mourad G, Garrigue V, Rostaing L, Kamar N, Loupy A, Legendre C, Morelon E, Buron F, Foucher Y and Dantan E. Propensity score-based comparison of the graft failure risk between kidney transplant recipients of standard and expanded criteria donor grafts : Toward increasing the pool of marginal donors. *Am J Transplant.* 2018 May ;18(5) :1151-1157.

Préambule

Mes travaux de recherche en biostatistique et en épidémiologie sont systématiquement appliqués à l'étude de pathologies chroniques. J'ai ainsi participé à plusieurs travaux de recherche en épidémiologie de la transplantation rénale. Le travail de Querard et al. [2018] est un reflet d'une expertise biostatistique en inférence causale plus récemment développée au sein de notre groupe de travail. Cette publication illustre également notre intérêt à publier dans des revues de spécialités médicales à l'aide d'approches statistiques récentes qui nous permettent ici une interprétation différente de certains résultats connus en transplantation rénale.

Nous proposons deux estimations de l'effet d'une greffe de donneur ECD comparativement à une greffe de donneur SCD (Standard Criteria Donor) à partir de la cohorte de patients transplantés rénaux DIVAT. Nous validons tout d'abord l'estimation de l'effet conditionnel d'une greffe de donneur ECD ($HR = 1.75, p < 0.0001$) à partir d'un modèle de Cox multivarié classique. Il semblerait qu'il y ait près de 1.75 fois plus de risque d'échec de greffe pour un patient s'il reçoit un rein de donneur ECD que d'un donneur SCD. De manière complémentaire, nous proposons d'estimer également l'effet marginal (ou effet populationnel) de ce type de greffe à l'aide d'un modèle de Cox pondéré sur un score de propension ($HR = 1.34, p = 0.0034$). Dans ce contexte, il apparaît pertinent d'estimer un effet marginal dit *ATT (Average Treatment ef-*

fect for the Treated) et non un effet *ATE (Average Treatment Effect*) [Austin, 2014; Pirracchio et al., 2016]. Un effet ATT estime un effet marginal du traitement dans la population des traités tandis qu'un effet ATE s'intéresse à l'effet marginal du traitement dans la population. Ainsi, il semblerait donc qu'il y ait près de 1.34 fois plus de risque d'échec de greffe pour l'ensemble des patients ayant reçu un rein ECD comparativement au fait qu'ils auraient pu recevoir un rein SCD. Nous estimons toujours un excès de risque d'échec de greffe lié à la greffe ECD mais l'effet marginal s'avère moindre que celui initialement admis par Port et al. [2002]. Cela ouvre une piste de réflexion pour l'extension de la notion de marginalité d'un greffon, ce qui pourrait s'avérer essentiel pour l'élargissement du pool de reins transplantés.

Received: 16 May 2017 | Revised: 19 December 2017 | Accepted: 21 December 2017

DOI: 10.1111/ajt.14651

ORIGINAL ARTICLE

AJT

Propensity score–based comparison of the graft failure risk between kidney transplant recipients of standard and expanded criteria donor grafts: Toward increasing the pool of marginal donors

A. H. Querard[1,2] | F. Le Borgne[2,3] | A. Dion[2] | M. Giral[4,5] | G. Mourad[6] | V. Garrigue[6] | L. Rostaing[7] | N. Kamar[7] | A. Loupy[8] | C. Legendre[8] | E. Morelon[9] | F. Buron[9] | Y. Foucher[2,10] | E. Dantan[2]

[1]Department of Nephrology, Dialysis and Transplantation, Departmental Hospital of Vendée, La Roche -sur- Yon, France

[2]INSERM, UMR 1246 - SPHERE, Nantes University, Tours University, Nantes, France

[3]IDBC/A2com, Pacé, France

[4]Center for Research in Transplantation and Immunology Inserm UMR 1064, Urology and Nephrology Institute, Nantes University Hospital, Nantes, France

[5]Biotherapy Clinical Investigation Center, Labex Transplantex, Nantes, France

[6]Nephrology, Dialysis and Transplantation Department, Lapeyronie University Hospital, Montpellier, France

[7]Nephrology, Dialysis, and Organ Transplantation Department, Rangueil University Hospital and University Paul Sabatier, Toulouse, France

[8]Kidney Transplant Center, Necker University Hospital, APHP, RTRS « Centaure », Paris Descartes and Sorbonne Paris Cité Universities, Paris, France

[9]Nephrology, Transplantation and Clinical Immunology Department, RTRS « Centaure », Edouard Herriot University Hospital, Hospices Civils, Lyon, France

[10]Nantes University Hospital, Nantes, France

Correspondence
Anne-Hélène Querard
Email: anne-helene.querard@chd-vendee.fr

From a prospective and multicentric French cohort, we proposed an external validation study for the expanded criteria donor (ECD), based on 4833 kidney recipients transplanted for the first time between 2000 and 2014. We estimated the subject-specific effect from a multivariable Cox model. We confirmed a 1.75-fold (95% confidence interval [CI] 1.53-2.00, $P < .0001$) increase in graft failure risk if a given patient received an ECD graft compared to a graft from a donor with standard criteria (standard criteria donor [SCD]). Complementarily, we estimated the population-average effect using propensity scores. We estimated a 1.34-fold (95% CI 1.09-1.64, $P = .0049$) increase in graft failure risk among ECD patients receiving an ECD graft compared to receiving a SCD graft. With a 10-year follow-up, it corresponded to a decrease of 8 months of the mean time to graft failure due to ECD transplantation (95% CI 2-14 months). The population-average relative risk due to ECD transplantation and the corresponding absolute effect seem finally not so high. Regarding the increase of quality of life in transplantation, our study constitutes an argument to extend the definition of marginality by considering more grafts at high risk and thereby enlarging the pool of kidney grafts.

KEYWORDS
clinical research/practice, donors and donation: extended criteria, donors and donation: standard criteria, epidemiology, graft survival, kidney transplantation/nephrology, organ procurement and allocation

Abbreviations: 95% CI, 95% confidence interval; ATT, average treatment effect on the treated; CDC, Complement Dependent Cytotoxicity; CIT, cold ischemia time; CVA, cerebrovascular accident; DIVAT, Données Informatisées et VAlidées en Transplantation; DSA, donor-specific antibody; ECD, expanded criteria; ELISA, Enzyme-Linked Immunosorbent Assay; HBP, high blood pressure; HR, hazard ratio; IPW, inverse probability weighted; KDRI, Kidney Donor Risk Index; OPTN, Organ Procurement and Transplantation Network; SCD, standard criteria donor; SD, standard deviation.

This work was supported by the French Biomedicine Agency.

© 2018 The American Society of Transplantation and the American Society of Transplant Surgeons

1 | INTRODUCTION

Kidney transplantation is recognized as the best treatment for end-stage renal disease in terms of morbidity, mortality, and quality of life.[1] Its development and expansion are limited on one hand by donor organ shortfalls, resulting in the search for new donors, and on the other hand by optimizing the use of available grafts. Due to an aging population, waiting lists are increasing, further exacerbating the need to increase successful transplants by extending the pool of available kidneys.[2] One possibility is to consider marginal grafts for kidney transplantation as they may deliver sufficient "renal function" to improve patient well-being.

In 2002, the expanded criteria donor (ECD) was proposed in the United States.[3] The ECD is defined as a donor older than 60 years, or between 50 and 59 years of age with at least 2 other comorbidities: serum creatinine higher than 1.5 mg/dL, a history of high blood pressure, or a cerebrovascular accident (CVA) as the cause of donor death. The ECD classification was proposed to obtain a subject-specific relative risk of graft failure equal to 1.7 compared to the standard criteria donor (SCD). More precisely, such a subject-specific effect can be interpreted as follows: on average, the graft failure risk for a given patient is multiplied by 1.7 if she/he received an ECD graft instead of an SCD graft.

We recently confirmed this subject-specific effect in a meta-analysis[4] based on 5 studies,[5-9] where we also demonstrated the lack of external validation studies for these criteria. Indeed, no European study was selected in this meta-analysis. However, Aubert et al[10] recently confirmed these results in France. More recently, Ma et al[11] described that this effect seemed more important among younger recipients, whereas it could not be significantly shown for older recipients. In kidney transplantation, it is well known that ECD grafts are preferentially attributed to older recipients, resulting in a possible shift of the recipient age distributions between ECD and SCD recipients. A consequence may be the difficulty of interpreting the ECD transplantation effect among young recipients. Nevertheless, no report in the literature has examined this positivity assumption that requires that one can observe both ECD and SCD recipients at every combination of the values of the observed confounders.[12]

From a French prospective and observational cohort of kidney recipients, we proposed an external validation study of the subject-specific effect of ECD transplantation. Beyond this scope, thanks to the recent developments in propensity score–based analyses and more generally in causal inference, the main objective of our study was to propose an alternative and complementary estimation by providing the absolute effect of ECD transplant organs on graft failure risk. More precisely, we compared for the first time the observed graft survival in ECD recipients and the expected graft survival if all ECD recipients had received SCD grafts. To ensure the robustness of this propensity score–based analysis, we paid special attention to the respect of the assumptions necessary to achieve causality.

2 | MATERIALS AND METHODS

2.1 | Study population

All data were extracted from the French, multicentric, observational and prospective DIVAT cohort of transplanted patients (Données Informatisées et VAlidées en Transplantation ; www.divat.fr, CNIL final agreement, decision DR-2025-087 [No914184] February 15, 2015). A total of 5048 patients met the following inclusion criteria: adult recipients of a single kidney, transplanted for the first time between January 2000 and December 2014, from heart-beating deceased donor, and ABO compatible. We used the ECD definition as proposed by Port al.[3] We did not include 215 patients because they had missing data on at least one variable involved in the ECD definition. Finally, 4833 patients constituted the studied population.

2.2 | Data collected

Recipient pretransplantation variables included the following: age, gender, BMI, initial nephropathy (relapsing disease or not), histories of either high blood pressure (HBP), diabetes, dyslipidemia, cardiovascular diseases, neoplasia, urological disease, dialysis technique before transplantation, time spent on dialysis, and cytomegalovirus serology. Immunization against class I or class II anti-HLA was scored positive if at least one donor-specific antibody (DSA) was identified by Luminex Single Antigen Bead technology within the 6 months before transplantation, unless if at least one DSA was not identified but later determined by Luminex screening or other technology (enzyme-linked immunosorbent assay [ELISA] or Complement Dependent Cytotoxicity [CDC]). Transplantation parameters were cold ischemia time (CIT), HLA A-B-DR incompatibilities, and type of induction treatment (depleting vs nondepleting).

2.3 | Statistical analyses

Patient characteristics were expressed as mean, standard deviation (SD), first and third quartile for continuous variables, or as count and percentage for categorical variables. Comparisons regarding the ECD and SCD status were performed using Student's t-tests, or chi-square tests for quantitative or categorical variables, respectively. For each variable, standardized differences between ECD and SCD recipients were calculated on the original sample. The median event-free follow-up time was obtained by the reverse Kaplan-Meier method.[13]

Similar to the initial study,[3] the main judgment criteria were the time from transplantation (baseline) to graft failure, which was defined as a return-to-dialysis or a death with a functioning graft. To estimate the subject-specific relative risk of graft failure related to the ECD, we used a multivariable Cox proportional hazards model.[14] In addition, we estimated the population-average effect of ECD [15,16] in a propensity score–based analysis by using the inverse probability weighted (IPW) estimator[17,18] with the robust variance estimator.[19] More precisely, we estimated the average treatment

effect on the treated (ATT) effect,[20-22] that is, the patient-graft survival of ECD recipients in a counterfactual world in which these recipients would have instead received SCD graft. Logistic regression was used to compute propensity scores. Positivity assumption was graphically evaluated (Figure S1). Standardized differences on the pseudo-population were all lower than 10%, illustrating that the propensity score–based analysis allows straightening up of the initial covariates' imbalance between ECD and SCD recipients. Adjusted survival curves were estimated by using the weighted Kaplan-Meier estimator.[23,24] The corresponding restricted mean survival times were also estimated.[25]

Concerning the subject-specific estimation, a pool of variables initially selected on univariable survival analysis ($P < .20$), as presented in Table S1, was reduced to a lower number by a descending selection procedure ($P < .05$) in a multivariable Cox model. Then, to appraise the efficient confounding variables, each significant remaining variable was removed if it did not affect the estimation of the hazard ratio associated with ECD transplantation of more than 10%. Concerning the population-average estimation, we first selected variables associated only with graft failure in univariable survival analysis ($P < .20$). Second, this pool of selected variables was included in a multivariable logistic regression and, then, reduced to a lower number by a descending selection procedure ($P < .05$), allowing the efficient confounding variables to be retained and the imprecision in the estimation to be limited.[26] The final retained variables were used to estimate the propensity score used to define ATT weights. Log-linearity assumption was graphically verified. Log-minus-log survival curves allowed graphical assessment of the proportional hazards assumption.

In addition, we studied the time from the transplantation to the death-censored graft failure and the time from the transplantation to the patient death with graft function (returns in dialysis were right censored). Note that we only reported relative risks by using a cause-specific approach to deal with competing events, that is, death with functioning graft and return in dialysis. Indeed, one can expect an important competition in the ECD population because of the age of recipients. We did not report the absolute effects, which would have been based on the cumulative incidence functions, for instance by using the Aalen-Johansen estimator.[27] However, to our knowledge, no one has developed this estimator in an inverse probability of treatment weighting framework.

In addition, we investigated the level of marginality within the ECD recipient cohort. We therefore performed a subgroup analysis by comparing the risk between recipients of ECD grafts from donors older than 75 years versus recipients of ECD graft from younger donors.

All analyses were performed using R software version 3.0.2.[28] The IPW-based analyses were performed using the 0.4 version of the *IPWsurvival* package (www.labcom-risca.com).

3 | RESULTS

3.1 | Description of kidney transplant recipients at baseline

Among the 4833 studied patients, 2105 (44%) received an ECD graft and 2728 (56%) received an SCD graft. We observed a larger proportion of patients receiving an ECD graft compared to the 15% of ECD transplantation in the initial work of Port et al,[3] confirming that marginal grafts are increasingly used. We observed 25% of patients with a donor age between 50 and 59 years, 35% of patients had a donor older than 60 years, while 5% of patients had received a graft from a donor older than 75 years. The maximum donor age was 89 years. Moreover, 14% of donors presented a serum creatinine higher than 1.5 mg/dL and 31% of donors had history of high blood pressure. Fifty-seven percent of donors died of CVA, 28% died of trauma (including 12% of public road accident), 12% died from anoxia, and 3% died from others causes (intoxication, meningitides, tumor, and so on). Demographic characteristics of recipients are summarized in Table 1. When comparing ECD and SCD recipients, significant differences were expected, notably concerning recipient age, since grafts were matched on age where possible. As also illustrated in Figure S2, ECD kidneys are preferentially proposed for older recipients with a mean recipient age estimated at 61 years for ECD and 45 years for SCD ($P < .0001$). As might be expected, ECD recipients presented significantly more comorbidity histories than SCD recipients: notably concerning diabetes (20% for ECD vs 10% for SCD, $P < .0001$), dyslipidemia (37% for ECD vs 24% for SCD, $P < .0001$), and cardiovascular diseases (38% for ECD vs 26% for SCD, $P < .0001$).

3.2 | Description of follow-up

A total of 792 patients returned to dialysis and 488 patients died with a functioning graft. The causes of the recipient death were not collected. The cumulative follow-up covered 22 406 patient-years. The median follow-up time was 5.22 years. The patient-graft survival probability at 3, 6, and 10 years posttransplantation was 85% (95% confidence interval [CI] 84%-86%), 73% (95% CI 72%-75%), and 58% (95% CI 56%-60%), respectively. Patient-graft survival curves for ECD and SCD kidney recipients are presented in Figure S3.

3.3 | Subject-specific effect on patient and graft survival

Table 2 presents results for the multivariable Cox model. The subject-specific relative risk related to ECD versus SCD was estimated at 1.75 (95% CI 1.53-2.00, $P < .0001$). In other words, we confirmed a 1.75-fold increase in the risk of graft failure if a given patient received an ECD graft compared to a graft from an SCD. This estimation should be interpreted independently of recipient age. Nevertheless, regarding the important shift of the recipient age distributions in the ECD and SCD groups (Figure S2), this effect should not be interpreted for young recipients, since they do not have an equal chance of receiving an ECD or SCD graft.

3.4 | Population-average effect on patient and graft survival

The recipient age was the only significant variable retained in the final logistic regression used to estimate the propensity score. This

TABLE 1 Description of recipient, donor, and transplantation characteristics for the sample of kidney transplanted patients (n = 4833) according the ECD/SCD transplantation

	Missing data	Global N = 4833	SCD recipients N = 2728 (56.4%)	ECD recipients N = 2105 (43.6%)	*P*-value	Standardized difference (%)
Quantitative characteristics: mean ± SD (first quartile - third quartile)						
Recipient age (years)	0	51.8 ± 13.4 (43-62)	44.7 ± 11.9 (36-54)	60.9 ± 8.9 (56-67)	<.0001	153.67
Recipient BMI (kg/m²)	40	24.4 ± 4.4 (21.3-27.1)	23.9 ± 4.5 (20.8-26.4)	25.1 ± 4.2 (22.1-27.6)	<.0001	26.56
Time spent on dialysis (years)	13	3.4 ± 3.2 (1.3-4.6)	3.5 ± 3.3 (1.3-4.7)	3.2 ± 3.1 (1.2-4.4)	.0052	8.10
CIT (hours)	30	19.9 ± 7.6 (14.4-24.0)	20.0 ± 7.9 (14.0-24.8)	19.8 ± 7.2 (14.8-23.5)	.2460	3.36
Categorical characteristics: N (%)						
Recipient men	0	2970 (61.5)	1661 (60.9)	1309 (62.2)	.3737	2.67
History of HBP	0	3644 (75.4)	2022 (74.1)	1622 (77.1)	.0206	6.84
History of diabetes	0	704 (14.6)	282 (10.3)	422 (20.0)	<.0001	27.30
History of dyslipidemia	0	1442 (29.8)	655 (24.0)	787 (37.4)	<.0001	29.31
History of cardiovascular diseases	0	1501 (31.1)	709 (26.0)	792 (37.6)	<.0001	25.18
History of neoplasia	0	395 (8.2)	138 (5.1)	257 (12.2)	<.0001	25.67
History of urological disease	0	753 (15.6)	405 (14.8)	348 (16.5)	.1182	4.64
Dialysis technique	5				.2131	5.09
Preemptive transplantation		378 (7.8)	199 (7.3)	179 (8.5)		
Peritoneal dialysis		324 (8.7)	177 (6.5)	147 (7.0)		
Hemodialysis		4126 (85.5)	2350 (86.2)	1776 (84.5)		
Relapsing initial disease	14	1326 (27.5)	883 (32.5)	443 (21.1)	<.0001	25.81
Recipient CMV infection	42	2880 (60.1)	1535 (56.7)	1345 (64.5)	<.0001	15.87
Detectable daily anti-HLA class I	1706	698 (22.3)	374 (21.1)	324 (24.0)	.0598	6.93
Detectable daily anti-HLA class II	1790	536 (17.6)	280 (16.2)	256 (19.4)	.0261	8.28
HLA A-B-DR incompatibilities ≥5	121	681 (14.5)	367 (13.8)	314 (15.3)	.1835	4.02
Depleting induction	31	1883 (39.2)	1060 (39.0)	823 (39.5)	.7875	0.87

TABLE 2 Multivariable Cox model to study the association of ECD and the risk of graft failure, adjusted on the year of transplantation (n = 4833)

	HR	95% CI	*P*
Marginal donor criteria (ECD vs SCD)	1.75	1.53-2.00	<.0001
Recipient age (years)	1.02	1.01-1.03	<.0001

variable also appeared significantly associated with the patient-graft survival (Table S1). Table 3 presents the distribution of the characteristics of the pseudo-population obtained by propensity score weighting. In this pseudo-population, where the characteristics at baseline between ECD and SCD were balanced and comparable with the observed characteristics of the ECD recipients (fifth column of Table 1), we estimated a population-average relative risk of graft failure between the ECD and SCD groups at 1.34 (95% CI 1.09-1.64, *P* = .0049). In other words, we estimated a 1.34-fold decrease in risk of graft failure between patients who had received an ECD and the same patients if they had instead received an SCD graft. The corresponding patient-graft adjusted survival curves are presented in Figure 1. The ECD curve represents the estimated patient-graft survival probability for ECD recipients of an ECD kidney while the SCD curve represents the estimated patient-graft survival probability for ECD patients in the hypothetical situation of receiving an SCD kidney. The survival rates at 3, 6, and 10 years posttransplantation were, respectively, 77%, 62%, and 40% for ECD recipients and 82%, 68%, and 53% for the comparable group of SCD recipients. Therefore, at 10 years posttransplantation, the absolute risk excess due to ECD transplantation was 13% (95% CI 2%-24%). This means that, if 100 patients followed-up to 10 years had received an SCD instead of an ECD graft, 13 graft failures would have been prevented. The corresponding number needed to treat was about 8 (95% CI 4-45), which means that we must transplant 8 patients with an SCD graft instead of an ECD graft to prevent one graft failure at 10 years posttransplantation. For a cohort with a 10-years follow-up, the mean time

TABLE 3 Description of recipient, donor, and transplantation characteristics for the pseudo-population obtained after ATT weighting

	SCD recipients in the pseudo-population N = 2054 (49.4%)	ECD recipients in the pseudo-population N = 2105 (50.6%)	Standardized difference (%)
Quantitative characteristics: mean ± SD			
Recipient age (years)	60.9 ± 9.1	60.9 ± 8.9	0.18
Recipient BMI (kg/m²)	25.1 ± 4.3	25.1 ± 4.2	1.10
Time spent on dialysis (years)	3.6 ± 3.7	3.2 ± 3.1	12.85
CIT (hours)	20.2 ± 8.0	19.8 ± 7.2	5.99
Categorical characteristics: N (%)			
Recipient men	1152 (56.1)	1309 (62.2)	12.42
History of HBP	1570 (76.4)	1622 (77.1)	1.51
History of diabetes	463 (22.5)	422 (20.0)	6.04
History of dyslipidemia	725 (35.3)	787 (37.4)	4.35
History of cardiovascular diseases	824 (40.1)	792 (37.6)	5.12
History of neoplasia	185 (9.0)	257 (12.2)	10.43
History of urological disease	303 (14.7)	348 (16.5)	4.94
Dialysis technique			4.35
Preemptive transplantation	156 (7.6)	179 (8.5)	
Peritoneal dialysis	132 (6.4)	147 (7.0)	
Hemodialysis	1767 (86.0)	1776 (84.5)	
Relapsing initial disease	467 (22.7)	443 (21.1)	3.92
Recipient CMV infection	1354 (66.1)	1345 (64.5)	3.49
Detectable daily anti-HLA class I	359 (27.5)	324 (24.0)	8.06
Detectable daily anti-HLA class II	284 (22.3)	256 (19.4)	7.21
HLA A-B-DR incompatibilities ≥5	324 (16.1)	314 (15.3)	2.22
Depleting induction	817 (39.9)	823 (39.5)	0.81

to graft failure was, respectively, 6.63 years for ECD recipients (95% CI 6.44-6.83 years) and would have been 7.31 years (95% CI 6.85-7.75 years) if they had received an SCD graft. It therefore corresponded to a decrease of 8 months due to ECD transplantation (95% CI 2-14 months).

3.5 | Cause-specific risk of graft failure and cause-specific risk of death with functioning graft

As presented in Table S2, we estimated a 1.84-fold increase in the population-average risk of death-censored graft failure between ECD and SCD recipients (95% CI 1.33-2.55). This population-average relative risk was 0.97 for the risk of death with graft function (95% CI 0.73-1.30).

3.6 | ECD grafts from donors older than 75 years versus ECD grafts from younger donors

We estimated a nonsignificant population-average relative risk of graft failure between patients with an ECD graft from donors older than 75 years versus recipients with an ECD graft from younger donors (hazard ratio [HR] 1.19, 95% CI 0.90-1.58). At 5 years posttransplantation, the graft survival was 46% (95% CI 28%-64%) in the recipients of older donors versus 60% (95% CI 45%-75%) in the recipients of younger donors. We therefore estimated an absolute risk increase due to donor age at 14% (95% CI-10%-38%). This nonsignificant difference was also observed for the risk of death with graft function (HR 0.98, 95% CI 0.64-1.51), and for the risk of death censored graft failure (HR 1.39, 95% CI 0.97-1.99). These estimates are also summarized in Table S2.

4 | DISCUSSION

From the French multicentric DIVAT cohort, we confirmed that the subject-specific relative risk of graft failure between ECD and SCD is close to the one obtained in the initial study: HR 1.75 (95% CI 1.53-2.00) versus 1.70 (95% CI not available). Nevertheless, to our knowledge, no study has estimated the population-average effect of ECD transplantation. Namely, what would the decrease in patient-graft survival be if patients received ECD kidneys instead of SCD kidneys? Alternatively, what would the results have been for a clinical trial among the potential ECD recipients with a randomization into the ECD and SCD groups? Although randomized trials are the gold standard

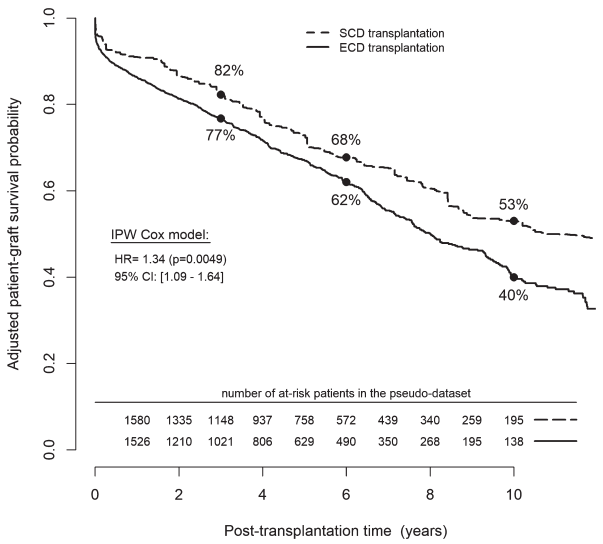


FIGURE 1 Adjusted patient graft survival curves estimated by the IPW Kaplan–Meier estimator for ECD kidney recipients (solid line) and SCD kidney recipients (dashed line)

study design for estimating a population-average effect, such a design is not conceivable here, since ethically the graft allocation could not be randomized. Our study proposes an interesting solution using a suitable methodology providing population-average effects. We estimated a 1.34-fold increase in graft failure risk between the ECD recipients and a comparable group of patients receiving SCD grafts (95% CI 1.09-1.64). This relative risk may possibly not be explained by a difference in the risk of death with a functioning graft (HR 0.97, 95% CI 0.73-1.30). Our propensity score–based approach presents a major interest, since it also provides confounder-adjusted absolute effects. At 10 years after transplantation, the decrease in the mean time to graft failure due to ECD transplantation was estimated at 8 months (95% CI 2-14 months). We also described a difference of 13% in terms of patient and graft survival at 10 years after transplantation (95% CI 2%-24%), corresponding to a number needed to treat at 8 patients (95% CI 4-45).

The population-average effects of ECD transplantation may appear not so high in our population. The absolute risk difference between marginal donor transplantation and nonmarginal donor transplantation could be authorized to be higher, and this may constitute an argument in favor of extending the definition of marginality criteria. To better discuss this possibility, we have further compared the recipients of a graft older than 75 years compared to patients receiving a younger ECD graft. The risk excess may be considered as low with a population-average relative risk of graft failure estimated at 1.19 (95% CI 0.90-1.58), whereas the confounder-adjusted patient and graft survival at 5 years posttransplantation was 46% (95% CI 28%-64%) in the oldest group versus 60% (95% CI 45%-75%) in the youngest group. These results definitively encourage an increase in the pool of marginal donors by considering higher risk grafts. One could for instance allow older donors. Nevertheless, to more precisely address the possible increase of the donor pool, future studies should consider patient survival if the sample of ECD recipients had received no graft, that is, the population-average of ECD versus dialysis.

Several limitations may be highlighted from our observational study. First, we could not exclude the potential bias induced by nonobserved confounders. Second, we did not include patients with missing data concerning variables used in the ECD definition. This concerns 215 patients, that is, less than 5% of the whole sample. Third, even if the ECD status has the advantage of simplicity as a binary criterion defined from only 4 donor characteristics, an important limit of our work is not to study a graduating scale of donor marginality. In 2014, the American Organ Procurement and Transplantation Network (OPTN) proposed the Kidney Donor Risk Index (KDRI),[29] a continuous score of donor marginality, with few external validations.[30] As a perspective of our work, future studies should externally validate the KDRI by studying both subject-specific and population-average effects. Finally, we believe that the most important limit of the ECD definition is to be only based on donor characteristics. This limit is also true for other scoring systems of graft quality, such as the Deceased Donor Score,[31] the Donor Risk Score,[32] the Maryland Aggregate Pathology Index,[33] or the KDRI.[29] None of these scoring systems considers the interactions between donor and recipient characteristics. However,, the kidney graft marginalization should not be determined as an absolute criterion but should depend on the recipients' features.

In conclusion, even if our study confirmed the subject-specific relative risk related to ECD versus SCD, the population-average effects of ECD transplantation seemed not so important. Using a propensity score–based analysis, we proposed an alternative estimation of ECD transplantation at a population level, not better than the traditional multivariate modeling giving subject-specific estimation, but allowing a complementary interpretation. This study reinforces the idea of further increasing the pool of marginal kidneys. This perspective probably needs to integrate both the donor and recipient characteristics to better identify recipients who will benefit from being transplanting by such highly marginal grafts.

ACKNOWLEDGMENTS

We wish to thank members of the clinical research assistant team (S. Le Floch, A. Petit, J. Posson, C. Scellier, V. Eschbach, K. Zurbonsen, C. Dagot, F. M'Raiagh, V. Godel, X. Longy, P. Przednowed). The DIVAT cohort is partially supported by Roche Laboratory since 1994.

DISCLOSURE

The authors of this manuscript have no conflicts of interest to disclose as described by the *American Journal of Transplantation*.

ORCID

C. Legendre http://orcid.org/0000-0002-1743-2948

REFERENCES

1. Tonelli M, Wiebe N, Knoll G, et al. Systematic review: kidney transplantation compared with dialysis in clinically relevant outcomes. *Am J Transplant*. 2011;11(10):2093-2109.
2. Hart A, Smith JM, Skeans MA, et al. OPTN/SRTR 2015 annual data report: Kidney. *Am J Transplant*. 2017;17:21-116.
3. Port FK, Bragg-Gresham JL, Metzger RA, et al. Donor characteristics associated with reduced graft survival: an approach to expanding the pool of kidney donors. *Transplantation*. 2002;74(9):1281-1286.
4. Querard A-H, Foucher Y, Combescure C, 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;29(4):403-415.
5. Molnar MZ, Streja E, Kovesdy CP, et al. Age and the associations of living donor and expanded criteria donor kidneys with kidney transplant outcomes. *Am J Kidney*. 2012;59(6):841-848.
6. Sung RS, Guidinger MK, Leichtman AB, et al. Impact of the expanded criteria donor allocation system on candidates for and recipients of expanded criteria donor kidneys. *Transplantation*. 2007;84(9):1138-1144.
7. Reeves-Daniel AM, DePalma JA, Bleyer AJ, et al. The APOL1 gene and allograft survival after kidney transplantation. *Am J Transplant*. 2011;11(5):1025-1030.
8. Mezrich JD, Pirsch JD, Fernandez LA, et al. Differential outcomes of expanded-criteria donor renal allografts according to recipient age. *Clin J Am Soc Nephrol*. 2012;7(7):1163-1171.
9. Woodside KJ, Merion RM, Leichtman AB, et al. Utilization of kidneys with similar kidney donor risk index values from standard versus expanded criteria donors. *Am J Transplant*. 2012;12(8):2106-2114.
10. Aubert O, Kamar N, Vernerey D, et al. Long term outcomes of transplantation using kidneys from expanded criteria donors: prospective, population based cohort study. *BMJ*. 2015;351:h3557.
11. Ma MKM, Lim WH, Craig JC, Russ GR, Chapman JR, Wong G. Mortality among younger and older recipients of kidney transplants from expanded criteria donors compared with standard criteria donors. *Clin J Am Soc Nephrol*. 2016;11(1):128-136.
12. Westreich D, Cole SR. Invited commentary: positivity in practice. *Am J Epidemiol*. 2010;171(6):674-677.
13. Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. *Control Clin Trials*. 1996;17(4):343-346.
14. Cox D. Regression models and life-tables. *J R Stat Soc Ser B*. 1972;34:187-229.
15. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc*. 1984;79(387):516-524.
16. Gayat E, Resche-Rigon M, Mary J-Y, Porcher R. Propensity score applied to survival data analysis through proportional hazards models: a Monte Carlo study. *Pharm Stat*. 2012;11(3):222-229.
17. Cole SR, Hernan MA. Adjusted survival curves with inverse probability weights. *Comput Methods Programs Biomed*. 2004;75(1):45-49.
18. Binder DA. Fitting Cox's proportional hazards models from survey data. *Biometrika*. 1992;79(1):139-147.
19. Lin DY, Wei LJ. The robust inference for the cox proportional hazards model. *J Am Stat Assoc*. 1989;84(408):1074-1078.
20. 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;33(7):1242-1258.
21. 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*. Oct 2016;25(5):1938-1954.
22. Austin PC. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Stat Med*. 2016;35(30):5642-5655.
23. Xie J, Liu C. Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Stat Med*. 2005;24(20):3089-3110.
24. 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;35:1103-1116.
25. Andersen PK, Perme MP. Pseudo-observations in survival analysis. *Stat Methods Med Res*. 2010;19(1):71-99.
26. Heinze G, Jüni P. An overview of the objectives of and the approaches to propensity score analyses. *Eur Heart J*. 2011;32(14):1704-1708.
27. Aalen OO, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand J Stat*. 1978;5(3):141-150.
28. R Development Core Team. R: A Language and Environment for Statistical Computing [Internet]. Computing RF for S, éditeur. Vienna, Austria; 2016. http://www.R-project.org/. Accessed January 22, 2018.
29. Rao PS, Schaubel DE, Guidinger MK, et al. A comprehensive risk quantification score for deceased donor kidneys: the kidney donor risk index. *Transplantation*. 2009;88(2):231-236.
30. Peters-Sengers H, Heemskerk MBA, Geskus RB, et al. Validation of the prognostic kidney donor risk index (KDRI) scoring system of deceased donors for renal transplantation in the Netherlands. *Transplantation*. 2018;102:162-170.
31. Nyberg SL, Matas AJ, Kremers WK, et al. Improved scoring system to assess adult donors for cadaver renal transplantation. *Am J Transplant*. 2003;3(6):715-721.
32. Schold JD, Kaplan B, Baliga RS, Meier-Kriesche H-U. The broad spectrum of quality in deceased donor kidneys. *Am J Transplant*. 2005;5(4 Pt 1):757-765.
33. Philosophe B, Malat GE, Soundararajan S, et al. Validation of the Maryland Aggregate Pathology Index (MAPI), a pre-implantation scoring system that predicts graft outcome. *Clin Transplant*. 2014;28(8):897-905.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Querard AH, Le Borgne F, Dion A, et al. Propensity score–based comparison of the graft failure risk between kidney transplant recipients of standard and expanded criteria donor grafts: Toward increasing the pool of marginal donors. *Am J Transplant*. 2018;18:1151–1157. https://doi.org/10.1111/ajt.14651

Troisième partie

Projets de recherche

- Chapitre 9 -

Pronostic en transplantation rénale

9.1 Introduction

Les chercheurs et décideurs en santé insistent de plus en plus pour promouvoir des soins centrés sur le patient, respectueux de ses préférences, de ses besoins et de ses valeurs [Methodology Committee of the Patient-Centered Outcomes Research Institute (PCORI), 2012; Basch, 2013; Ziegelstein, 2015]. Une amélioration des connaissances du patient au sujet de sa maladie pourrait l'inciter à mieux adhérer aux traitements proposés et contribuer à une meilleure décision médicale partagée faisant ainsi évoluer la relation traditionnelle et paternaliste entre patient et médecin [Gordon et al., 2013; Jamieson et al., 2016]. Ce paradigme est développé dans certaines maladies chroniques comme l'insuffisance rénale [O'Hare, 2018] ou la transplantation rénale [Sautenet et al., 2018]. Il a récemment été montré en transplantation rénale que les patients priorisaient la survie du greffon avant notamment tout risque d'infections ou de cancer [Howell et al., 2017]. Fournir une prédiction individuelle de la survie du greffon pourrait contribuer à améliorer les connaissances des patients prêts à entendre une telle information au sujet de leur état de santé.

Au vu des statistiques fournies par les registres nationaux et les cohortes, la plupart des cliniciens ont une idée du devenir du greffon au niveau populationel. En revanche, au niveau individuel, cette information est plus difficile à appréhender du fait de la multiplicité des facteurs de risque impliqués dans l'échec de greffe. Informer les patients à chaque visite de suivi sur leur possible devenir à moyen terme pourrait les aider à comprendre et accepter un suivi renforcé en cas de risque de perte du greffon et les impliquer dans ce processus décisionnel, ou bien limiter un stress inutile face à un risque de perte du greffon faible ou stable dans le temps. Une récente revue systématique des outils de prédiction en transplantation rénale souligne d'ailleurs la nécessité de développer des outils de prédiction dynamique [Kaboré et al., 2017].

Le sérum de créatinine, marqueur collecté au cours du suivi du patient transplanté, pourrait par exemple être utile pour calculer des prédictions dynamiques et enrichir les outils de prédictions à temps fixe. Afin de prendre en considération de telles mesures longitudinales, les modèles conjoints pour l'évolution d'un marqueur et le risque de survenue d'un événement sont une approche intéressante [Rizopoulos, 2012; Brankovic et al., 2018]. Quelques travaux sont basés sur de tels modèles en transplantation rénale [Garre et al., 2008; Rizopoulos and Ghosh, 2011; Moranne et al., 2013], mais, à notre connaissance, aucun d'entre eux n'a été utilisé pour développer des prédictions dynamiques.

Dans ce contexte, en collaboration avec Yohann Foucher et Magali Giral, j'ai pu encadrer la thèse de Marie-Cécile Fournier. La section 9.2 comprend son dernier travail portant sur le développement d'un outil de prédiction dynamique de l'échec de greffe rénale. Ce travail est présenté dans cette section car il était en cours de finalisation au moment de la rédaction de mon manuscrit d'HDR, il a depuis été accepté dans la revue *Nephrology, Dialysis and Transplantation* [Fournier et al., 2019]. Son travail de thèse a également ouvert la porte à une collaboration internationale avec Özgür Asar, enseignant chercheur à l'Université d'Acibadem en Turquie. Dans les sections 9.3 et 9.4, je présente les projets en cours de développement dans le cadre de cette collaboration internationale. Il s'agit de travaux en Biostatistique dont les résultats pourraient contribuer à l'amélioration de l'outil de prédiction dynamique de l'échec de greffe rénale.

9.2 Prédictions dynamiques de l'échec de greffe rénale

Le premier travail de thèse de Marie-Cécile Fournier a été de proposer un modèle conjoint à effets aléatoires partagés pour l'évolution du sérum de créatinine et le risque d'échec de greffe (voir le chapitre 5 pour plus de détails) [Fournier et al., 2016]. En plus des mesures répétées de sérum de créatinine, les caractéristiques inclues dans ce modèle sont l'âge, le genre, les antécédents de diabète et de maladies cardiovasculaires du receveur, le rang de la greffe, l'immunisation pré-greffe anti-HLA de classe I, l'âge, le genre et le statut vital du donneur, la survenue d'au moins un rejet aigu dans la première année de greffe et les niveaux de sérum de créatinine à 3 et 6 mois post-transplantation. L'objectif du travail présenté ici a été de proposer un outil simple de prédiction dynamique de survie à un échec de greffe à moyen terme (*Dynamic prediction of the Patient and Graft survival - DynPG*). Pour cela, nous avons simplifié le modèle conjoint initialement proposé en réduisant le nombre de prédicteurs. Nous validons par ailleurs les capacités pronostiques de cet outil de prédiction dynamique.

9.2.1 Matériels et méthodes

Population étudiée

Les données ont été extraites de la cohorte DIVAT (`www.divat.fr`). Un total de 4121 patients répondaient aux critères d'inclusion suivant : receveurs adultes d'une première ou seconde transplantation rénale entre janvier 2000 et août 2013 d'un donneur vivant ou décédé à cœur battant, vivant avec un greffon fonctionnel à un an post-transplantation, et maintenu sous Tacrolimus et Mycofenolate. Cet échantillon a été divisé aléatoirement : 2/3 des patients (n=2749) constituent l'échantillon d'apprentissage, 1/3 des patients (n=1372) auxquels nous ajoutons une mise à jour de l'extraction des données sur la période de transplantation allant de septembre 2013 à octobre 2016 (n=1217) constitueront l'échantillon de validation (n=2589).

Critères de jugement

Le critère de jugement est le délai de survenue d'un échec de greffe défini comme le délai entre la date de début du suivi du patient (premier anniversaire de la transplantation) et le premier événement entre un retour en dialyse, une retransplantation préemptive ou un décès avec un greffon fonctionnel. Pour certains patients, ce délai de suivi est possiblement censuré à droite par la date de dernières nouvelles. Nous nous intéresserons au pronostic d'un échec de greffe dans la phase chronique de l'évolution de l'état de santé du patient. Pour cela, nous écartons la première année de greffe, phase un peu particulière d'adaptation du patient à sa condition de transplanté et d'adaptation des traitements. Ainsi, dans les critères d'inclusion, nous ne gardons que des patients vivants avec un greffon fonctionnel à un an post-transplantation. Nous considérons par ailleurs l'évolution du sérum de créatinine (en μmol/L) avec des mesures répétées collectées annuellement jusqu'à l'échec de greffe. Nous avons retenu une fenêtre de prédiction de 5 ans comme temps d'horizon afin de prédire le risque d'échec de greffe à moyen terme.

Analyses statistiques

Le nombre important de prédicteurs (12 variables) retenus dans le modèle conjoint initialement proposé était lié à l'objectif étiologique de cette précédente étude, à savoir décrire précisément les facteurs associés à l'évolution du sérum de créatinine et/ou au risque d'échec de greffe. Dans une perspective pronostique, seules les variables liées au critère de survie sont considérées. En effet, l'évolution du sérum de créatinine étant sur le chemin causal entre certaines variables disponibles à l'origine et l'échec de greffe, il n'apparaît pas nécessaire d'ajuster sur les variables préalablement mis en évidence comme associées à l'évolution du sérum de créatinine. Nous utilisons le même échantillon d'apprentissage que dans l'article précédemment publié afin de construire un modèle plus parcimonieux [Fournier et al., 2016]. Finalement, le modèle conjoint

retenu pour définir les prédictions dynamiques de survie à un échec de greffe est composé des mesures annuelles de sérum de créatinine et de 6 variables disponibles au début de l'étude : l'âge du receveur, le rang de la greffe, les antécédents de maladies cardiovasculaires, le niveau de sérum de créatinine à 3 mois post-transplantation, la survenue d'au moins un rejet aigu au cours de la première année de greffe et l'immunisation pré-greffe anti-HLA de classe I.

La prédiction dynamique de survie à un échec de greffe (DynPG) est définie comme la probabilité de ne pas connaître un échec de greffe dans une fenêtre de 5 ans suivant le temps de landmark, i.e. le temps auquel la prédiction est effectuée. Le nombre de sujets à risque à la sortie de la fenêtre des 5 ans pour une prédiction réalisée à 6 ans post-transplantation étant seulement de 178 dans l'échantillon de validation, nous ne réalisons des prédictions que pour des temps de landmark variant entre 1 et 6 ans post-transplantation. Les performances prédictives ont été évaluées selon les recommendations TRIPOD [Collins et al., 2015; Moons et al., 2015]. Les performances globales de l'outil que nous proposons sont évaluées à l'aide du critère R^2 que nous avons également proposé (voir le chapitre 6 pour plus de détails) [Fournier et al., 2018]. Les capacités discriminantes sont étudiées à l'aide de l'aire sous la courbe ROC pour des prédictions dynamiques [Blanche et al., 2015]. La calibration est décrite en comparant les valeurs prédites au sein de sous-groupes définis par des quantiles de prédictions aux survies observées calculées par l'estimateur de Kaplan-Meier.

9.2.2 Résultats

Description

La Table 9.1 présente la comparaison des caractéristiques des patients des échantillons d'apprentissage (n=2749) et de validation (n=2589). Les patients de l'échantillon de validation sont plus fréquemment immunisés contre les antigènes HLA de classe I (40.3% versus 32.7%) et II (35.5% versus 29.8%), avec des comorbidités plus fréquentes (14.1% versus 11.6% pour les antécédents de diabètes ; 37.0% versus 31.3% pour les antécédents de dysplipidémie ; 11.0% versus 8.3% pour les antécédents de cancer), ont un temps d'ischémie moyen plus court (16 heures versus 17.8 heures). Ces différences pourraient être expliquées en partie par l'inclusion de patients transplantés plus récemment dans l'échantillon de validation.

Au cours du suivi, dans l'échantillon de validation, 259 patients retournent en dialyse, 6 sont re-greffés pré-emptivement et 196 décèdent avec un greffon fonctionnel. Dans l'échantillon d'apprentissage, 275 patients retournent en dialyse, 3 sont re-greffés pré-emptivement et 203 décèdent avec un greffon fonctionnel. Le délai de suivi médian, calculé par la méthode inverse de Kaplan-Meier, est estimé à 3.1 ans [Shuster, 1991]. La probabilité de survie pour un échec de greffe à 8 ans post-transplantation est estimé à 71.4% (Intervalle de confiance à 95% -IC95%- de 68.8% à 74.1%) pour l'échantillon de validation contre 71.8% (IC95% de 69.3% à 74.5%) pour l'échantillon d'apprentissage (Test du Log-Rank, p=0.5191).

Table 9.1: *Descriptif des caractéristiques des receveurs et donneurs et des caractéristiques de la transplantation selon l'échantillon d'apprentissage (n=2749) et l'échantillon de validation (n=2589).*

	Base d'apprentissage (n=2749)			Base de validation (n=2589)			p-value
	NA	Caractéristiques		NA	Caractéristiques		
Variables quantitatives : moyenne (écart-type)							
Age du receveur (années)	0	49.71	(13.59)	0	50.63	(14.25)	0.0158
IMC du receveur (kg/m^2)	10	23.99	(4.24)	7	24.65	(4.43)	<0.0001
Age du donneur (années)	1	50.74	(15.52)	4	52.12	(15.95)	0.0013
SCr du donneur (μmol/L)	25	89.91	(52.77)	21	88.24	(53.46)	0.2557
Temps d'ischémie froide (heures)	10	17.76	(9.79)	7	15.96	(9.51)	<0.0001
Temps passés en dialyse (années)	36	2.98	(3.08)	23	2.93	(3.19)	0.5649
SCr à 3 mois post-greffe (μmol/L)	38	138.30	(53.38)	36	138.77	(55.10)	0.7528
SCr à 6 mois post-greffe (μmol/L)	75	136.64	(53.18)	73	135.43	(48.91)	0.3930
Variables catégorielles : N (%)							
Receveurs hommes	0	1674	(60.89)	0	1596	(61.65)	0.5737
Secondes greffes	0	474	(17.24)	0	442	(17.07)	0.8689
Technique de dialyse :	4			0			0.4841
Greffe préemptive		342	(12.46)		346	(13.36)	
Hémodialyse		2191	(79.82)		2032	(78.49)	
Dialyse péritonéale		212	(7.72)		211	(8.15)	
Maladie initiale de type récidivante	0	799	(29.07)	0	687	(26.54)	0.0393
Antécédent de diabète	0	319	(11.60)	0	365	(14.10)	0.0064
Antécédent d'hypertension	0	2272	(82.65)	0	2120	(81.88)	0.4654
Antécédent cardiovasculaire	0	933	(33.94)	0	911	(35.19)	0.3380
Antécédent de dyslipidémie	0	860	(31.28)	0	957	(36.96)	<0.0001
Antécédent de néoplasie	0	228	(8.29)	0	284	(10.97)	0.0009
Incompatibilités HLA-ABDR >4	7	350	(12.76)	3	399	(15.43)	0.0052
Immunisé en classe I en pré-greffe	66	876	(32.65)	30	1032	(40.33)	<0.0001
Immunisé en classe II en pré-greffe	87	792	(29.75)	45	904	(35.53)	<0.0001
Donneurs hommes	8	1545	(56.37)	3	1468	(56.77)	0.7680
Statut du donneur	6			7			0.0026
Donneurs vivants		418	(15.21)		481	(18.63)	
Décès cérébrovasculaire		1309	(47.72)		1151	(44.58)	
Décès non cérébrovasculaire		1016	(37.05)		950	(36.79)	
Retard au démarrage du greffon	15	714	(26.12)	9	414	(27.67)	0.2001
Survenue d'un rejet aigu avant 1 an post-transplantation	0	591	(21.50)	0	499	(19.27)	0.0439
Greffes antérieures à 2008	0	2091	(39.17)	0	1369	(49.80)	<0.0001

NA : manquant ; IMC : Indice de masse corporelle ; SCr : sérum de créatinine ; HLA : Human Leucocytes Antigen

Modèle conjoint simplifié

La Table 9.2 présente les paramètres du modèle conjoint simplifié estimé sur l'échantillon d'apprentissage. Ce modèle contient 6 variables de moins que le modèle initialement publié dans l'European Journal of Epidemiology [Fournier et al., 2016]. L'âge du receveur, le rang de la greffe, les antécédents de maladies cardiovasculaires, le niveau de sérum de créatinine à 3 mois post-transplantation, la survenue d'au moins un rejet aigu dans la première année de greffe et l'immunisation pré-greffe anti-HLA de classe I semblent être significativement associés au risque d'échec de greffe ($p < 0.05$). Quel que soit le temps au-delà du premier anniversaire de la greffe, l'échec de greffe semble dépendre du niveau courant de sérum de créatinine et de la pente courante d'évolution de sérum de créatinine. Si un patient a un niveau de sérum de créatinine

25% plus élevé, son risque d'échec de greffe est 2 fois plus important (HR=1.96, IC95% de 1.79 à 2.15). Pour une valeur de sérum de créatinine donnée, si un patient a une augmentation de sérum de créatinine plus marquée, son risque d'échec de greffe est plus important (HR=1.84, IC95% de 1.11 à 3.04, pour une augmentation de 25% en 1 an).

Table 9.2: *Modèle conjoint simplifié pour l'évolution du logarithme de sérum de créatinine (SCr) et le risque d'échec de greffe chez des patients transplantés rénaux (n=2584, 165 patients exclus pour cause de données manquantes).*

	Processus de survie		
	HR	IC95%	p-value
Valeur courante de SCr (μmol/L), pour une augmentation de 25%	1.96	[1.79 ;2.15]	<0.0001
Augmentation du niveau de SCr (μmol/L), de 25%/an	1.84	[1.11 ;3.04]	0.0176
Age du receveur (années, standardisé)	1.49	[1.33 ;1.66]	<0.0001
Antécédent cardiovasculaire (positif versus négatif)	1.41	[1.16 ;1.71]	0.0007
SCr à 3 mois post-greffe (μmol/L, standardisé)	0.83	[0.74 ;0.93]	0.0011
Survenue d'au moins un rejet aigu dans la première année (positif versus négatif)	1.46	[1.16 ;1.82]	0.0011
Immunisation pré-greffe anti-HLA de classe I	1.54	[1.22 ;1.94]	0.0002
Rang de la greffe (seconde versus première)	1.31	[1.01 ;1.71]	0.0433

Intercept à 1 an post-transplantation du log(SCr) : 4.86, IC95% [4.85 ;4.87] ; Pente du log(SCr) : 0.024, IC95% [0.021 ;0.028]
Distribution de Weibull pour le risque de base du sous-modèle de survie (intercept : -20.72±0.97 ; log(shape) :0.33±0.05)
Sous-modèle de survie ajusté sur un effet période (avant 2008 versus après 2008) : HR=0.74, IC95% [0.58 ;0.95]

Capacités pronostiques

La Figure 9.1 ne révèle pas de problèmes majeurs de calibration, même si nous pouvons noter de meilleures performances prédictives pour les temps de landmark les plus précoces. Les performances globales de prédiction de la survie à un échec de greffe sur un horizon de 5 ans semblent relativement similaires quelque soit le temps de landmark (Figure 9.2, Partie A). Les valeurs de R^2 sont estimées à 14% (IC95% de 7% à 21%) et 15% (IC95% de -2% à 33%) pour des prédictions faites à 1 et 6 ans post-transplantation respectivement. Au vu d'une calibration qui semble raisonnable, à 6 ans post-transplantation, la survie patient-greffon pourrait être en moyenne 15% plus basse chez les patients qui expérimentent réellement un échec de greffe comparés à ceux qui ne l'expérimentent pas (voir le chapitre 6 pour plus de détails concernant les interprétations du R^2). Les capacités discriminantes augmentent avec le temps de landmark (Figure 9.2, Partie B). Les aires sous la courbe ROC sont estimées à 0.72 (IC95% de 0.67 à 0.78) et 0.76 (IC95% de 0.68 à 0.85) pour des prédictions faites à 1 et 6 ans post-transplantation respectivement. Cela signifie qu'à 6 ans post-transplantation, il y a près de 76% de chance que la probabilité de survie pour un sujet subissant un échec de greffe dans les 5 ans soit plus faible que la probabilité de survie pour un sujet ne connaissant pas d'échec de greffe sur cette même période. Enfin, nous pouvons également noter que les capacités pronostiques en termes de calibration et de discrimination obtenues à l'aide du modèle conjoint simplifié apparaissent similaires à celles que nous aurions à partir du modèle conjoint précédemment proposé (Figure 9.3).

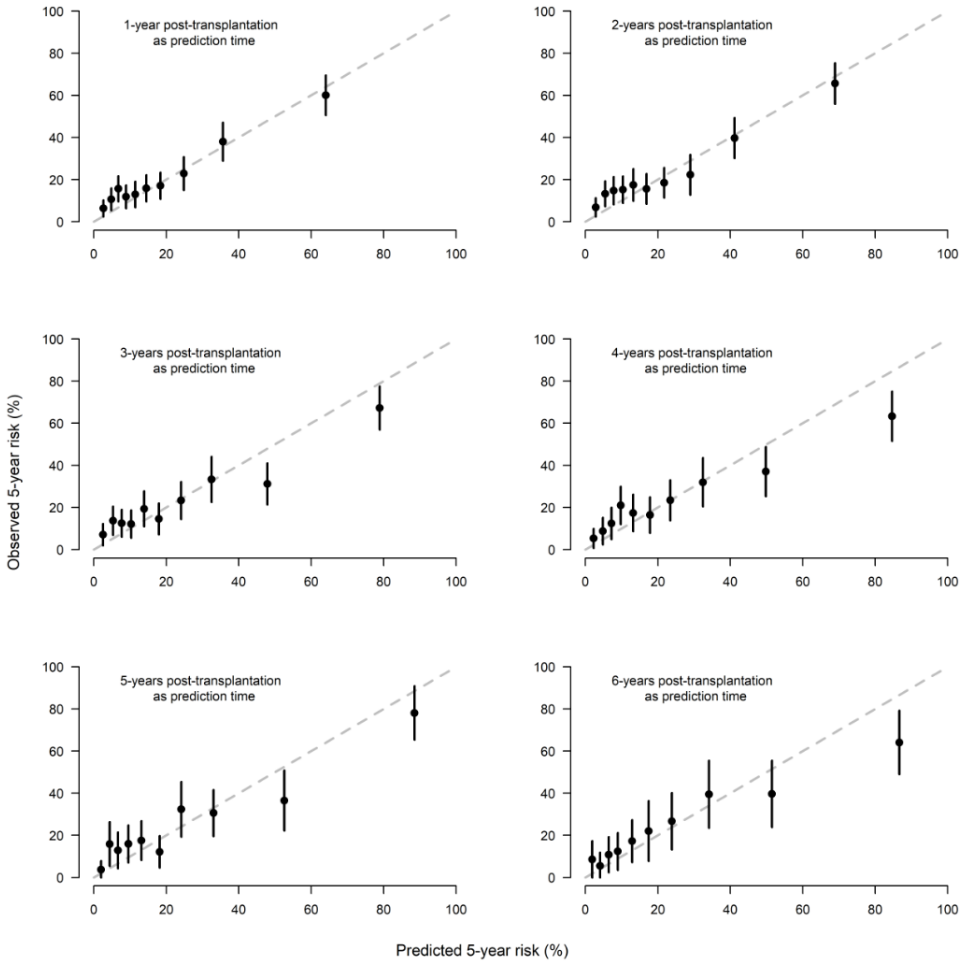


Figure 9.1: *Calibration des prédictions dynamiques sur l'échantillon de validation (n=2523, 66 sujets supprimées pour cause de données manquantes) pour des temps de landmark allant de 1 à 6 ans post-transplantation. Les risques moyens prédits et observés (Kaplan-Meier) sont présentés pour des sous-groupes définis par les quantiles de prédictions.*

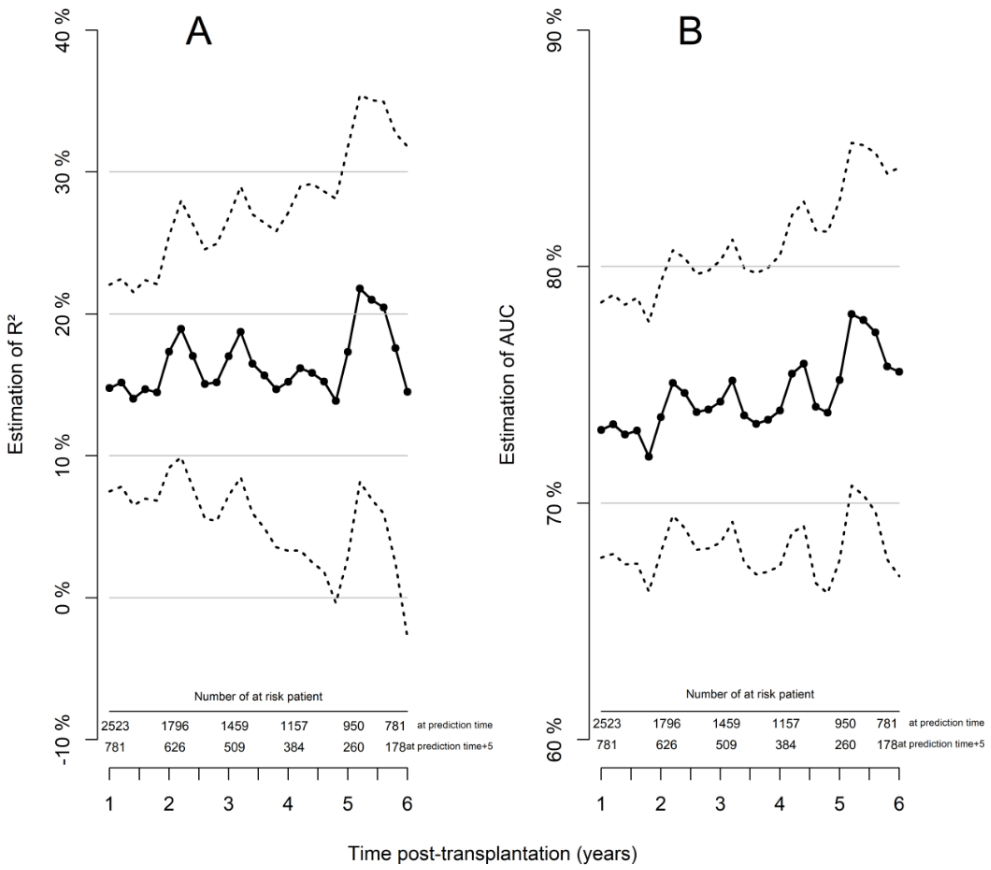


Figure 9.2: *Capacités pronostiques des prédictions dynamiques sur un horizon de 5 ans (n=2523, 66 sujets supprimées pour cause de données manquantes) estimées pour des temps de landmark allant de 1 à 6 ans post-transplantation. La courbe de R^2 est un indicateur de performance globale (Partie A), tandis que l'aire sous la courbe ROC (AUC) évalue les performances de discrimination (Partie B). Les estimations sont représentées par les lignes en traits plein, les bandes de confiance à 95% sont dessinées en trait pointillé.*

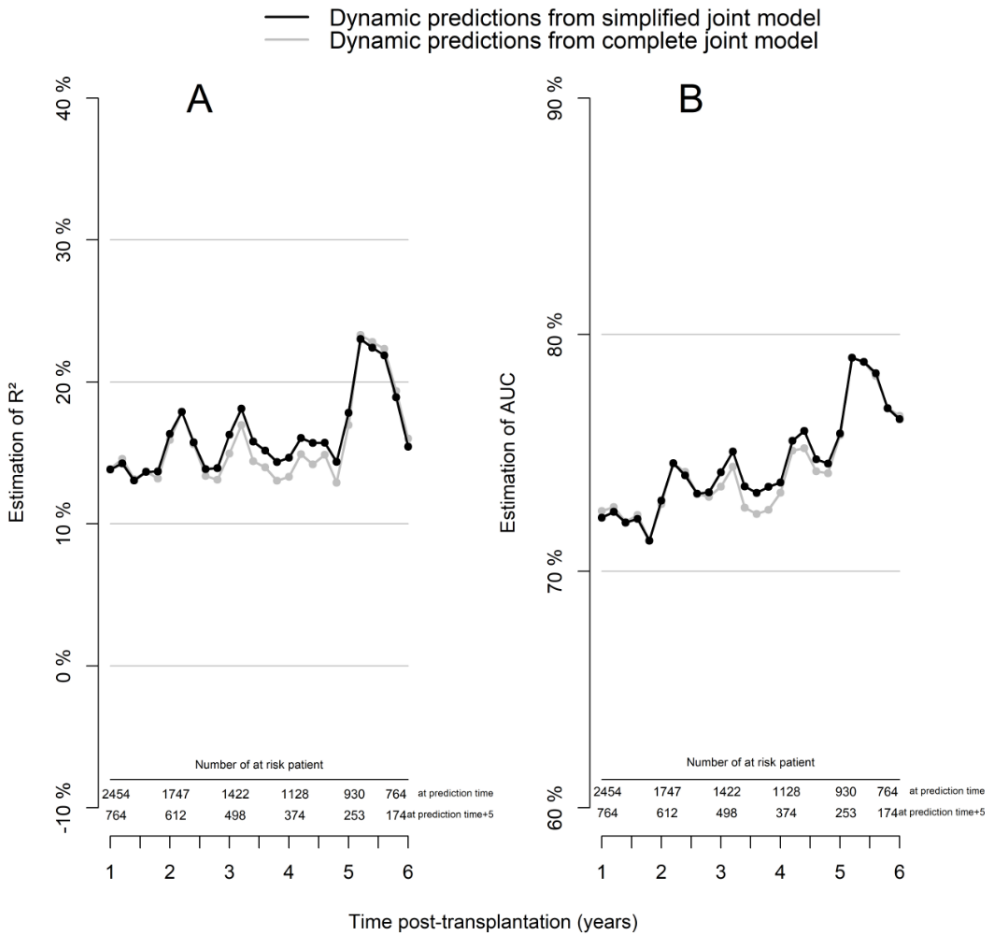


Figure 9.3: *Capacités pronostiques des prédictions dynamiques obtenues à partir du modèle conjoint publié par Fournier et al. [2016] et comparées à celles obtenues à partir du modèle conjoint simplifié. Ces prédictions sont calculées à partir de l'échantillon de validation (n=2454, 135 sujets supprimés pour cause de données manquantes) estimées pour des temps de landmark allant de 1 à 6 ans post-transplantation. La courbe de R^2 est un indicateur de performance globale (Partie A), tandis que l'aire sous la courbe ROC (AUC) évalue les performances de discrimination (Partie B).*

Exemples de prédictions dynamiques

Pour illustrer l'intérêt des prédictions dynamiques, nous présentons deux cas cliniques de patients issus de l'échantillon de validation (Figures 9.4 et 9.5). Nous calculons leurs prédictions dynamiques à chaque temps de mesure de sérum de créatinine. Les mesures de sérum de créatinine et l'évolution estimée sont représentées sur la partie gauche du graphique et la probabilité estimée de survie à un échec de greffe est représentée sur la partie droite. Nous y représentons également la probabilité de survie moyenne brute comme référence. Calculée chez les patients en vie avec un greffon fonctionnel au temps de landmark considéré, cette probabilité est estimée à partir de l'estimateur de Kaplan-Meier sur l'échantillon d'apprentissage.

Cas A - Femme de 51 ans transplantée pour la première fois en 2005, immunisée en classe I en pré-greffe, sans antécédent de maladies cardiovasculaires, ni rejet aigu au cours de la première année de greffe et avec un niveau de sérum de créatinine à 88 µmol/L à 3 mois post-transplantation.

A 3 ans post-transplantation, cette patiente a 87% de chances d'être en vie avec un greffon fonctionnel 5 ans plus tard, c'est-à-dire à 8 ans post-transplantation (IC95% de 46% à 97%, Figure 9.4). Pour ce temps de landmark, sa courbe de survie estimée est supérieure à celle de la population à risque au même temps de prédiction. Un an plus tard, à 4 ans post-transplantation, nous pouvons observer une augmentation du niveau de sérum de créatinine conduisant à une diminution de la probabilité de survie à un échec de greffe. En dépit d'une stabilisation du niveau de sérum de créatinine dans les années à suivre, (à 5 et 6 ans post-transplantation), son risque d'échec de greffe reste élevé. A 6 ans post-transplantation, nous estimons qu'il y a près de 98% de risque de perte du greffon ou de décès dans les 5 années à suivre. Bien évidemment, même sans l'utilisation de l'outil DynPG, les cliniciens et patients seraient inquiets face à une telle dégradation de la fonction rénale observée à partir de 4 ans post-greffe. Néanmoins, il nous semble difficile pour le médecin d'estimer précisément le risque d'échec de greffe et de répondre aux interrogations du patient concernant ses chances de conserver son greffon fonctionnel dans un futur proche. L'outil DynPG que nous proposons peut aider à fournir de telles prédictions.

Cas B - Femme de 60 ans transplantée pour la seconde fois en 2007, immunisée en classe I en pré-greffe, sans antécédent de maladies cardiovasculaires, ayant connu au moins un épisode de rejet aigu au cours de la première année de greffe et avec un niveau de sérum de créatinine à 100µmol/L à 3 mois post-transplantation.

En contraste avec le cas clinique A, nous n'observons pas d'augmentation du sérum de créatinine selon le temps post-transplantation pour cette patiente (Figure 9.5). A 5 ans post-transplantation, la probabilité de survie à 10 ans post-transplantation est estimée à 89% (IC95% de 68% à 97%, Figure 9.5). Alors que les caractéristiques au début du suivi de cette patiente semblent de moins bon pronostic comparativement au cas A, le maintien du niveau de sérum de créatinine au cours du temps conduit le risque d'échec de greffe à rester relativement stable au

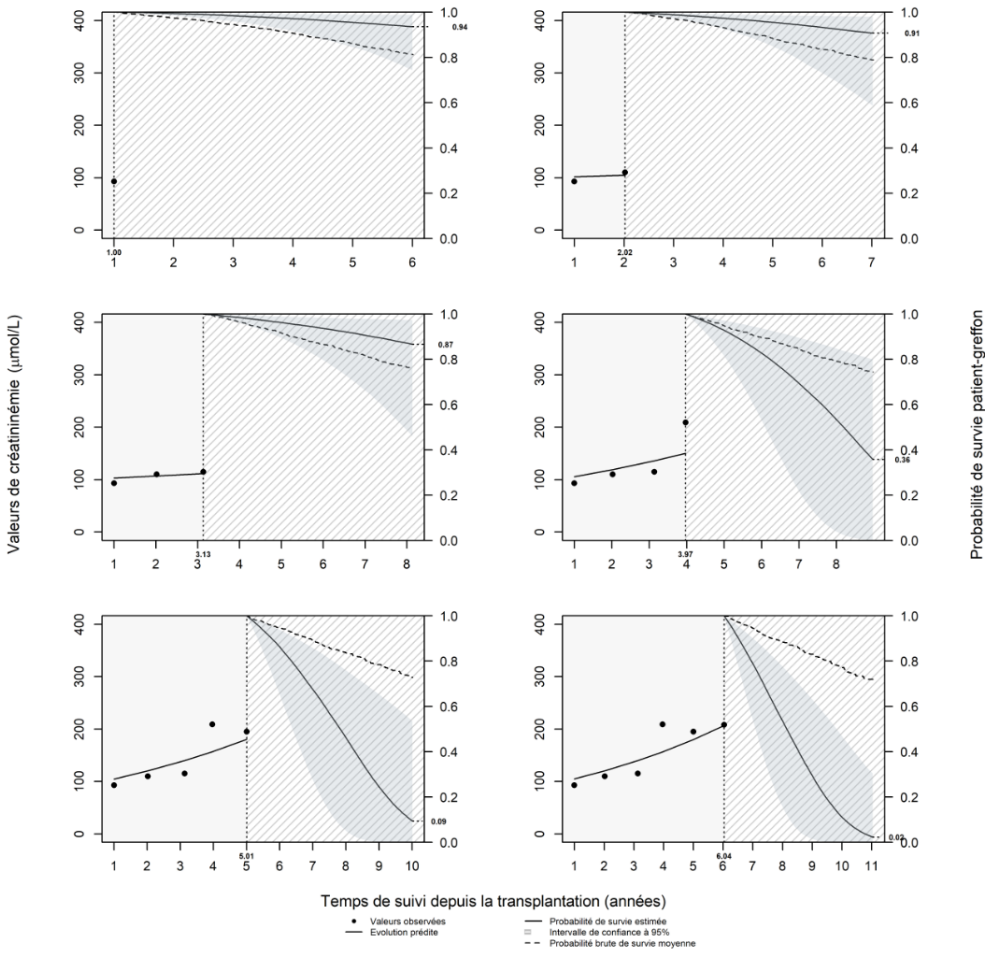


Figure 9.4: *Prédictions dynamiques individuelles obtenues à partir du modèle conjoint simplifié pour des temps de landmark allant de 1 à 6 ans post-transplantation pour une femme de 51 ans transplantée pour la première fois en 2005, immunisée en classe I en pré-greffe, sans antécédent de maladies cardiovasculaires, ni rejet aigu au cours de la première année de greffe et avec un niveau de sérum de créatinine à 88 µmol/L à 3 mois post-transplantation (Cas A). Cette patiente subit finalement un retour en dialyse à 9.3 ans post-transplantation.*

cours du suivi. Considérer l'évolution du sérum de créatinine illustre l'importance de la mise à jour d'une prédiction.

Validation externe

Nous avons par ailleurs débuté un travail de validation externe de notre outil de prédiction dynamique de l'échec de greffe rénale. En collaboration avec le Pr. Marc Hazzan et le Dr. Rémi Lenain du service de néphrologie et transplantation du CHU de Lille, nous souhaitons étudier les capacités pronostiques de DynPG à partir d'une cohorte de patients transplantés rénaux indépendants de ceux du réseau DIVAT. Nous regarderons les propriétés de calibration et de discrimination sur un échantillon de patients du CHU de Lille répondant aux critères d'inclusion suivant : receveurs adultes d'une première ou seconde transplantation rénale seule après janvier 2000 d'un donneur vivant ou décédé à cœur battant, vivant avec un greffon fonctionnel à un an post-transplantation, et maintenus sous Tacrolimus et Mycofenolate.

9.2.3 Discussion

La médecine 4P, prédictive, préventive, personnalisée et participative, est un concept largement développé dans la littérature mais encore trop peu mis en pratique [Flores et al., 2013]. Informer le receveur d'un rein sur son pronostic à moyen terme pourrait accroître l'adhérence du patient à ses traitements et lui donner un rôle actif dans le processus décisionnel médical [Gordon et al., 2013]. L'outil DynPG permet de fournir des prédictions utiles aux cliniciens et aux patients. Par exemple, concernant la patiente A, alors que nous prédisons un risque élevé d'échec de greffe à partir de 4 ans post-transplantation, ces estimations pourraient être utiles au patient pour comprendre les changements thérapeutiques qui lui sont proposés et se préparer psychologiquement et émotionnellement à son possible devenir à moyen terme. Cela pourrait contribuer à rendre le patient acteur de son suivi de patient transplanté. Concernant la patiente B, les prédictions dynamiques montrent un risque d'échec de greffe relativement stable dans le temps. Fournir la prédiction individuelle au patient pourrait contribuer à le rassurer concernant son possible devenir à moyen terme. Ces exemples soulignent l'utilité de l'outil de prédiction dynamique comparativement aux outils de prédiction à temps fixe.

Plusieurs limites à ce travail doivent tout de même être mentionnées. Premièrement, les prédictions dynamiques sont basées sur la seule évolution du sérum de créatinine. D'autres marqueurs longitudinaux, comme la protéinurie, l'immunisation post-greffe, la survenue d'infections, pourraient contribuer à améliorer les capacités pronostiques de l'outil que nous proposons. L'inclusion de multiples marqueurs dans un modèle conjoint resterait à mettre en oeuvre et fait actuellement l'objet de développement dans la littérature en Biostatistique. Deuxièmement, définir un échec de greffe comme étant le premier événement entre un retour en dialyse, une retransplantation préemptive et un décès avec un greffon fonctionnel est discutable puisqu'un

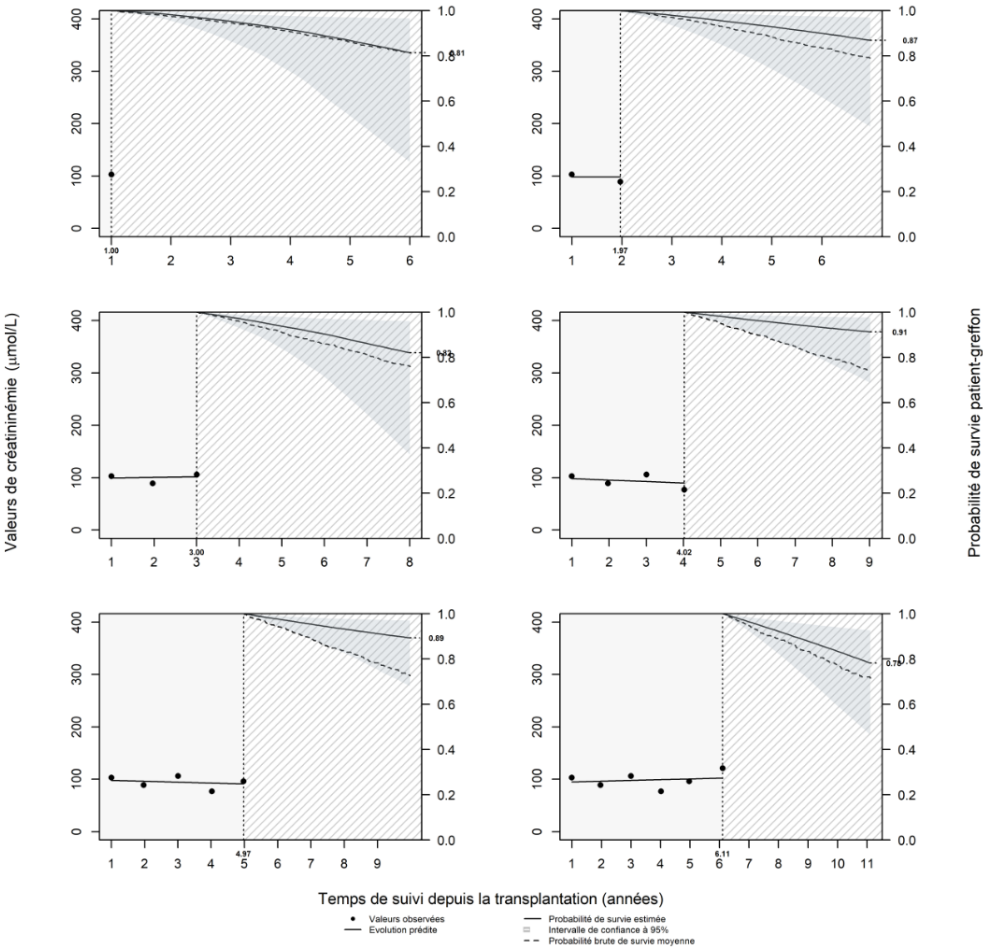


Figure 9.5: *Prédictions dynamiques individuelles obtenues à partir du modèle conjoint simplifié pour des temps de landmark allant de 1 à 6 ans post-transplantation pour une femme de 60 ans transplantée pour la seconde fois en 2007, immunisée en classe I en pré-greffe, sans antécédent de maladies cardiovasculaires, ayant connu au moins un épisode de rejet aigu au cours de la première année de greffe et avec un niveau de sérum de créatinine à 100µmol/L à 3 mois post-transplantation (Cas B). Cette patiente est toujours en vie avec un greffon fonctionnel à 10 post-transplantation.*

nombre non négligeable de décès avec un greffon fonctionnel est associé à une autre cause que la transplantation. Une solution alternative pourrait être de censurer à droite la survenue d'un décès mais là encore ce choix pourrait être discuté. Troisièmement, nous avons choisi une fenêtre de prédiction de 5 ans pour prédire le devenir du patient à moyen terme. D'autres fenêtres d'horizon auraient pu être retenues. Une fenêtre plus courte aurait pu être intéressante pour préparer le patient à être réinscrit sur liste d'attente précocement ou pour envisager une retransplantation préemptive si le risque individuel d'échec de greffe dans un horizon court apparaît élevé. Enfin, l'outil de prédiction dynamique que nous proposons repose sur un modèle conjoint à effets aléatoires partagés soumis à certaines hypothèses qui peuvent être discutées. Nous abordons ce point dans la section suivante.

En conclusion, nous proposons et validons un outil de prédiction dynamique de l'échec de greffe rénale. Nous avons développé un calculateur en ligne permettant de facilement obtenir ces prédictions dynamiques (`https://shiny.idbc.fr/DynPG`). Cet outil de prédiction semble prometteur pour informer le patient sur son possible devenir et pourrait contribuer à une amélioration des prises de décision médicale partagée entre patients et cliniciens. Indépendamment du développement de cet outil de prédiction, il est également nécessaire d'entamer une réflexion sur son utilisation qui pose plusieurs questions éthiques. Par exemple, la question du mode de transmission d'une telle information au patient doit être posée et réfléchie. Il semble important d'avoir une réflexion sur le développement de programme d'éducation et de soutien psychologique pour aider les patients à comprendre et intégrer le risque d'échec de greffe rénale sans surinterprétation. Notre hypothèse de départ repose sur le fait que fournir de telles prédictions pourraient aider le patient à mieux appréhender l'issu de sa greffe. Il sera également nécessaire de se donner les moyens d'évaluer l'impact de la transmission d'une telle information sur la qualité de vie du patient, son anxiété, son stress, etc.

9.3 Prédictions dynamiques en présence de mesures longitudinales extrêmes

Nous constatons un intérêt grandissant pour les méthodes permettant d'obtenir des prédictions dynamiques de survenue d'un événement. Parmi celles-ci, les modèles conjoints pour données longitudinales et données de survie sont très populaires [Rizopoulos, 2012; Proust-Lima and Taylor, 2009; Taylor et al., 2013]. Dans leur formulation classique, le processus longitudinal est modélisé à l'aide de modèles mixtes supposant des distributions Gaussiennes pour les effets aléatoires et les erreurs résiduelles, hypothèses que nous avons considérées dans le travail publié dans l'European Journal of Epidemiology (Chapitre 5) ou dans le cadre du développement des prédictions dynamiques de l'échec de greffe rénale (Section 9.2).

En analyse de données longitudinales, certaines observations peuvent être quelque peu in-

attentues. Ainsi, deux types d'observations longitudinales *extrêmes* peuvent être définis : i) au niveau populationnel, des sujets extrêmes sont ceux n'ayant pas une trajectoire individuelle similaire à la tendance en population, ii) au niveau individuel, des observations ponctuelles extrêmes pour un sujet sont celles s'écartant très fortement de la trajectoire du sujet [Pinheiro et al., 2001; Wu, 2009]. Face à ce type de valeurs extrêmes, l'hypothèse de normalité des effets aléatoires et erreurs residuelles pourrait ne pas être la plus adaptée et donner des poids inappropriés aux individus et observations [Sutradhar and Ali, 1986; Lange et al., 1989]. En conséquence, les prédictions dynamiques obtenues à partir d'un modèle conjoint classique pourraient ne pas être optimales.

Les modèles conjoints robustes sont une extension des modèles conjoints classiques. Ils font l'hypothèse de distributions non-Gaussiennes pour les effets aléatoires et/ou erreurs résiduelles. Un choix naturel pour ces distributions est la distribution de Student (ou distribution t) présentant des queues de distribution plus importantes (i.e. *heavy tails distribution*) que la loi Normale, permettant ainsi de donner des poids plus appropriés aux sujets et valeurs extrêmes [Wu, 2009; Li et al., 2009]. Plusieurs auteurs ont conclu à de meilleures performances des modèles conjoints robustes comparativement aux modèles conjoints classiques concernant l'estimation des paramètres et particulièrement concernant l'estimation de leurs erreurs standards [Li et al., 2009; Huang et al., 2010; Baghfalaki et al., 2013, 2014; Kim and Albert, 2016]. A notre connaissance, Taylor et al. [2013] est le seul travail considérant un modèle conjoint robuste à des fins de prédictions de probabilité de survie. Les auteurs supposent une distribution Gaussienne pour les effets aléatoires et une distribution t pour le terme d'erreur résiduelle avec un nombre de degré de liberté fixé à 5. L'utilisation d'approches par modèles conjoints robustes reste encore un sujet de recherche en Biostatistique afin d'obtenir des prédictions dynamiques individuelles.

L'objectif principal de notre travail collaboratif avec Özgür Asar est d'étudier les performances de prédictions dynamiques en présence de valeurs extrêmes. Nous nous demandons à quel point la présence de valeurs extrêmes peut affecter les prédictions dynamiques. Pour cela, nous considérerons un modèle conjoint robuste pour données longitudinales et données de survie avec des distributions t pour les effets aléatoires et erreurs résiduelles du modèle. Les propriétés de calibration et discrimination obtenues à partir d'un modèle conjoint classique et d'un modèle conjoint robuste seront comparées. Le modèle conjoint robuste proposé est un modèle original : il suppose l'indépendance des distributions t pour les effets aléatoires et erreurs résiduelles, les erreurs résiduelles sont également supposées indépendantes pour un même sujet. L'objectif principal de notre travail collaboratif avec Özgür Asar est d'étudier les performances de prédictions dynamiques en présence de valeurs extrêmes. Nous nous demandons à quel point la présence de valeurs extrêmes peut affecter les prédictions dynamiques. Pour cela, nous considérerons un modèle conjoint robuste pour données longitudinales et données de survie avec des distributions t pour les effets aléatoires et erreurs résiduelles du modèle. Les propriétés de calibration et discrimination obtenues à partir d'un modèle conjoint classique et d'un modèle conjoint robuste seront comparées. Le modèle conjoint robuste proposé est un modèle original : il suppose

l'indépendance des distributions t pour les effets aléatoires et erreurs résiduelles, les erreurs résiduelles sont également supposées indépendantes pour un même sujet.

9.3.1 Modèle conjoint pour données longitudinales et données de survie

Notations

L'estimation d'un modèle conjoint pour données longitudinales et données de survie peut se faire à partir d'un échantillon de n sujets identiquement et indépendamment distribués pour lesquels on observe les données $\{T_i, \delta_i, Y_i, t_i, X_i; i = 1, ..., n\}$ où $Y_i = \{Y_{ij}; j, ..., n_i\}$ est un marqueur quantitatif longitudinal mesuré au temps $t_i = \{t_{ij}; j = 1, ..., n_i\}$, X_i est une matrice de variables explicatives, T_i est le délai de survenue de l'événement. Ce délai est classiquement censuré à droite avec $T_i = \min(T_i^*, C_i)$ où C_i est un délai de censure supposé indépendant. Enfin, $\delta_i = \mathbb{1}(T_i^* \leq C_i)$ est la fonction indicatrice de survenue de l'événement.

Dans ce travail, nous considérons un modèle conjoint à effets aléatoires partagés [Wulfsohn and Tsiatis, 1997] dont la formulation est basée sur le lien entre le processus longitudinal et le processus de survie. Ce modèle peut être défini ainsi :

$$\begin{cases} Y_{ij} = Y_{ij}^* + \epsilon_{ij} = X_{Yij}^\top \beta + Z_{ij}^\top b_i + \epsilon_{ij} & (9.1) \\ \lambda_i(t) = \lambda_0(t) \exp(X_{Ti}^\top \alpha + g\left(b_i, t, \eta\right)) & (9.2) \end{cases}$$

L'équation 9.1 correspond au modèle linéaire mixte pour le processus longitudinal. Les valeurs observées du marqueur longitudinal, Y_{ij}, sont des mesures bruitées d'un processus latent, Y_{ij}^*, avec ϵ_{ij} l'erreur résiduelle ou erreur de mesure. $X_{Yij} = \{X_{ij1}...X_{ijp}\}^\top$ est une matrice d'éléments de X_i et de temps t_i, Z_{ij} une sous-matrice d'éléments de X_{Yij}. β et b_i sont respectivement les vecteurs de paramètres populationnels et de variables latentes spécifiques au sujet.

L'équation 9.2 correspond au modèle de survie avec une fragilité dépendante du temps. $\lambda_i(t)$ est le risque instantané de connaître l'événement au temps t avec $\lambda_0(t)$ le risque de base pouvant être défini d'après une distribution de paramètres notés ζ. $X_{Ti} = \{X_{i1}...X_{ir}\}^\top$ est une matrice d'éléments de X_i et α est un vecteur de paramètres associés. $g\left(b_i, t, \eta\right)$ est une fonction définissant la structure de dépendance entre le temps de survenue de l'événement et le processus longitudinal. Plusieurs structures de dépendance ont déjà été proposées dans la littérature [Rizopoulos, 2012]. Les exemples les plus fréquemment rencontrés comprennent une dépendance sur la valeur courante du marqueur, i.e. $g\left(b_i, t, \eta\right) = \eta Y_i(t)^*$, ou sur la valeur courante du marqueur et son intensité d'évolution, i.e. $g\left(b_i, t, \eta\right) = \eta_1 Y_i(t)^* + \eta_2 \frac{\partial Y_i(t)^*}{\partial t}$. Cette dernière est celle que nous avions retenue dans Fournier et al. [2016] et dans le modèle conjoint simplifié utilisé pour développer des prédictions dynamiques.

Hypothèses sur les distributions

Les modèles conjoints classiques supposent b_i et ϵ_{ij} distribués normalement ($b_i \sim \mathcal{MVN}(0, \Sigma)$ et $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$). Les effets aléatoires b_i sont classiquement considérés indépendants des erreurs résiduelles, i.e. $b_i \perp \epsilon_{ij}$, les erreurs résiduelles sont également supposées indépendantes entre elles, i.e. $\epsilon_{ij} \perp \epsilon_{ij'}$ pour $j \neq j'$.

Comme évoqué en introduction, l'hypothèse de normalité peut être trop restrictive dans certaines applications, par exemple en présence de valeurs extrêmes. Pinheiro et al. [2001] proposent deux types d'observations longitudinales extrêmes : i) les sujets extrêmes, ii) et les observations ponctuelles extrêmes d'un sujet. Les sujets extrêmes pourraient par exemple comprendre des sujets avec un niveau initial du marqueur extrêmement élevé/faible et/ou avec une rapide évolution au cours du temps. En d'autres termes, les sujets extrêmes correspondent à des valeurs extrêmes d'effets aléatoires b_i. Les observations ponctuelles extrêmes correspondent à quelques observations pour un sujet qui sont très différentes du reste des observations de ce même sujet. En d'autres mots, ces observations correspondent à des valeurs extrêmes d'erreurs résiduelles ϵ_{ij}. A titre d'illustration, nous avons simulé un échantillon de 30 sujets à partir d'un modèle conjoint robuste dont l'évolution du marqueur est supposée d'intercept nul et de pente nulle (Figure 9.6). Comme attendu, la plupart des sujets ont une évolution homogène du marqueur, alors que quelques sujets semblent présenter des trajectoires ou observations extrêmes. Par exemple, les sujets 10, 26 ou encore 27 présentent une pente d'évolution relativement forte comparativement aux autres sujets ; concernant le sujet 28, trois observations ponctuelles semblent s'écarter de la trajectoire individuelle.

Nous développerons un modèle conjoint robuste pour lequel nous relâcherons les distributions normales des effets aléatoires b_i et erreurs résiduelles ϵ_{ij}. Nous considérerons des distributions t en supposant $b_i = \sqrt{V_i} b_i^*$ et $\epsilon_{ij} = \sqrt{W_{ij}} \sigma \epsilon_{ij}^*$ où V_i et W_{ij} sont des variables aléatoires définies selon des distributions Inverse Gamma avec $V_i \sim \mathcal{IG}(\phi/2, \phi/2)$ et $W_{ij} \sim \mathcal{IG}(\nu/2, \nu/2)$, et les propriétés suivantes, $V_i \perp W_{ij}$, $W_{ij} \perp W_{ij'}$ pour $j \neq j'$, $b_i^* \sim \mathcal{MVN}(0, \Sigma)$ et $\epsilon_{ij}^* \sim \mathcal{N}(0, 1)$. D'après ces précédentes définitions, nous retrouvons ainsi $b_i \sim \mathcal{MV}t(0, \Sigma, \phi)$ et $\epsilon_{ij} \sim t(0, \sigma^2, \nu)$ avec les propriétés suivantes $b_i \perp \epsilon_{ij}$ et $\epsilon_{ij} \perp \epsilon_{ij'}$ pour $j \neq j'$. ϕ et ν sont les degrés de libertés des distributions t. Rappelons que si ces paramètres convergent vers l'infini, les distributions t convergent vers des lois normales.

A notre connaissance, aucun travail dans la littérature en Biostatistique ne considère un tel modèle conjoint robuste avec les propriétés évoquées ci-dessus. Li et al. [2009], Huang et al. [2010] et Taylor et al. [2013] considère un modèle conjoint robuste avec une hypothèse de normalité pour les effets aléatoires b_i et avec une distribution t avec un degré de liberté fixé pour l'erreur résiduelle. Kim and Albert [2016] considèrent une distribution Normale pour les effets aléatoires b_i et une distribution t généralisée pour l'erreur résiduelle. Enfin, Baghfalaki et al. [2013] et Baghfalaki et al. [2014] considèrent un modèle similaire à celui que nous proposons mais avec la

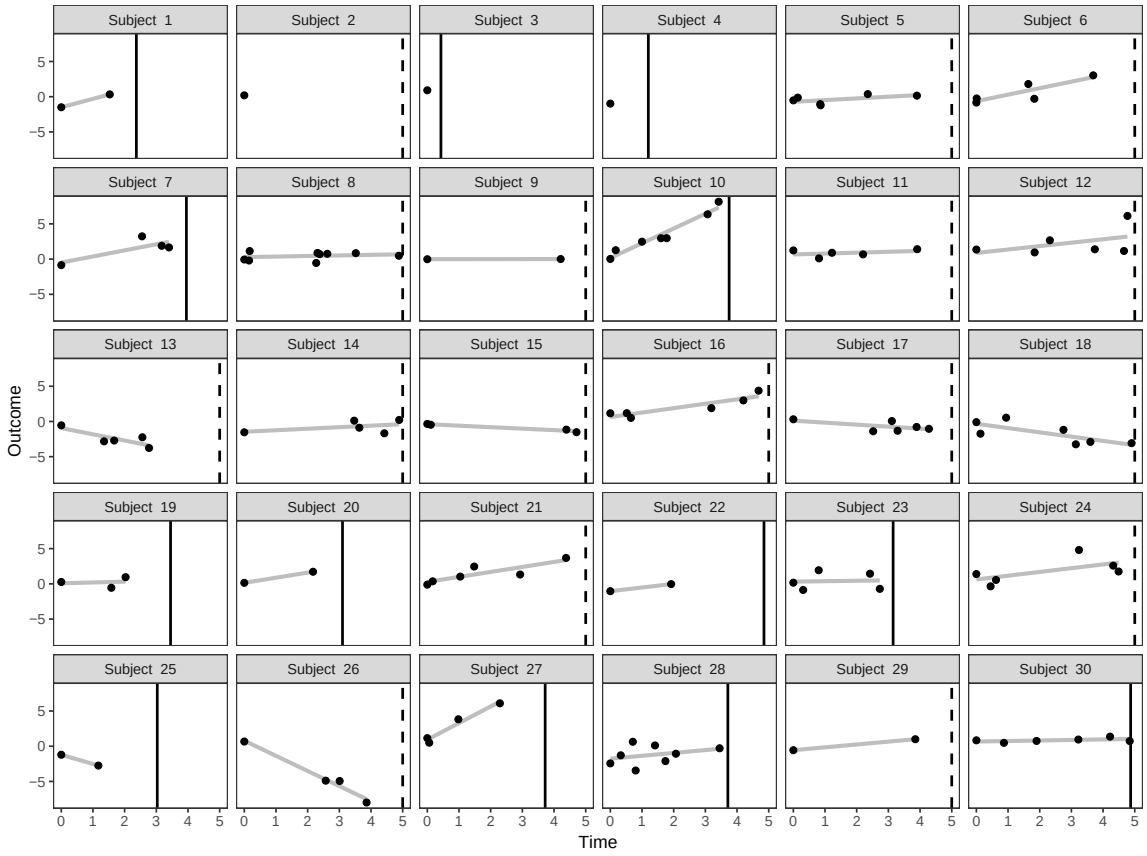


Figure 9.6: *Simulation de 30 sujets à partir d'un modèle conjoint robuste (Intercept et pente d'évolution du marqueurs supposés égaux à 0).*

contrainte $W_{ij} = W_i$ conduisant à une dépendance des erreurs résiduelles.

Inférence bayésienne

La distribution jointe a posteriori des paramètres et des variables latentes peut être définie ainsi :

$$\begin{aligned} f(\beta, \Sigma, \phi, \sigma, \nu, \zeta, \alpha, \eta, B, V, W | Y, T, \delta, X_Y, Z, X_T) \propto & \; f(Y|\beta, \sigma, B, W, X_Y, Z) \\ & \times f(T, \delta|\zeta, \alpha, \eta, \beta, B, X_Y, Z, X_T) \\ & \times f(B|V, \Sigma) f(V|\phi) f(W|\nu) \\ & \times f(\beta, \Sigma, \phi, \sigma, \nu, \zeta, \alpha, \eta) \end{aligned} \tag{9.3}$$

où $f(\cdot)$ est la notation générique pour une fonction de densité. L'inférence des modèles conjoints considérés (classique et robuste) sera réalisée à l'aide d'une approche bayésienne pour l'estimation des paramètres des modèles et pour l'estimation des prédictions dynamiques, Özgür Asar ayant une bonne expérience de ce type de méthodes. Des échantillonnage seront réalisés à partir de la distribution a posteriori jointe des processus d'après la méthode *No-U-Turn Sampler (NUTS)* [Hoffman and Gelman, 2011], une version adaptée de l'algorithme *Hamiltonian Monte Carlo (HMC)* [Brooks et al., 2011].

9.3.2 Application au pronostic en transplantation rénale

Afin d'évaluer le possible impact de valeurs extrêmes sur les prédictions dynamiques de l'échec de greffe rénale, nous étudierons les performance prédictives du modèle conjoint robuste à partir des données de la cohorte DIVAT utilisées dans le dernier travail de thèse de Marie-Cécile Fournier. Sur la même cohorte d'apprentissage (N=2749), nous estimerons tout d'abord le même modèle conjoint classique mais avec une inférence bayésienne. Cela pourrait notamment conduire à avoir une plus grande précision des estimations et ainsi réduire la variabilité autour des prédictions dynamiques, ce qui reste pour l'instant une limite importante de l'approche que nous avons présentée en section 9.2. Nous pourrons ainsi comparer les estimations obtenues à l'aide des deux approches (fréquentiste et bayésienne). Puis, toujours sur l'échantillon d'apprentissage, nous estimerons un modèle conjoint robuste considérant des distributions t pour les effets aléatoires et erreurs résiduelles. Les covariables considérées seront celles retenues dans le travail de Marie-Cécile Fournier. Nous étudierons plus particulièrement à quel point considérer les valeurs extrêmes avec des distributions plus appropriées pourrait améliorer les prédictions dynamiques. Pour cela, nous comparerons les performances en termes de calibration et de discrimination issues des deux approches (modèles conjoints classique et robuste). Nous reproduirons ainsi les graphiques de calibration et de courbes de R^2, d'aires sous la courbe ROC (Figures 9.1 et 9.2). Enfin, l'application en ligne fournissant les prédictions dynamiques pourra être mise à

jour afin de prendre en considération les développements méthodologiques présentés ici.

9.4 Prédiction dynamiques : prise en considération du délai entre deux visites de suivi

Le descriptif des données de la cohorte DIVAT utilisées dans ce travail, nous a conduit à soulever une hypothèse que nous souhaiterions approfondir. Dans la cohorte DIVAT, les données des patients sont normalement collectées à chaque date anniversaire de la transplantation jusqu'à un échec de greffe (retour en dialyse, retransplantation préemptive, ou décès du patient avec greffon fonctionnel) ou une date de censure. En regardant les dates de mesures de sérum de créatinine en fonction du statut de survenue de l'événement (Figure 9.7), nous pouvons voir que, chez les patients censurés au cours de leur suivi, nous retrouvons approximativement ces dates de collecte de données. En revanche, la répartition des dates de mesures chez les patients connaissant un échec de greffe ne paraît pas définie selon le schéma de suivi évoqué. Cela pose la question de l'informativité de ces délais de mesure. Notre hypothèse est que les patients en moins bonne santé pourraient venir plus précocement à l'hôpital. Les délais entre deux visites pourraient apporter une information indirecte sur l'état de santé du patient et s'avérer distribués non aléatoirement. Ces délais pourraient ainsi contribuer à améliorer les performances prédictives de l'outil proposé jusqu'à présent.



Figure 9.7: *Mesures observées du log de sérum de créatinine en fonction du délai écoulé (en années) avant la date de dernier suivi (Censored subjects) ou avant la date d'échec de greffe rénale (Graft failure) sur l'échantillon d'apprentissage (n=2749). Le délai est exprimé avec la date de dernier suivi comme origine de temps. Le trait plein représente l'évolution moyenne estimée à partir d'un modèle linéaire généralisé pénalisé intégrant une fonction de lissage [Wood, 2004].*

Dans ce projet, nous pensons modéliser simultanément 3 processus à l'aide d'un modèle conjoints à effet aléatoires partagés : i) le processus longitudinal, ii) le processus de dates de visite,

iii) le processus de survenue d'un échec de greffe. Inspiré d'un travail proposé en cancérologie par Król et al. [2016], ce modèle conjoint pourrait être défini ainsi :

$$\begin{cases} Y_{ij} = Y_{ij}^* + \epsilon_{ij} = X_{Yij}^\top \beta + Z_{ij}^\top b_i + \epsilon_{ij} & (9.4) \\ \lambda_{ij}^S(s) = u_i \lambda_0^S(s) \exp(X_{Sij}^\top \alpha_1) & (9.5) \\ \lambda_i^T(t) = \lambda_0^T(t) \exp(X_{Ti}^\top \alpha + u_i \eta_1 + g_2(b_i, t, \eta_2)) & (9.6) \end{cases}$$

L'équation 9.4 correspond comme précédemment au modèle linéaire mixte pour le processus longitudinal. L'équation 9.5 correspond à un modèle de survie avec fragilité pour le délai S entre les visites de suivi. L'effet aléatoire u_i permet de prendre en compte la dépendance des délais entre visites pour un sujet donné. Plusieurs distributions pour u_i sont envisageables et la distribution Gamma est un choix fréquemment rencontré [Hougaard, 1995]. Enfin, l'équation 9.6 correspond au modèle de survie pour le risque de survenue d'un échec de greffe rénale. Ce risque dépend comme précédemment d'une fragilité dépendante du temps liée à l'évolution du sérum de créatinine, mais également de la spécificité individuelle du risque de survenue d'une visite.

Selon les résultats du travail présenté en section 9.3, nous considérerons des effets aléatoires et erreurs résiduelles à l'aide de distributions appropriées. Ce modèle devra être implémenté dans un package R. Nous validerons l'estimation des paramètres d'un tel modèle à partir d'une étude de simulation. Nous étudierons les performances d'un outil de prédictions dynamiques à partir de ce modèle sur les données de la cohorte DIVAT afin de voir si la prise en compte des délais de visites au cours du suivi les améliorent.

- Chapitre 10 -

Modélisation en réanimation

10.1 Revue de la littérature : Pronostic en réanimation

Le travail que j'ai récemment débuté dans le contexte de la réanimation se développe dans le cadre de la thèse de sciences de Jeanne Simon-Pimmel que j'encadre avec Yohann Foucher et Denis Frasca. Nous avons commencé par un travail de revue de la littérature des outils pronostiques en réanimation. Plus précisément, nous menons simultanément deux revues : l'une concernant les traumatismes crâniens (TC), l'autre concernant les accidents vasculaires cérébraux hémorragiques (AVC) et hémorragies méningées (HM), trois types d'agressions cérébrales touchant particulièrement une population jeune et partageant les mêmes types de complications.

La principale problématique en réanimation est liée au mauvais pronostic des patients. En effet, l'un des enjeux majeurs est le risque élevé de mortalité, de handicap sévère ou de survenue d'un état végétatif. L'incidence du traumatisme crânien est de près de 258 patients pour 100000 individus par an [Brazinova et al., 2016], avec près de 4% considérés comme sévères [Tagliaferri et al., 2006]. L'incidence de survenue d'hémorragie intracérébrale est de près de 25 patients pour 100000 individus par an avec un taux de mortalité de près de 40% à 1 mois [van Asch et al., 2010]. Le défi pour les soignants et personnes de confiance du patient est de prendre la difficile décision de poursuivre ou non les soins spécialisés en fonction du possible devenir du patient [Hamel et al., 1995]. A la difficulté médicale, s'ajoute une problématique éthique. La qualité de vie du patient en état végétatif ou atteint d'un handicap sévère, ainsi que celle de son entourage, sont en effet inévitablement altérées et la question du bien fondé de la réanimation prolongée peut se poser en fonction de la gravité de l'atteinte cérébrale [Gillett et al., 2010].

A l'ère de la médecine 4P (prédictive, préventive, personnalisée et participative) [Flores et al., 2013], la compréhension des patients et de l'entourage et leur considération pour les conséquences possibles d'un traitement ou d'une intervention est primordiale. Dans ce contexte,

le développement d'outils pronostiques présente un véritable intérêt dans la prise en charge réanimatoire des agressions cérébrales. L'objectif réside dans la capacité de ces outils à pronostiquer correctement le devenir possible des patients (décès, état fonctionnel dégradé, etc.). Ces outils pourraient aider à une meilleure décision médicale partagée relative à la question de la poursuite des soins ou non. Le groupe PROGRESS a publié en 2013 une série d'articles de recommendations sur le développement, la validation et l'utilisation d'outils pronostiques à des fins de médecine personnalisée [Hemingway et al., 2013; Riley et al., 2013; Steyerberg et al., 2013; Hingorani et al., 2013]. En réponse au constat accablant de la faiblesse méthodologique de bons nombres d'articles scientifiques présentant des outils/scores comme étant pronostiques, les recommendations TRIPOD ont récemment été publiées afin de contribuer à l'amélioration de la qualité des informations reportées concernant des modèles de prédiction [Collins et al., 2015]. L'objectif de ce travail de revue de la littérature est de recenser les outils pronostiques développés dans le contexte de la réanimation médicale et portant sur 1) les TC et 2) les AVC/HM. Ce travail permettra de mieux positionner les futurs développements prédictifs de Jeanne Simon-Pimmel.

10.2 La cohorte AtlanREA

Plusieurs projets autour du pronostic en réanimation pourront être envisagés et appliqués aux données de la cohorte AtlanREA (`www.atlanrea.org`). Il s'agit d'une cohorte observationnelle prospective de patients avec des lésions cérébrales sévères (TC ou AVC), hospitalisés et intubés dans un service de réanimation. AtlanREA est une cohorte multicentrique française composée aujourd'hui de 5 centres hospitaliers (Nantes, Poitiers, Angers, Rennes et Tours). Depuis l'initiation du réseau en mars 2013, 2750 patients ont été inclus dans la cohorte. Les patients sont suivis depuis l'admission jusqu'à la sortie du service de réanimation ou la survenue d'un décès. AtlanREA contient de nombreuses variables cliniques et biologiques (âge, genre, score SAPS II, score SOFA, score de Glasgow pour les 24 premières heures, etc.). Des données sont collectées au cours du suivi comme la pression intracrânienne du patient. Certaines données sont également collectées au-delà de la sortie de réanimation comme le score de Glasgow (Glasgow Outcome Scale, GOS) à 3 mois post-entrée en réanimation, ou encore le questionnnaire de qualité de vie EQ-5D.

10.3 Modélisation de la survie nette

L'urgence médicale rencontrée en réanimation et la gravité de l'état de santé des patients conduisent les cliniciens à être essentiellement intéressés par des critères de jugement à court terme. Des résultats préliminaires de notre revue de la littérature nous permettent d'identifier les critères de jugement les plus fréquemment rencontrés dans les études. Sur les 132 articles

inclus dans notre revue, 107 (82%) présentaient un critère de mortalité dont 47 (36%) portaient sur la mortalité en service de réanimation, 16 (12%) sur la mortalité entre 1 et 3 mois post-sortie, 29 (22%) sur la mortalité à 6 mois. Il apparaît évident que la priorité clinique est la survie du patient. Nous pouvons remarquer qu'un seul article s'intéressait au devenir à très long terme du patient.

Pour les patients ayant survécus après un séjour en réanimation, la question de la récupération fonctionnelle, cognitive et psychique se pose. Dans ce contexte, nous pensons qu'il serait intéressant de pronostiquer le devenir à long terme des patients vivants à la sortie de réanimation et présentant un niveau de séquelle fonctionnelle ayant un potentiel de récupération. Les critères d'inclusion d'une telle étude pourraient être : patients adultes ayant subi un TC (ou un AVC), ayant réalisés un séjour en service de réanimation, vivants avec un score de Glasgow de 3 ou plus (non inclusion des patients en état végétatif). Le score de Glagow est un score calculé chez des patients cérébrolésés composé de 5 catégories (Tableau 10.1) [Jennett and Bond, 1975].

Table 10.1: *Description des catégories du Glasgow Outcome Scale (GOS).*

Niveau de GOS	Etat du patient
1	Décès
2	Etat végétatif persistant (absence d'activité corticale)
3	Handicap sévère (conscient mais dépendant : atteinte mentale ou motrice ou les deux)
4	Handicap modéré : patient autonome dans la vie quotidienne (dysphasie, hémiparésie, ataxie, troubles intellectuels ou de mémoire, troubles de la personnalité)
5	Retour à la vie normale (déficits neurologiques ou psychologiques mineurs)

Les patients de la cohorte AtlanREA sont suivis jusqu'à la sortie du service de réanimation ou la survenue d'un décès en réanimation. Afin d'étudier le risque de survenue d'un décès à long terme chez les patients ayant séjourné en service de réanimation, il est nécessaire de les suivre bien au-delà de ce temps d'horizon. Nous avons entamé des démarches adminitratives auprès du Centre d'épidémiologie sur les causes médicales de Décès (CépiDc) afin d'obtenir les dates de décès des patients d'AtlanREA. Nous serons ainsi en mesure d'étudier la survie des patients avec un GOS 3-5 en sortie de réanimation. Nous étudierons les facteurs de risque associés à la survenue d'un décès à long terme.

Nous envisageons également de poursuivre ces recherches à l'aide de modèles de survie relative. Ces modèles présentent des concepts de modélisation qui pourraient s'avérer pertinents dans les problématiques du droit à l'oubli pour des patients ayant subi une atteinte cérébrale. Les modèles de survie relative ont été initialement développés pour prendre en considération la mortalité spécifiquement liée à la pathologie étudiée [Esteve et al., 1990]. Lorsqu'il sont à risques additifs (leur formulation la plus connue), ils permettent d'estimer la survie nette, i.e. la survie dans le cas où les seules causes de décès sont liées à la maladie. La survie observée est corrigée par la survie attendue estimée à partir d'un groupe contrôle. Ces modèles sont de plus en plus utilisés notamment en cancérologie. En appliquant ce type d'approche au contexte de la

réanimation, nous pourrions étudier l'excès de risque de mortalité dû à un séjour en réanimation. L'idée serait de supprimer une part de la mortalité indépendante de la maladie à partir de tables de mortalité de la population générale. Moins répandus, les modèles de survie relative à risques multiplicatifs [Breslow and Day, 1975; Andersen et al., 1985] permettent aussi de comparer l'effet de facteurs de risque entre un groupe d'intérêt et un groupe de référence. L'un des avantages de cette approche est de ne pas imposer le risque instantané observé $\lambda_O(t)$ à être supérieur au risque instantané attendu dans la population de référence $\lambda_A(t)$.

Les modèles de survie relative pourraient être une approche méthodologique réellement intéressante dans les problématiques de droit à l'oubli. Du point de vue du patient et de son entourage, il nous apparaît important de pouvoir apporter des informations quant à leur potentiel de récupération. Notre hypothèse est qu'un patient ayant eu un TC (ou AVC) sévère puisse possiblement retrouver petit à petit un risque de décès proche de celui de la population générale. L'objectif serait alors d'estimer le délai de guérison ou le délai de retour à un risque normal, i.e. délai pour lequel le risque observé tend vers le risque attendu dans la population de référence.

Plusieurs travaux méthodologiques se sont intéressés à cette problématique [De Angelis et al., 1999; Yu et al., 2004]. En particulier, les modèles de survie relative à risques additifs ont déjà été utilisés dans cette démarche [Andersson et al., 2011; Boussari et al., 2018]. L'idée générale de cette approche est d'estimer le délai auquel le risque instantané en excès lié à la pathologie est proche de 0. Les modèles de survie relative à risques multiplicatifs pourraient également être exploités dans l'optique d'estimer le délai de guérison. Plus précisément, il s'agirait d'estimer le temps auquel le risque instantané relatif à la pathologie $\lambda_R(t)$ est proche de 1. Nous souhaiterions explorer de manière plus avancée l'utilisation d'une telle approche dans cette problématique de droit à l'oubli. Nous pensons comparer les approches par modèles à risques additifs et à risques multiplicatifs à partir d'une étude de simulation et les appliquer dans le contexte de la réanimation à partir des donnée de la cohorte AtlanREA.

Dans le contexte de la réanimation, plusieurs contraintes devront probablement être prises en compte dans la modélisation. Il est probable que certains patients graves ne récupèrent jamais, que certains patients moins graves récupèrent très lentement et que d'autres se retrouvent avec un excès de risque lié au séjour en réanimation qui soit nul. Cette possible hétérogénéité des excès de risque devra être prise en compte dans la modélisation. Les modèles de survie à fraction de personnes guéries (i.e. cured survival model [Farewell, 1982; Shao and Zhou, 2004]) pourraient permettre de tenir compte de cette hétérogénéité [Andersson et al., 2011].

10.4 GOS pondéré

Dans les études en réanimation, plusieurs critères d'intérêt sont de nature ordinale. Le Rankin modifié (modified Rankin Scale, mRS) est une des échelles les plus utilisées pour étudier les

patients ayant subi un AVC [Lees et al., 2012]. Allant de 0 (aucun symptôme) à 6 (décès), ce score mesure le degré de handicap ou de dépendence dans la vie courante. Le score de GOS (ou sa version étendue GOS-E [Teasdale et al., 1998]), au travers d'une échelle ordinale allant de 1 (décès) à 5 (vie normale), catégorise la sévérité du handicap chez les patients ayant subi un TC (voir la table 10.1) [Jennett and Bond, 1975]. Très souvent, un tel critère de jugement est étudié après dichotomisation de la variable (favorable versus non favorable). Mais les analyses de ce nouveau critère binaire peuvent conduire à des conclusions erronées du fait de la perte d'information induite par la dichotomisation. La puissance pour détecter un effet du traitement est réduite lorsque l'échelle est dichotomisée [Altman and Royston, 2006]. L'étude d'une telle variable dans son échelle ordinale à l'aide par exemple d'un modèle des côtes proportionnelles présente certains avantages et a fait l'objet de nombreux articles [Lees et al., 2012; Optimising Analysis of Stroke Trials (OAST) Collaboration et al., 2007; McHugh et al., 2010; Roozenbeek et al., 2011].

Toutefois, ces critères de fonctionnalité n'intègrent pas de dimensions de qualité de vie. Les états de santé favorable et non favorable devraient effectivement être définis du point de vue du patient (et de son entourage) en intégrant notamment la considération de leur état de santé. La considération des critères de qualité de vie dans l'analyse d'essais cliniques portant sur les AVC est de plus en plus reconnue [Carod-Artal and Egido, 2009; Ali et al., 2013]. Les mesures d'utilité permettent de qualifier la préférence des patients pour les différents états de santé envisagebles [Feeny, 2000]. Ainsi, Chaisinanunkul et al. [2015] définissent une nouvelle échelle dans le contexte de l'AVC : le mRS pondéré par l'utilité correspondante (Utility-Weighted modified Rankin Scale, UW-mRS). Les auteurs réanalysent les résultats de différents essais cliniques en comparant les approches par dichotomisation, analyse de la réponse ordinale du mRS et analyse du UW-mRS pour la détection d'un effet traitement. Dijkland et al. [2018] complètent cette première proposition à partir d'une étude de simulation. Les auteurs concluent à des meilleures performances du score UW-mRS intégrant les préférences des patients pour détecter un effet traitement comparativement aux approches étudiant le critère dichotomisé. En revanche, à partir de leur étude de simulation, ils soulignent que l'analyse du score UW-mRS ne semble pas plus efficace que celle du score mRS : la puissance pourrait être inférieure à celle obtenue dans le cas de l'analyse du score ordinal. Dans ces deux travaux, une limite majeure est la non considération de la variabilité individuelle des utilités. En effet, pour chaque modalité du mRS, le score est pondéré par la moyenne des utilités des patients ayant la modalité considérée. Ceci ne peut donc pas représenter correctement la variabilité autour de la perception des états de santé. Par ailleurs, les auteurs de ces deux travaux suggèrent que le score UW-mRS puissent être utilisé dans le calcul de gain/perte de QALYs.

Nous envisageons de repartir de cette idée pour pour proposer un score UW-GOS chez les patients ayant subi un TC. Dans la cohorte AtlanREA, le score GOS et le score EQ-5D sont recueillis à 3 mois post-sortie de réanimation. Il nous sera possible d'étudier un GOS pondéré par l'utilité après conversion des mesures d'EQ-5D. Plutôt que de considérer la moyenne des utilités

pour chacune des modalités du GOS, nous calculerons le GOS d'un patient pondéré directement par sa mesure d'utilité individuelle. Nous mettrons en oeuvre une étude de simulation pour démontrer de l'intérêt de notre approche comparativement aux approches traditionnelles.

- Chapitre 11 -

Modèle conjoint pour l'évolution longitudinale d'un score d'utilité et du risque de survenue d'un événement

11.1 Introduction

Dans de nombreuses études de recherche clinique, les principaux critères de jugement sont basés sur la survie des patients. Par exemple, en cancérologie, les résultats d'essais cliniques sont essentiellement présentés selon la survie libre de métastases, la survie libre de progression de la maladie, ou la survie globale. L'intervention la plus efficace caractérisée par un gain de survie pourrait toutefois ne pas être la meilleure du point de vue du patient si ce gain implique une importante déterioration de la qualité de vie due aux effets secondaires du traitement [Basch, 2013]. Afin de mieux considérer la voix des patients face aux options thérapeutiques qui s'offrent à eux, les recherches centrées sur le patient méritent un plus grand intérêt [Basch, 2013; Methodology Committee of the Patient-Centered Outcomes Research Institute (PCORI), 2012]. Même si des recommandations existent, les critères de jugement liés à la qualité de vie des patients sont encore trop rarement considérés dans les essais cliniques et les cohortes observationnelles. Lorsque des critères de survie et de qualité de vie sont collectés, leurs analyses respectives sont souvent réalisées séparément. Dans une mini-revue de la littérature que nous avons menée dans le contexte du cancer de la prostate, aucune étude ne décrivait précisément le possible arbitrage entre quantité et qualité de vie [Foucher et al., 2018]. Nous recommandions que cet arbitrage entre survie et qualité de vie puisse être étudié au travers d'un unique critère de jugement, par exemple le nombre d'années ajustées par la qualité de vie (QALYs) [Weinstein et al., 2009]. Les QALYs sont une mesure des états de santé bien connue, initialement désignés pour l'évaluation économique de programmes de santé [Weinstein, 1980]. Leur nature composite les rend aussi très intéressants pour l'aide à la prise de décision clinique [Kind et al., 2009]. Plus précisément, le principe sous-jacent des QALYs est de convertir une succession d'états de santé au cours du temps en un nombre équivalent d'années de vie vécues en parfaite santé. Les QALYs sont

calculés en assignant à chaque état de santé du patient une mesure de qualité de vie, appelée score d'utilité, allant de 0 (décès) à 1 (parfaite santé) de manière à ce que chaque année de vie soit pondérée par l'utilité correspondante. Un score d'utilité élevé indique un état de santé préferable. Un QALY représente une année de vie en parfaite santé. Par exemple, comme illustré sur la Figure 11.1, un patient vivant 10 ans avec un score d'utilité de 0.8 équivaut à 8 QALYs, i.e. 8 années vécues en parfaite santé. Cette valeur serait plus faible pour un patient vivant 12 ans mais avec un score d'utilité de 0.6, le nombre de QALYs étant alors de 7.2 du fait d'un traitement plus efficace mais avec plus d'effets secondaires sur la qualité de vie.

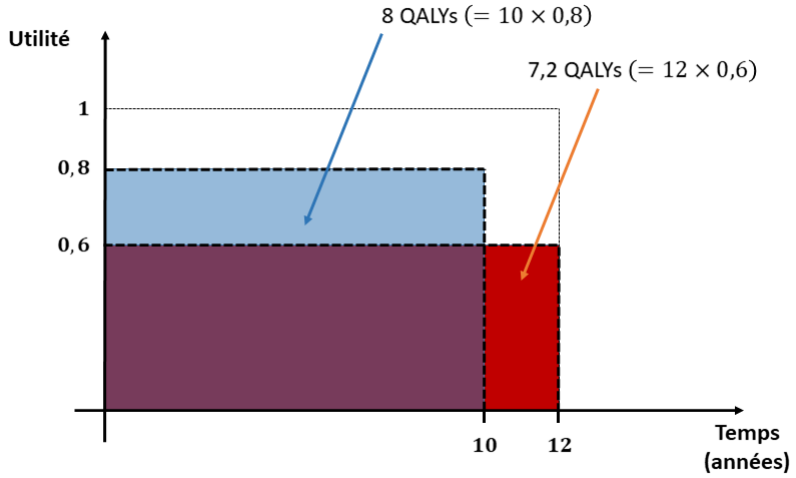


Figure 11.1: *Illustration du calcul du nombre espéré de QALYs.*

Dans un récent travail, nous avons eu pour objectif de proposer un cadre méthodologique pour une prise de décision médicale stratifiée centrée sur le patient. Plus précisément, nous avons proposé une fonction d'utilité dépendante du temps pour données censurées définie à partir de la distribution jointe d'une décision médicale entre deux traitements et du temps de survenue d'un événement pondéré par l'utilité du traitement [Dantan et al., 2018] (voir le chapitre 7 pour plus de détails). Cette méthodologie permet l'estimation du nombre d'année de vie vécues en parfaite santé dans une population. En maximisant cette fonction d'utilité, nous estimons la meilleure situation médicale entre deux traitements possibles. Deux principales limites doivent être soulignées. Premièrement, nous avons utilisé des scores d'utilité issus de la littérature. Il aurait été préférable de collecter les données individuelles afin de prendre en considération la variabilité individuelle des scores d'utilité au cours du temps. Les scores d'utilité auraient pu être mesurés en interrogeant les patients. Pour cela, plusieurs méthodes existent comme celle du *Standard Gamble*, du *time trade-off* [Kind et al., 2009], ou en utilisant des questionnaires d'utilité des états de santé tels que l'*Euroqol EQ-5D* [Towse, 2010]. Deuxièmement, nous supposions que le score d'utilité pour un état de santé chronique était constant toute la durée de cet état de santé et identique pour les patients dans chacun des bras de traitement, alors que la qualité de vie est, par essence, un processus évoluant de manière continue. Il s'agit d'une variable endogène, i.e. une variable générée par le patient lui-même [Kalbfleisch and Prentice, 2011]. En conséquence,

l'évolution du score d'utilité est un processus lui-même continu qui a une variabilité inhérente entre les patients.

D'autres approches méthodologiques ont déjà abordées la question de l'analyse de survie ajustée par la qualité de vie au moyen de mesures longitudinales d'utilité. En 1989, Goldhirsch et al. ont introduit le concept de Q-TWiST (Quality-adjusted Time Without Symptoms or Toxicity) dans le contexte du cancer du sein [Goldhirsch et al., 1989]. Les auteurs remettent en question l'utilisation d'une chimiothérapie adjuvante chez les patientes ménopausées atteintes d'un cancer du sein à un stade précoce puisque les avantages en termes de gain de survie pourraient être compensés par la toxicité du traitement. Dans cette perspective, ils ont proposé une approche originale consistant à répartir le suivi des patients sur plusieurs états de santé pour lesquels la qualité de vie correspondante diffère. Leur proposition initiale a ensuite été généralisée dans un cadre non paramétrique par Glasziou et al. [1998]. Une hypothèse fondamentale porte sur le fait que, pour un état de santé donné, le nombre de QALYs se décompose comme le produit de l'utilité $U(t)$ pour un état de santé au temps t et de la proportion de patients vivants $S(t)$ au temps t :

$$QALY(\tau) = \int_0^\tau S(t) \times U(t)dt$$

Ainsi, l'idée des Q-TWiST est de considérer une séquence d'états de santé dans le suivi du patient pondérée par un score d'utilité correspondant. L'espérance de vie corrigée en fonction de la qualité de vie peut donc être estimée comme suit : $\widehat{Q} = \sum_s D_s U_s$ avec D_s la durée passée dans l'état s et U_s l'utilité supposé constante sur la durée de cet état. Plus concrètement, Goldhirsch et al. [1989] considèrent la toxicité du traitement, la récidive de maladie et le décès avec T_1 le temps de traitement, T_2 le temps de récidive de maladie et T_3 le temps de décès. En conséquence, différentes durées peuvent être définies : $D_{TOX} = T_1$ la durée de la toxicité du traitement, $D_{TWIST} = T_2 - T_1$ la durée sans symptôme ou toxicité et $D_{REL} = T_3 - T_2$ la durée de vie après récidive. Pour chaque groupe de traitement $x \in \{1, 2\}$, l'espérance de vie ajustée par la qualité de vie peut être définie par

$$Q_x = u_{TOX,x} \int_0^\tau \hat{S}_{1,x}(t)dt + \int_0^\tau \hat{S}_{2,x}(t)dt - \hat{S}_{1,x}(t)dt + u_{REL,x} \int_0^\tau \hat{S}_{3,x}(t) - \hat{S}_{2,x}(t)dt$$

où $\hat{S}_{i,x}(t) = P(T_i > t|x)$ est l'estimateur de Kaplan-Meier pour l'état i et le traiment x et $u_{TOX,x}$ et $u_{REL,x}$ les utilités pour les états de toxicité et de récidive comparativement à la période sans symptôme. La durée passée dans chaque état peut être obtenue par l'espérance de vie restreinte, i.e. l'aire sous la courbe de survie $\hat{S}_{i,x}(t)$ jusqu'au temps d'horizon τ. La décomposition sur les états de santé peut donc être généralisé en supposant l'utilité constante sur la durée d'un état de santé :

$$QALY(\tau) = \sum_s u_s \int_0^\tau (S_s(t) - S_{s-1}(t))dt$$

avec $S_0(t) = 0$.

Les méthodes actuelles utilisées pour estimer le nombre espéré de QALYs souffrent de plusieurs limites. L'approche par Q-TWiST repose sur la définition d'une séquence d'états de santé. Ces états doivent être clairement identifiables, ce qui peut être difficile en pratique. La variabilité individuelle de l'évolution de l'utilité n'est pas prise en considération alors qu'il s'agit d'une variable endogène ayant une évolution continue. Or, considérer une séquence discrète d'états de santé successifs pourrait conduire à des biais si l'écart est important avec le processus continu réel latent. Les utilités sont supposées constantes pour chaque état de santé. Il s'agit d'une hypothèse forte qui ne reflète pas la continuité du processus. Dans l'approche par Q-TWiST, il n'y a pas de mention d'une possible dépendance entre l'évolution de la qualité de vie et le risque de survenue de l'événement. Enfin, l'approche par Q-TWiST est une approche non-paramétrique pour laquelle la prise en considération de variables explicatives est une difficulté importante. Or, pour apprécier le nombre de QALYs entre deux bras de traitement, il peut être nécessaire de considérer l'existence de possible facteurs de confusion.

Alors que de nombreuses études d'évaluation économique sont basées sur l'estimation de QALYs et que de plus en plus d'études de recherche clinique pourraient être basées sur les QALYs, leur modélisation apparaît un sujet important. Les modèles conjoints pour l'analyse de données longitudinales et données de survie fournissent un cadre méthodologique intéressant pour répondre à plusieurs de ces difficultés [Wulfsohn and Tsiatis, 1997; Rizopoulos, 2012]. En particulier, cela pourrait être une approche pertinente pour considérer la dépendance mutuelle entre l'évolution de l'utilité des états de santé et le risque de survenue d'un événement clinique. L'objectif de ce projet est donc de proposer un cadre méthodologique pour estimer le nombre espéré d'années de vie en parfaite santé (QALYs) à partir d'un modèle conjoint original pour l'évolution de l'utilité et le processus de survie.

11.2 Méthodes

En analyse de survie, les données recueillies sont très souvent incomplètes, la censure à droite étant l'exemple le plus fréquemment rencontré [Andersen, 1993; Kalbfleisch and Prentice, 2011]. Le délai de survenue de l'événement T n'est alors pas observé pour tous les sujets. Notons $Y = \min(T, C)$ où C est un délai de censure supposé indépendant. Nous pouvons définir le couple (Y, δ) où $\delta = \mathbb{1}(T \leq C)$ est la fonction indicatrice. Soit la variable aléatoire U qualifiant le processus d'évolution de la qualité de vie mesurée au travers de scores d'utilité répétés au cours du temps. Considérons la situation de décision médicale où deux options thérapeutiques se présentent aux patients. Soit X une variable aléatoire binaire caractérisant les deux modalités de traitement. Notons que la question de l'arbitrage entre quantité et qualité de vie ne se pose évidemment que lorsque l'efficacité du traitement en terme de survie est accompagnée d'une détérioration de la qualité de vie. En effet, dans le cas concordant où le traitement entraine un gain de survie et un gain de qualité de vie, le débat sur l'utilité clinique du traitement n'a pas

réellement lieu.

Approche par modélisation conjointe

Dans une première approche, nous envisageons de modéliser conjointement l'évolution de l'utilité et le risque de survenue d'un événement clinique. Soit un échantillon de n sujets identiquement et indépendamment distribués pour lesquels on observe $\{y_i, \delta_i, u_i, t_i, x_i; i = 1, ..., n\}$ où (y_i, δ_i) la réalisation de (Y, δ) qualifiant la survenue de l'événement pour le sujet i, $u_i = \{u_{ij}; j, ..., n_i\}$ est un score d'utilité mesuré aux temps $t_i = \{t_{ij}; j = 1, ..., n_i\}$. Notons que dans un contexte multivarié, x_i peut également être défini comme une matrice de variables explicatives disponibles à l'inclusion dans l'étude et incluant un ensemble de covariables x_{Li} dont le traitement associées au processus longitudinal et x_{Si} celles associées au processus de survie.

Traditionnellement, un modèle conjoint à effets aléatoires partagés pour données longitudinales et données de survie est formulé sous l'hypothèse d'un modèle linéaire mixte pour l'évolution d'un marqueur quantitatif continu gaussien et sous l'hypothèse d'un modèle à risques proportionnels pour le risque de survenue de l'événement censuré à droite [Rizopoulos, 2012; Commenges and Jacqmin-Gadda, 2015]. La formulation classique d'un modèle conjoint à effets aléatoires partagés est détaillée en section 9.3.1. Dans notre contexte, plusieurs contraintes méthodologiques devront être précisées, notamment concernant la nature du marqueur longitudinal. En effet, le score d'utilité est un marqueur quantitatif prenant des valeurs entre 0 et 1. La distribution retenue devra tenir compte de cet aspect et la distribution Beta est une solution envisageable [Brazier et al., 2017]. [Ferrari and Cribari-Neto, 2004] définit la fonction de densité Beta comme suit :

$$f(u; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} u^{\mu\phi-1}(1-u)^{(1-\mu)\phi-1}$$

où $0 < \mu < 1$ et $\phi > 0$, $\Gamma(\cdot)$ la fonction Gamma, $E(u) = \mu$ et $V(u) = \mu(1-\mu)/(1+\phi)$. Le score d'utilité serait ainsi supposé suivre une densité Beta, $u_i \sim \mathcal{B}(\mu_i, \phi)$.

Le modèle que nous envisageons repose sur l'estimation conjointe d'un modèle de regression Beta défini à l'aide d'un modèle linéaire généralisé à effets aléatoires pour l'évolution de l'utilité et d'un modèle à risque proportionnel pour la survenue d'un événement clinique :

$$\begin{cases} \omega(\mu(t_{ij})) = Q_{ij}^\top \beta + Z_{ij}^\top b_i \\ \lambda(t|R_i, b_i) = \lambda_0(t) \exp(R_i^\top \alpha + g\left(b_i, t, \eta\right)) \end{cases}$$

avec $R_i = \{R_{i1}...R_{ir}\}^\top$ constitué d'élements de X_i et α le vecteur de paramètres associés, $Q_{ij} = \{Q_{ij1}...Q_{ijp}\}^\top$ une matrice composé de X_i, du temps t_i et d'interactions entre les deux, Z_{ij} une sous matrice de Q_{ij}, β et b_i respectivement les vecteurs de coefficients fixes et d'effets

aléatoires associés et $\omega(\cdot)$ une fonction de lien strictement croissante et dérivable deux fois associant l'espérance d'utilité spécifique au sujet $\mu(t_{ij})$ au temps t_{ij} et la matrice $Q_{ij}s$. Différentes fonctions de lien pourraient convenir : logit $\omega(\mu) = \log(\mu/(1-\mu))$; probit $\omega(\mu) = \Phi^{-1}(\mu)$ où $\Phi(\cdot)$ est une densité normale ; complémentaire log-log $\omega(\mu) = \log\{-\log(1-\mu)\}$; log-log $\omega(\mu) = -\log\{-\log(\mu)\}$; Cauchy $\omega(\mu) = \tan\{\pi(\mu - 0.5)\}$, comme le suggère Cribari-Neto and Zeileis [2010].

La log-vraisemblance d'un tel modèle conjoint repose sur une hypothèse d'indépendance conditionnelle du marqueur et de l'événement étudié sachant les effets aléatoires b_i et s'écrira :

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \log \left(\int f_U(u_i|Q_{ij}, b_i; \theta) f_T(t|R_i, b_i; \theta) f_b(b_i; \theta) db_i \right)$$

avec θ le vecteur de l'ensemble des paramètres du modèle, $f_U(\cdot)$ la densité pour l'utilité, $f_T(\cdot)$ la densité pour la survenue de l'événement et $f_b(\cdot)$ la densité pour les effets aléatoires. Soulignons également que l'événement étudié ici est le décès. Dans une situation de risques compétitifs, la vraisemblance du modèle conjoint serait alors plus complexe.

Le modèle conjoint proposé ici devrait permettre de relâcher certaines des limites évoquées préalablement. Les mesures individuelles répétées d'utilité sont modélisées à l'aide d'un modèle linéaire généralisé à effets aléatoires afin de considérer la variabilité individuelle, plutôt qu'une utilité constante. Au lieu d'une séquence discrète d'états de santé, l'évolution du processus latent est modélisée. L'utilisation d'un modèle conjoint permet de considérer la dépendance entre l'évolution de l'utilité et le processus de survie au travers d'effets aléatoires partagés. Cela devrait permettre une meilleure estimation du nombre espéré de QALYs et de la variabilité qui entoure son estimation. Enfin, notre approche paramétrique permet de considérer des covariables dans chacun des sous-modèles d'évolution et de survie.

Dans le cadre d'essai clinique de taille suffisante, les patients traités dans chacun des groupes sont comparables à la mise en place du traitement. L'effet univarié du traitement estimé à partir du modèle conjoint serait alors un effet marginal. En conséquence, le nombre espéré de QALYs jusqu'au temps τ pourrait être estimé de la manière suivante :

$$\begin{aligned} QALY(\tau|X = x; \theta) &= E_{b_i}\left[\int_0^{\tau} S(t|X = x, b_i; \theta)\mu(t|X = x, b_i; \theta)dt\right] \\ &= E_{b_i}\left[\int_0^{\tau} \exp\left(-\int_0^{t} \lambda(\nu|X = x, b_i; \theta)d\nu\right)\mu(t|X = x, b_i; \theta)dt\right] \\ &= \int_{-\infty}^{+\infty}\int_0^{\tau} \exp\left(-\int_0^{t} \lambda(\nu|X = x, b_i; \theta)d\nu\right)\mu(t|X = x, b_i; \theta) f_b(b_i; \theta)dt db_i \end{aligned}$$

Nous proposons donc d'évaluer la balance bénéfice/risque d'une décision médicale en comparant le nombre esperé de QALY dans chacun des groupes de traitement.

Modélisation conjointe après changement d'échelle de temps

Dans une seconde approche, nous pensons repartir de la définition même du nombre de QALYs pour proposer un autre modèle conjoint original. En effet, le temps de survie en parfaite santé noté Q correspond au nombre de QALYs, i.e. au nombre d'années de vie pondérées par la qualité de vie. Pour un sujet i, il peut être défini par $Q_i = \int_0^Y u_i(v)dv$, où $u_i(v)$ est l'utilité au temps v du sujet i. Nous pensons qu'il serait intéressant d'approfondir les réflexions suivantes et de modéliser conjointement l'évolution au cours du temps des scores d'utilité et le risque de survenue d'un événement au temps t dans un état de parfaite santé. Ce modèle pourrait être défini ainsi :

$$\begin{cases} \omega(\mu(t_{ij})) = Q_{ij}^\top \beta + Z_{ij}^\top b_i \\ \lambda_Q(t|R_i, b_i) = \lambda_{0Q}\left(\int_0^t u_i(v)dv\right)\exp(R_i^\top \alpha) \end{cases}$$

L'originalité du modèle conjoint envisagé repose sur l'utilisation de la définition d'un QALY pour un individu. La dépendance entre les deux processus ne passe plus au travers de la fonction $g(\cdot)$. Nous préfererons étudier le délai en parfaite santé de survenue de l'événement à la place du temps observé, la dépendance entre les deux processus passe par cette définition $Q = \int_0^Y u_i(v)dv$. Les interprétations issues d'un tel modèle conjoint seraient alors différentes de celles obtenues à l'aide du premier modèle conjoint envisagé. Le nombre espéré de QALYs pourrait alors être estimé pour un temps de suivi τ de la manière suivante :

$$QALY(t) = E(\min(Q, \tau)) = \int_0^\tau S_Q(t|X)dt$$

où $S_Q(t|X = x) = \mathbb{P}(Q > t|X = x)$ la probabilité de survivre jusqu'en t dans un état de parfaite santé conditionnellement au traitement x. Il reste que ce modèle soulève des interrogations en termes d'identifiabilité et son estimation pourrait poser des problèmes d'ordre computationnel.

Dans le cadre de l'analyse de données d'essais cliniques, les sujets randomisés dans chacun des bras de traitement sont comparables à l'inclusion dans l'étude à l'exception du traitement reçu. L'effet univarié du traitement estimé à partir des modèles conjoints envisagés est donc un effet marginal. L'élaboration du modèle sous-jacent à l'estimation du nombre de QALYs ne requiert pas d'ajustements spécifiques sur des covariables pouvant confondre l'effet du traitement. En revanche à partir de données observationnelles, estimer un effet marginal peut être plus compliqué en présence de possibles facteurs de confusion qu'il sera nécessaire à prendre en considération. L'estimation d'un effet marginal pourra se faire à partir d'un échantillon contrefactuel construit à partir du redressement des différences initialement observées [Cole and Hernán, 2004; Binder, 1992], par exemple par pondération inverse sur un score de propension, ou encore par une méthode d'estimation d'un effet marginal à partir d'un modèle multivarié appelé *G-computation* [Robins, 1986; Snowden et al., 2011; Wang et al., 2017].

La réalisation d'un tel projet comporte plusieurs étapes. Il nous faudra développer de manière plus précise l'écriture analytique du modèle conjoint envisagé. Nous réaliserons une étude de simulation dont l'objectif sera de valider l'estimation du modèle conjoint proposé. Afin de démontrer l'intérêt de l'approche proposée, nous étudierons les propriétés en termes de calculs de biais, de variance dans l'estimation des QALYs. Nous comparerons ces estimations notamment avec l'approche proposée par Glasziou et al. [1998] pour l'estimation des QALYs. Par ailleurs, nous pensons développer un package R afin de rendre plus facile l'usage d'une telle approche. Au-delà du cadre théorique qui mérite d'être complété, des applications sur données réelles seront réalisées pour montrer l'intérêt de l'évaluation de l'effet d'un traitement à partir d'une approche méthodologique intégrant les préférences des patients pour les conséquences des décisions thérapeutiques.

11.3 Applications en cancérologie

Nous illustrerons l'intérêt des approches envisagées dans le contexte de la cancérologie en étudiant l'utilité clinique de traitements à partir de données d'essais cliniques. En collaboration avec l'institut Curie, l'essai clinique STIC CTC phase 3 (NCT01710605) a initialement été mis en place afin de savoir si le taux de cellules tumorales circulantes (CTC - Circulating Tumors Cells) pouvait être utilisé pour orienter le choix entre une hormonothérapie et une chimiothérapie dans le cancer du sein. Les patientes ont été randomisées entre deux bras. Dans le premier, le choix de la thérapie est décidé cliniquement (bras standard). Dans le deuxième, l'hormothérapie est proposée si le taux de CTC est inférieur à 5CTC/7.5ml alors que la chimiothérapie est donnée dans le cas contraire (bras CTC). Les résultats initiaux indiquent une non infériorité du bras CTC en terme de survie comparativement au bras standard. Par ailleurs, l'hormonothérapie est plus fréquemment proposée dans le bras standard (72.6%) comparativement au bras CTC (62.6%). Supposant que la qualité de vie des patients pour ces deux thérapies puissent être différentes, il apparaît nécessaire d'étudier l'intérêt d'une telle stratégie de choix médical en considérant la qualité de vie des patients. Notre approche pourrait s'avérer pertinente dans un tel contexte.

En collaboration avec l'Institut du Cancer de Montpellier, nous pensons réanalyser les données de l'essai clinique randomisé PRODIGE 4/ACCORD 11 (ClinicalTrials.gov, number NTC00112658). L'objectif de cet essai clinique était de comparer un traitement par Folforinox (combinaison de régimes de chimiothérapie) par rapport au traitement de première ligne Gemcitabine chez des patients avec un cancer métastatique pancréatique. Dans un premier article, les auteurs concluent à un gain de survie chez les patients avec Folforinox [Conroy et al., 2011]. En parallèle, une analyse de la qualité de vie a été réalisée [Gourgou-Bourgade et al., 2013]. L'estimation du nombre espéré de QALYs pour chacun de bras de traitement permettra de mieux évaluer l'utilité clinique du traitement par Folforinox.

Toujours en collaboration avec l'Institut du Cancer de Montpellier, nous nous intéresserons à l'essai PRODIGE 5/ACCORD 17 (ClinicalTrials.gov, number NTC00861094) dans le cancer de l'oesophage. Dans cet essai, la survie pour la progression de la maladie n'était pas significativement plus élevée dans le bras de traitement par FOLFOX comparativement au bras de traitement pas fluorouracil et cisplatin. Mais les analyses longitudinales de la qualité de vie semblent indiquer une détérioration plus importante de la qualité de vie pour les patients du groupe FOLFOX pour certaines dimensions du QLQ-C30. Les analyses séparées de la survie et de la qualité de vie rendent difficiles les conclusions puisque le traitement par FOLFOX semble indiquer une meilleur espérance de vie accompagnée d'une déterioration de la qualité de vie. Après conversion des mesures répétées de QLQ-C30 en mesures d'utilité, le modèle conjoint que nous proposons permettra une nouvelle interprétation en termes de QALYs espérés.

11.4 Applications en transplantation rénale

Depuis 2016, des questionnaires de qualité de vie EQ-5D sont annuellement collectés chez les patients transplantés de la cohorte DIVAT. Actuellement, quatre centres de transplantation recueillent ces informations (Nantes, Lyon, Paris-Necker et Nice). Plus de 1500 patients transplantés d'un rein ont au moins eu une mesure d'EQ-5D et le nombre maximal de mesures est aujourd'hui de 4. Ce recueil de données se poursuit et devrait permettre d'envisager une modélisation longitudinale des scores d'utilité.

Une étude en cours souligne l'importante hétérogénéité dans le choix de la thérapie d'induction entre les centres de transplantation du réseau DIVAT. Elle illustre la nécessité de nouveaux outils ou nouvelles analyses permettant d'homogénéiser les pratiques cliniques sur le choix du traitement d'induction. Le traitement par Anti-Thymocyte Globuline (ATG) est un traitement d'induction déplétant notamment utilisé pour ses propriétés pharmacologiques dans la prévention du rejet aigu et du délai de retard au démarrage, facteurs de risque d'échec de greffe. Son utilité clinique peut être discutée au vu des effets secondaires qui pourraient détériorer la qualité de vie des patients. A l'aide de l'approche proposée, nous serons en mesure d'estimer le nombre de QALYs pour les patients recevant de l'ATG et celui pour ceux recevant un traitement d'induction non déplétant par Basiliximab afin de déterminer si l'arbitrage entre quantité et qualité de vie est favorable à l'un ou l'autre des deux traitements.

Quatrième partie

Conclusions et Perspectives

Conclusion et Perspectives

La rédaction de ce manuscrit d'Habilitation à Diriger des Recherches aura été une formidable occasion de regarder en arrière et de m'interroger sur l'évolution de mes travaux de recherche. Depuis près de dix ans, mes recherches concernent le développement et l'utilisation de méthodes permettant d'étudier l'évolution de l'état de santé de patients atteints de maladies chroniques. Trois thématiques ont progressivement émergées : la modélisation de l'histoire naturelle de maladies chroniques, la recherche autour d'outils prédictifs, et le développement d'approches pour une aide à la décision médicale centrée sur le patient.

Dans une démarche étiologique, je me suis intéressé aux modèles conjoints pour données longitudinales et données de survie ainsi qu'aux modèles multi-états. Ces modèles permettent de mieux appréhender le rôle de facteurs de risque en présence de données longitudinales incomplètes. Ils m'apparaissent extrêmement importants pour la compréhension des mécanismes d'évolution d'une maladie chronique. Pour autant, la place concrète que pourrait trouver ces travaux auprès des patients ne me semble pas immédiate. Les recherches que nous menons dans un cadre pronostique pourrait être plus à même de trouver un écho auprès des cliniciens pour une utilisation concrète et avoir de plus grandes retombées cliniques. A travers différents projets de recherche, j'ai participé à l'élaboration de scores prédictifs ainsi qu'à l'évaluation de leurs performances prédictives. Le projet de recherche de l'équipe SPHERE porte sur le développement de méthodes dans le domaine de la recherche en santé centrée sur le patient. Dans ce contexte, l'un des objectifs majeurs des recherches que j'envisage concerne la place du patient dans ces recherches en pronostic qui, il me semble, est encore trop souvent négligée. L'identification d'un seuil de discrimination pour un score pronostique à partir de la maximisation d'une fonction d'utilité intégrant les préférences des patients pour les différentes strates de prise en charge en est un exemple.

Motivés par les applications dans le domaine de la santé, mes travaux s'intéressent aux problématiques méthodologiques rencontrées à partir de données réelles. En transplantation rénale, les questions de recherche clinique sont nombreuses et la cohorte DIVAT est un excellent outil pour tenter d'y apporter des réponses. L'épidémiologie de la transplantation rénale est ainsi un fil conducteur depuis plusieurs années, mais j'élargis de plus en plus mon expertise à

d'autres pathologies comme la cancérologie ou la réanimation, qui posent de nouvelles questions sur le plan méthodologique.

Mon programme de recherche pour les années à venir s'articulera autour de plusieurs axes. A court terme, il s'agira de poursuivre et finaliser les projets déjà débutés. Je pense en particulier aux travaux sur la modélisation en réanimation et ceux sur le pronostic dynamique en transplantation rénale (DynPG). Dans le cadre de travaux collaboratifs avec le Dr. Özgür Asar, nous espérons pouvoir améliorer l'outil pronostique de l'échec de greffe en considérant par exemple la présence de mesures longitudinales extrêmes. Un outil qui intégrerait plusieurs marqueurs longitudinaux (sérum de créatinine, protéinurie, etc.) serait également une avancée notable. Par ailleurs, le développement de ce score est une première étape qui ne peut suffire. Il est à présent temps d'envisager son utilisation et sa validation en pratique clinique. Cela pose de nombreuses questions autour de l'utilisation d'un tel score comme outil d'information au patient, de la mise en place de programme d'éducation thérapeutique pour aider les patients à comprendre le risque d'échec de greffe, de l'évaluation d'une telle information sur la qualité de vie du patient.

J'aimerais également renforcer les collaborations existantes (Yohann Foucher, Magali Giral, Philippe Tessier, etc.) et consolider ainsi les recherches mises en oeuvre au sein de notre groupe de travail, et par ailleurs développer de nouvelles collaborations. Nous avons, par exemple, récemment été sollicités par le Dr. Samy Hadjaj pour travailler au développement d'un outil pronostique dynamique à partir de patients atteints d'un diabète de type II de la cohorte SURDIAGENE. En collaboration avec le Dr. Célia Touraine, nous travaillerons à la modélisation des QALYs à partir de modèles conjoints pour l'évolution longitudinale d'un score d'utilité et du risque de survenue d'un événement clinique. Il est envisagé d'appliquer ces développements dans le domaine de la cancérologie. En cohérence avec le projet d'équipe SPHERE, ce travail sera essentiel dans le futur proche. J'espère que ce projet de modélisation ambitieux pourra être supporté par l'obtention de financements. Deux lettres d'intention ont été acceptées et les projets complets sont à l'étude auprès de l'Agence Nationale de la Recherche pour l'appel d'offre *Jeunes Chercheuses et Jeunes Chercheurs* ainsi qu'à l'Institut National du Cancer pour l'appel d'offre *Sciences Humaines et Sociales, Epidémiologie et Santé Publique.*

La pérénisation de notre groupe de travail autour de la médecine de précision va bien au-délà de cette Habilitation à Diriger des Recherches. Etroitement liée à l'équipe SPHERE, l'évolution de notre groupe de recherche dépendra du futur de nos chercheurs juniors (doctorants et post-doctorants). La pérénnisation de postes de certains collaborateurs contractuels ou le recrutement de chercheurs extérieurs me semblent être essentiels. Dans ce contexte, l'accompagnement de nos tutelles (Université, INSERM) sera primordiale. En tant qu'enseignant-chercheur, une dimension importante de mon activité professionnelle concerne la transmission de connaissances. La direction de recherche et l'accompagnement d'étudiants en thèse est en lien très étroit avec ces activités. Notre dynamisme actuel avec une multiplication des projets de recherche repose pour beaucoup sur les chercheurs juniors et doit s'accompagner d'une structuration scientifique de

notre groupe de travail. A titre personnel, l'obtention d'une Habilitation à Diriger des Recherches me permettra d'acquérir une plus grande autonomie dans la gestion de nos travaux en cours et à venir et ainsi que de contribuer à l'évolution future de notre équipe de recherche.

Table des figures

3.1 Modèle conjoint à état latent pour l'accélération du déclin cognitif, la survenue d'une démence et la survenue d'un décès, avec α_{01}, α_{12} et α_3 les intensités de transitions entre les differents états (voir Dantan et al. [2011] pour plus de détails). 28

6.1 Illustration d'une prédiction de l'échec de greffe estimée au temps de landmark $s = 4$ pour une fenêtre d'horizon t de 5 ans, et calculée en fonction de l'évolution du sérum de créatinine jusqu'en s et de covariables X_i connues à l'inclusion. . . . 62

9.1 Calibration des prédictions dynamiques sur l'échantillon de validation (n=2523, 66 sujets supprimées pour cause de données manquantes) pour des temps de landmark allant de 1 à 6 ans post-transplantation. Les risques moyens prédits et observés (Kaplan-Meier) sont présentés pour des sous-groupes définis par les quantiles de prédictions. 107

9.2 Capacités pronostiques des prédictions dynamiques sur un horizon de 5 ans (n=2523, 66 sujets supprimées pour cause de données manquantes) estimées pour des temps de landmark allant de 1 à 6 ans post-transplantation. La courbe de R^2 est un indicateur de performance globale (Partie A), tandis que l'aire sous la courbe ROC (AUC) évalue les performances de discrimination (Partie B). Les estimations sont représentées par les lignes en traits plein, les bandes de confiance à 95% sont dessinées en trait pointillé. 108

9.3 Capacités pronostiques des prédictions dynamiques obtenues à partir du modèle conjoint publié par Fournier et al. [2016] et comparées à celles obtenues à partir du modèle conjoint simplifié. Ces prédictions sont calculées à partir de l'échantillon de validation (n=2454, 135 sujets supprimés pour cause de données manquantes) estimées pour des temps de landmark allant de 1 à 6 ans post-transplantation. La courbe de R^2 est un indicateur de performance globale (Partie A), tandis que l'aire sous la courbe ROC (AUC) évalue les performances de discrimination (Partie B). 109

9.4 Prédictions dynamiques individuelles obtenues à partir du modèle conjoint simplifié pour des temps de landmark allant de 1 à 6 ans post-transplantation pour une femme de 51 ans transplantée pour la première fois en 2005, immunisée en classe I en pré-greffe, sans antécédent de maladies cardiovasculaires, ni rejet aigu au cours de la première année de greffe et avec un niveau de sérum de créatinine à 88 μmol/L à 3 mois post-transplantation (Cas A). Cette patiente subit finalement un retour en dialyse à 9.3 ans post-transplantation. 111

9.5 Prédictions dynamiques individuelles obtenues à partir du modèle conjoint simplifié pour des temps de landmark allant de 1 à 6 ans post-transplantation pour une femme de 60 ans transplantée pour la seconde fois en 2007, immunisée en classe I en pré-greffe, sans antécédent de maladies cardiovasculaires, ayant connu au moins un épisode de rejet aigu au cours de la première année de greffe et avec un niveau de sérum de créatinine à 100μmol/L à 3 mois post-transplantation (Cas B). Cette patiente est toujours en vie avec un greffon fonctionnel à 10 post-transplantation. 113

9.6 Simulation de 30 sujets à partir d'un modèle conjoint robuste (Intercept et pente d'évolution du marqueurs supposés égaux à 0). 118

9.7 Mesures observées du log de sérum de créatinine en fonction du délai écoulé (en années) avant la date de dernier suivi (Censored subjects) ou avant la date d'échec de greffe rénale (Graft failure) sur l'échantillon d'apprentissage (n=2749). Le délai est exprimé avec la date de dernier suivi comme origine de temps. Le trait plein représente l'évolution moyenne estimée à partir d'un modèle linéaire généralisé pénalisé intégrant une fonction de lissage [Wood, 2004]. 120

11.1 Illustration du calcul du nombre espéré de QALYs. 130

Liste des tableaux

9.1 Descriptif des caractéristiques des receveurs et donneurs et des caractéristiques de la transplantation selon l'échantillon d'apprentissage (n=2749) et l'échantillon de validation (n=2589). 105

9.2 Modèle conjoint simplifié pour l'évolution du logarithme de sérum de créatinine (SCr) et le risque d'échec de greffe chez des patients transplantés rénaux (n=2584, 165 patients exclus pour cause de données manquantes). 106

10.1 Description des catégories du Glasgow Outcome Scale (GOS). 125

Bibliographie

[1] M. Ali, R. Fulton, T. Quinn, M. Brady, and VISTA Collaboration. How well do standard stroke outcome measures reflect quality of life ? A retrospective analysis of clinical trial data. Stroke, 44(11) :3161–3165, Nov. 2013. 127

[2] D. G. Altman and P. Royston. The cost of dichotomising continuous variables. BMJ (Clinical research ed.), 332(7549) :1080, May 2006. 127

[3] P. K. Andersen. Statistical Models Based on Counting Processes. Springer-Verlag, Jan. 1993. 132

[4] P. K. Andersen, K. Borch-Johnsen, T. Deckert, A. Green, P. Hougaard, N. Keiding, and S. Kreiner. A Cox regression model for the relative mortality and its application to diabetes mellitus survival data. Biometrics, 41(4) :921–932, Dec. 1985. 126

[5] T. M. L. Andersson, P. W. Dickman, S. Eloranta, and P. C. Lambert. Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. BMC medical research methodology, 11 :96, June 2011. 126

[6] Ö. Asar, J. Ritchie, P. A. Kalra, and P. J. Diggle. Joint modelling of repeated measurement and time-to-event data : An introductory tutorial. International Journal of Epidemiology, 44(1) :334–344, Feb. 2015. 47

[7] P. C. Austin. 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. 90

[8] T. Baghfalaki, M. Ganjali, and D. Berridge. Robust joint modeling of longitudinal measurements and time to event data using normal/independent distributions : A Bayesian approach. Biometrical Journal, 55(6) :844–865, Nov. 2013. 115, 117

[9] T. Baghfalaki, M. Ganjali, and R. Hashemi. Bayesian Joint Modeling of Longitudinal Measurements and Time-to-Event Data Using Robust Distributions. Journal of Biopharmaceutical Statistics, 24(4) :834–855, July 2014. 115, 117

[10] E. Basch. Toward patient-centered drug development in oncology. The New England Journal of Medicine, 369(5) :397–400, Aug. 2013. 101, 129

[11] D. A. Binder. Fitting Cox's Proportional Hazards Models from Survey Data. Biometrika, 79(1) :139–147, 1992. 135

[12] P. Blanche, C. Proust-Lima, L. Loubère, C. Berr, J.-F. Dartigues, and H. Jacqmin-Gadda. Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. Biometrics, 71 (1) :102–113, Mar. 2015. 30, 104

[13] O. Boussari, G. Romain, L. Remontet, N. Bossard, M. Mounier, A.-M. Bouvier, C. Binquet, M. Colonna, and V. Jooste. A new approach to estimate time-to-cure from cancer registries data. Cancer Epidemiology, 53 :72–80, Apr. 2018. 126

[14] M. Brankovic, I. Kardys, E. J. Hoorn, S. Baart, E. Boersma, and D. Rizopoulos. Personalized dynamic risk assessment in nephrology is a next step in prognostic research. Kidney International, 94(1) :214–217, July 2018. 102

[15] J. Brazier, J. Ratcliffe, J. Saloman, and A. Tsuchiya. Measuring and Valuing Health Benefits for Economic Evaluation. Oxford University Press, 2017. 133

[16] A. Brazinova, V. Rehorcikova, M. S. Taylor, V. Buckova, M. Majdan, M. Psota, W. Peeters, V. Feigin, A. Theadom, L. Holkovic, and A. Synnot. Epidemiology of Traumatic Brain Injury in Europe : A Living Systematic Review. Journal of Neurotrauma, Aug. 2016. 123

[17] N. E. Breslow and N. E. Day. Indirect standardization and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data. Journal of Chronic Diseases, 28(5-6) :289–303, June 1975. 126

[18] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. Handbook of Markov Chain Monte Carlo. CRC Press, May 2011. 119

[19] D. Cantarovich, B. Hodemon-Corne, K. Trébern-Launay, M. Giral, Y. Foucher, and E. Dantan. Early steroid withdrawal compared with steroid avoidance correlates with graft failure among kidney transplant recipients with an history of diabetes. Transplantation proceedings, 45(4) :1497–1502, May 2013. 33

[20] F. J. Carod-Artal and J. A. Egido. Quality of life after stroke : The importance of a good recovery. Cerebrovascular Diseases (Basel, Switzerland), 27 Suppl 1 :204–214, 2009. 127

[21] E. Caruana, Y. Foucher, P. Tessier, J.-S. Frenel, J.-M. Classe, and E. Dantan. Patient-centered simulations to assess the usefulness of the 70-gene signature for adjuvant chemotherapy administration in early-stage breast cancer. Breast Cancer Research and Treatment, Jan. 2019. 33

[22] N. Chaisinanunkul, O. Adeoye, R. J. Lewis, J. C. Grotta, J. Broderick, T. G. Jovin, R. G. Nogueira, J. J. Elm, T. Graves, S. Berry, K. R. Lees, A. D. Barreto, J. L. Saver, DAWN Trial and MOST Trial Steering Committees, and Additional contributors from DAWN Trial Steering Committee. Adopting a Patient-Centered Approach to Primary Outcome Analysis of Acute Stroke Trials Using a Utility-Weighted Modified Rankin Scale. Stroke, 46(8) :2238–2243, Aug. 2015. 127

[23] I. Chalmers and P. Glasziou. Avoidable waste in the production and reporting of research evidence. Lancet (London, England), 374(9683) :86–89, July 2009. 61

[24] I. Chalmers, M. B. Bracken, B. Djulbegovic, S. Garattini, J. Grant, A. M. Gülmezoglu, D. W. Howells, J. P. A. Ioannidis, and S. Oliver. How to increase value and reduce waste when research priorities are set. Lancet (London, England), 383(9912) :156–165, Jan. 2014. 61

[25] S. R. Cole and M. A. Hernán. Adjusted survival curves with inverse probability weights. Computer Methods and Programs in Biomedicine, 75(1) :45–49, July 2004. 135

[26] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) : The TRIPOD statement. Annals of Internal Medicine, 162(1) :55–63, Jan. 2015. 61, 104, 124

[27] C. Combescure, T. V. Perneger, D. C. Weber, J.-P. Daurès, and Y. Foucher. 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, Jan. 2014. 30

[28] C. Combescure, J. P. Daures, and Y. Foucher. A literature-based approach to evaluate the predictive capacity of a marker using time-dependent summary receiver operating characteristics. Statistical Methods in Medical Research, 25(2) :674–685, Apr. 2016. 30

[29] D. Commenges and H. Jacqmin-Gadda. Dynamical Biostatistical Models. CRC Press, Oct. 2015. 133

[30] T. Conroy, F. Desseigne, M. Ychou, O. Bouché, R. Guimbaud, Y. Bécouarn, A. Adenis, J.-L. Raoul, S. Gourgou-Bourgade, C. de la Fouchardière, J. Bennouna, J.-B. Bachet, F. Khemissa-Akouz, D. Péré-Vergé, C. Delbaldo, E. Assenat, B. Chauffert, P. Michel, C. Montoto-Grillot, M. Ducreux, Groupe Tumeurs Digestives of Unicancer, and PRODIGE Intergroup. FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. The New England Journal of Medicine, 364(19) :1817–1825, May 2011. 136

[31] G. Couvrat-desvergnes, A. Masseau, O. Benveniste, A. Bruel, B. Hervier, J.-m. Mussini, D. Buob, E. Hachulla, P. Rémy, R. Azar, E. M. Namara, B. Macgregor, L. Daniel, A. Lacraz, T. D. Broucker, P. Rouvier, P. Carli, M. Laville, E. Dantan, M. Hamidou, A. Moreau,

and F. Fakhouri. The Spectrum of Renal Involvement in Patients With Inflammatory Myopathies. Medicine, 93(1) :33–41, Jan. 2014. 26

[32] G. Couvrat-Desvergnes, Y. Foucher, F. Le Borgne, A. Dion, G. Mourad, V. Garrigue, C. Legendre, L. Rostaing, N. Kamar, M. Kessler, M. Ladrière, E. Morelon, F. Buron, M. Giral, and E. Dantan. Comparison of graft and patient survival according to the transplantation centre policy for 1-year screening biopsy among stable kidney recipients : A propensity score-based study. Nephrology, Dialysis, Transplantation : Official Publication of the European Dialysis and Transplant Association - European Renal Association, July 2018. 23, 33, 34

[33] D. Cox. Regression models and life-tables. Journal of the Royal Statistical Society : Series B, 34 :187–229, 1972. 29

[34] F. Cribari-Neto and A. Zeileis. Beta Regression in R. Journal of Statistical Software, 34 (2), 2010. 134

[35] Z. Daher Abdi, M. Essig, D. Rizopoulos, Y. Le Meur, A. Prémaud, J. B. Woillard, J. P. Rérolle, P. Marquet, and A. Rousseau. Impact of longitudinal exposure to mycophenolic acid on acute rejection in renal-transplant recipients using a joint modeling approach. Pharmacological Research, 72 :52–60, June 2013. 47

[36] E. Dantan, C. Proust-Lima, L. Letenneur, and H. Jacqmin-Gadda. Pattern mixture models and latent class models for the analysis of multivariate longitudinal data with informative dropouts. The international journal of biostatistics, 4(1) :Article 14, 2008. 28

[37] E. Dantan, P. Joly, J.-F. Dartigues, and H. Jacqmin-Gadda. Joint model with latent state for longitudinal and multistate data. Biostatistics (Oxford, England), 12(4) :723–736, Oct. 2011. 28, 145

[38] E. Dantan, C. Combescure, M. Lorent, J. Ashton-Chess, P. Daguin, J.-M. Classe, M. Giral, and Y. Foucher. An original approach was used to better evaluate the capacity of a prognostic marker using published survival curves. Journal of Clinical Epidemiology, 67 (4) :441–448, Apr. 2014. 15, 30, 38

[39] E. Dantan, Y. Foucher, M. Lorent, M. Giral, and P. Tessier. Optimal threshold estimator of a prognostic marker by maximizing a time-dependent expected utility function for a patient-centered stratified medicine. Statistical Methods in Medical Research, 27(6) : 1847–1859, June 2018. 16, 32, 73, 130

[40] R. De Angelis, R. Capocaccia, T. Hakulinen, B. Soderman, and A. Verdecchia. Mixture models for cancer survival analysis : Application to population-based data with covariates. Statistics in Medicine, 18(4) :441–454, Feb. 1999. 126

[41] A. Debout, Y. Foucher, K. Trébern-Launay, C. Legendre, H. Kreis, G. Mourad, V. Garrigue, E. Morelon, F. Buron, L. Rostaing, N. Kamar, M. Kessler, M. Ladrière, A. Poignas, A. Blidi, J.-P. Soulillou, M. Giral, and E. Dantan. Each additional hour of cold ischemia time significantly increases the risk of graft failure and mortality following renal transplantation. Kidney International, 87(2) :343–349, Feb. 2015. 24, 33

[42] S. A. Dijkland, D. C. Voormolen, E. Venema, B. Roozenbeek, S. Polinder, J. A. Haagsma, D. Nieboer, V. Chalos, A. J. Yoo, J. Schreuders, A. van der Lugt, C. B. L. M. Majoie, Y. B. W. E. M. Roos, W. H. van Zwam, R. J. van Oostenbrugge, E. W. Steyerberg, D. W. J. Dippel, H. F. Lingsma, and MR CLEAN Investigators. Utility-Weighted Modified Rankin Scale as Primary Outcome in Stroke Trials : A Simulation Study. Stroke, 49(4) :965–971, Apr. 2018. 127

[43] J. Esteve, E. Benhamou, M. Croasdale, and L. Raymond. Relative survival and the estimation of net survival : Elements for further discussion. Stat Med, 9(5) :529–38, 1990. 125

[44] V. T. Farewell. The use of mixture models for the analysis of survival data with long-term survivors. Biometrics, 38(4) :1041–1046, Dec. 1982. 126

[45] D. Feeny. A utility approach to the assessment of health-related quality of life. Medical Care, 38(9 Suppl) :II151–154, Sept. 2000. 127

[46] S. Ferrari and F. Cribari-Neto. Beta Regression for Modelling Rates and Proportions. Journal of Applied Statistics, 31(7) :799–815, Aug. 2004. 133

[47] M. Flores, G. Glusman, K. Brogaard, N. D. Price, and L. Hood. P4 medicine : How systems medicine will transform the healthcare sector and society. Personalized Medicine, 10(6) :565–576, 2013. 30, 112, 123

[48] Y. Foucher and R. Danger. Time dependent ROC curves for the estimation of true prognostic capacity of microarray data. Statistical Applications in Genetics and Molecular Biology, 11(6) :Article 1, Nov. 2012. 30

[49] Y. Foucher, M. Giral, J.-P. Soulillou, and J.-P. Daures. A semi-Markov model for multistate and interval-censored data with multiple terminal events. Application in renal transplantation. Statistics in Medicine, 26(30) :5381–5393, Dec. 2007. 22, 29

[50] Y. Foucher, P. Daguin, A. Akl, M. Kessler, M. Ladrière, C. Legendre, H. Kreis, L. Rostaing, N. Kamar, G. Mourad, V. Garrigue, F. Bayle, B. H de Ligny, M. Büchler, C. Meier, J. P. Daurès, J.-P. Soulillou, and M. Giral. A clinical scoring system highly predictive of long-term kidney graft survival. Kidney International, 78(12) :1288–1294, Dec. 2010. 62

[51] Y. Foucher, M. Giral, J. P. Soulillou, and J. P. Daures. A flexible semi-Markov model for interval-censored data and goodness-of-fit testing. Statistical Methods in Medical Research, 19(2) :127–145, Apr. 2010. 22, 29

[52] Y. Foucher, M. Giral, J. P. Soulillou, and J. P. Daures. Time-dependent ROC analysis for a three-class prognostic with application to kidney transplantation. Statistics in Medicine, 29(30) :3079–3087, Dec. 2010. 30

[53] Y. Foucher, M. Lorent, P. Tessier, S. Supiot, V. Sébille, and E. Dantan. A mini-review of quality of life as an outcome in prostate cancer trials : Patient-centered approaches are needed to propose appropriate treatments on behalf of patients. Health and Quality of Life Outcomes, 16(1) :40, Mar. 2018. 32, 129

[54] M.-C. Fournier, Y. Foucher, P. Blanche, F. Buron, M. Giral, and E. Dantan. A joint model for longitudinal and time-to-event data to better assess the specific role of donor and recipient factors on long-term kidney transplantation outcomes. European Journal of Epidemiology, 31(5) :469–479, May 2016. 20, 29, 47, 102, 103, 105, 109, 116, 145

[55] M.-C. Fournier, E. Dantan, and P. Blanche. An R2 -curve for evaluating the accuracy of dynamic predictions. Statistics in Medicine, 37(7) :1125–1133, Mar. 2018. 20, 31, 63, 104

[56] M.-C. Fournier, Y. Foucher, P. Blanche, C. Legendre, S. Girerd, M. Ladrière, E. Morelon, F. Buron, L. Rostaing, N. Kamar, G. Mourad, V. Garrigue, G. Couvrat-Desvergnes, M. Giral, E. Dantan, and DIVAT Consortium. Dynamic predictions of long-term kidney graft failure : An information tool promoting patient-centred care. Nephrology, Dialysis, Transplantation : Official Publication of the European Dialysis and Transplant Association - European Renal Association, Mar. 2019. 20, 31, 63, 102

[57] L. Frank, E. Basch, J. V. Selby, and Patient-Centered Outcomes Research Institute. The PCORI perspective on patient-centered outcomes research. JAMA, 312(15) :1513–1514, Oct. 2014. 31

[58] M. Gail. A review and critique of some models used in competing risk analysis. Biometrics, 31(1) :209–222, Mar. 1975. 29

[59] F. G. Garre, A. H. Zwinderman, R. B. Geskus, Y. W. J. Sijpkens, and TNO Kwaliteit van Leven. A joint latent class changepoint model to improve the prediction of time to graft failure, Jan. 2008. 47, 102

[60] T. A. Gerds and M. Schumacher. Consistent estimation of the expected Brier score in general survival models with right-censored event times. Biometrical Journal. Biometrische Zeitschrift, 48(6) :1029–1040, Dec. 2006. 30, 63

[61] F. Gillaizeau, E. Dantan, M. Giral, and Y. Foucher. A multistate additive relative survival semi-Markov model. Statistical Methods in Medical Research, 26(4) :1700–1711, Aug. 2017. 22, 29

[62] F. Gillaizeau, T. Sénage, F. Le Borgne, T. Le Tourneau, J.-C. Roussel, K. Leffondrè, R. Porcher, B. Giraudeau, E. Dantan, and Y. Foucher. Inverse probability weighting to control confounding in an illness-death model for interval-censored data. Statistics in Medicine, 37(8) :1245–1258, Apr. 2018. 22, 29

[63] G. R. Gillett, S. Honeybul, K. M. Ho, and C. R. P. Lind. Neurotrauma and the RUB : Where tragedy meets ethics and science. Journal of Medical Ethics, 36(12) :727–730, Dec. 2010. 123

[64] M. Giral, K. Renaudin, M. Naesens, R. Luning, D. Anglicheau, E. Morelon, A. Huneau, C. Paul, S. Brouard, G. Couvrat-Desvergnes, Y. Foucher, and E. Dantan. The 1-year Renal Biopsy Index (1-RBI) : A scoring system to drive biopsy indication at 1-year post-kidney transplantation. Transplant International : Official Journal of the European Society for Organ Transplantation, June 2018. 15, 25, 31, 33, 34

[65] P. P. Glasziou, R. J. Simes, and R. D. Gelber. Quality adjusted survival analysis. Statistics in Medicine, 9(11) :1259–1276, Nov. 1990.

[66] P. P. Glasziou, B. F. Cole, R. D. Gelber, J. Hilden, and R. J. Simes. Quality adjusted survival analysis with repeated quality of life measures. Statistics in Medicine, 17(11) : 1215–1229, June 1998. 131, 136

[67] A. Goldhirsch, R. D. Gelber, R. J. Simes, P. Glasziou, and A. S. Coates. Costs and benefits of adjuvant therapy in breast cancer : A quality-adjusted survival analysis. Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology, 7(1) : 36–44, Jan. 1989. 131

[68] E. J. Gordon, Z. Butt, S. E. Jensen, A. Lok-Ming Lehr, J. Franklin, Y. Becker, L. Sherman, W. J. Chon, N. Beauvais, J. Hanneman, D. Penrod, M. G. Ison, and M. M. Abecassis. Opportunities for shared decision making in kidney transplantation. American Journal of Transplantation : Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons, 13(5) :1149–1158, May 2013. 30, 101, 112

[69] S. Gourgou-Bourgade, C. Bascoul-Mollevi, F. Desseigne, M. Ychou, O. Bouché, R. Guimbaud, Y. Bécouarn, A. Adenis, J.-L. Raoul, V. Boige, J. Bérille, and T. Conroy. Impact of FOLFIRINOX compared with gemcitabine on quality of life in patients with metastatic pancreatic cancer : Results from the PRODIGE 4/ACCORD 11 randomized trial. Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology, 31 (1) :23–29, Jan. 2013. 136

[70] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine, 18(17-18) : 2529–2545, Sept. 1999. 30, 62, 63

[71] M. B. Hamel, L. Goldman, J. Teno, J. Lynn, R. B. Davis, F. E. Harrell, A. F. Connors, R. Califf, P. Kussin, and P. Bellamy. Identification of comatose patients at high risk for death or severe disability. SUPPORT Investigators. Understand Prognoses and Preferences for Outcomes and Risks of Treatments. JAMA, 273(23) :1842–1848, June 1995. 123

[72] P. J. Heagerty, T. Lumley, and M. S. Pepe. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics, 56(2) :337–344, June 2000. 16, 38

[73] H. Hemingway, P. Croft, P. Perel, J. A. Hayden, K. Abrams, A. Timmis, A. Briggs, R. Udumyan, K. G. M. Moons, E. W. Steyerberg, I. Roberts, S. Schroter, D. G. Altman, R. D. Riley, and PROGRESS Group. Prognosis research strategy (PROGRESS) 1 : A framework for researching clinical outcomes. BMJ (Clinical research ed.), 346 :e5595, Feb. 2013. 124

[74] R. Henderson, P. Diggle, and A. Dobson. Joint modelling of longitudinal measurements and event time data. Biostatistics (Oxford, England), 1(4) :465–480, Dec. 2000. 47

[75] A. D. Hingorani, D. A. van der Windt, R. D. Riley, K. Abrams, K. G. M. Moons, E. W. Steyerberg, S. Schroter, W. Sauerbrei, D. G. Altman, H. Hemingway, and PROGRESS Group. Prognosis research strategy (PROGRESS) 4 : Stratified medicine research. BMJ (Clinical research ed.), 346 :e5793, Feb. 2013. 124

[76] M. D. Hoffman and A. Gelman. The No-U-Turn Sampler : Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. Nov. 2011. 119

[77] P. Hougaard. Frailty models for survival data. Lifetime Data Analysis, 1(3) :255–273, 1995. 121

[78] H. C. V. Houwelingen. Dynamic Prediction by Landmarking in Event History Analysis. Scandinavian Journal of Statistics, 34(1) :70–85, Mar. 2007. 62

[79] M. Howell, G. Wong, J. Rose, A. Tong, J. C. Craig, and K. Howard. Patient Preferences for Outcomes After Kidney Transplantation : A Best-Worst Scaling Survey. Transplantation, 101(11) :2765–2773, Nov. 2017. 101

[80] X. Huang, G. Li, and R. M. Elashoff. A joint model of longitudinal and competing risks survival data with heterogeneous random effects and outlying longitudinal measurements. Statistics and its interface, 3(2) :185–195, 2010. 115, 117

[81] A. I. Iglesias, R. Mihaescu, J. P. A. Ioannidis, M. J. Khoury, J. Little, C. M. van Duijn, and A. C. J. W. Janssens. Scientific reporting is suboptimal for aspects that characterize

genetic risk prediction studies : A review of published articles based on the Genetic RIsk Prediction Studies statement. Journal of Clinical Epidemiology, 67(5) :487–499, May 2014. 61

[82] J. P. A. Ioannidis. Evolution and translation of research findings : From bench to where ? PLoS clinical trials, 1(7) :e36, Nov. 2006. 61

[83] J. P. A. Ioannidis. Expectations, validity, and reality in omics. Journal of Clinical Epidemiology, 63(9) :945–949, Sept. 2010. 61

[84] N. J. Jamieson, C. S. Hanson, M. A. Josephson, E. J. Gordon, J. C. Craig, F. Halleck, K. Budde, and A. Tong. Motivations, Challenges, and Attitudes to Self-management in Kidney Transplant Recipients : A Systematic Review of Qualitative Studies. American Journal of Kidney Diseases : The Official Journal of the National Kidney Foundation, 67 (3) :461–478, Mar. 2016. 30, 101

[85] B. Jennett and M. Bond. Assessment of outcome after severe brain damage. Lancet (London, England), 1(7905) :480–484, Mar. 1975. 125, 127

[86] M. J. Joyner and N. Paneth. Seven Questions for Personalized Medicine. JAMA, 314(10) : 999–1000, Sept. 2015. 32, 61, 74

[87] R. Kaboré, M. C. Haller, J. Harambat, G. Heinze, and K. Leffondré. Risk prediction models for graft failure in kidney transplantation : A systematic review. Nephrology, Dialysis, Transplantation : Official Publication of the European Dialysis and Transplant Association - European Renal Association, 32(suppl_2) :ii68–ii76, Apr. 2017. 101

[88] J. D. Kalbfleisch and R. L. Prentice. The Statistical Analysis of Failure Time Data. John Wiley & Sons, Jan. 2011. 130, 132

[89] S. Kim and P. S. Albert. A class of joint models for multivariate longitudinal measurements and a binary event. Biometrics, 72(3) :917–925, Sept. 2016. 115, 117

[90] P. Kind, J. E. Lafata, K. Matuszewski, and D. Raisch. The use of QALYs in clinical and patient decision-making : Issues and prospects. Value in Health : The Journal of the International Society for Pharmacoeconomics and Outcomes Research, 12 Suppl 1 : S27–30, Mar. 2009. 129, 130

[91] A. Król, L. Ferrer, J.-P. Pignon, C. Proust-Lima, M. Ducreux, O. Bouché, S. Michiels, and V. Rondeau. Joint model for left-censored longitudinal data, recurrent events and terminal event : Predictive abilities of tumor burden for cancer evolution with application to the FFCD 2000–05 trial. Biometrics, 72(3) :907–916, Sept. 2016. 121

[92] K. L. Lange, R. J. A. Little, and J. M. G. Taylor. Robust Statistical Modeling Using the t Distribution. Journal of the American Statistical Association, 84(408) :881–896, 1989. 115

[93] F. Le Borgne, C. Combescure, F. Gillaizeau, M. Giral, M. Chapal, B. Giraudeau, and Y. Foucher. Standardized and weighted time-dependent receiver operating characteristic curves to evaluate the intrinsic prognostic capacities of a marker by taking into account confounding factors. Statistical Methods in Medical Research, p. 962280217702416, Jan. 2017. 16, 30

[94] K. R. Lees, P. M. W. Bath, P. D. Schellinger, D. M. Kerr, R. Fulton, W. Hacke, D. Matchar, R. Sehra, D. Toni, and European Stroke Organization Outcomes Working Group. Contemporary outcome measures in acute stroke research : Choice of primary outcome measure. Stroke, 43(4) :1163–1170, Apr. 2012. 127

[95] N. Li, R. M. Elashoff, and G. Li. Robust joint modeling of longitudinal measurements and competing risks failure time data. Biometrical Journal. Biometrische Zeitschrift, 51(1) : 19–30, Feb. 2009. 115, 117

[96] M. Lorent, M. Giral, and Y. Foucher. Net time-dependent ROC curves : A solution for evaluating the accuracy of a marker to predict disease-related mortality. Statistics in Medicine, 33(14) :2379–2389, June 2014. 16, 30

[97] M. Lorent, H. Maalmi, P. Tessier, S. Supiot, E. Dantan, and Y. Foucher. Meta-analysis of predictive models to assess the clinical validity and utility for patient-centered medical decision making : Application to the CAncer of the Prostate Risk Assessment (CAPRA). BMC medical informatics and decision making, 19(1) :2, Jan. 2019. 33

[98] G. S. McHugh, I. Butcher, E. W. Steyerberg, A. Marmarou, J. Lu, H. F. Lingsma, J. Weir, A. I. R. Maas, and G. D. Murray. A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury : Results from the IMPACT Project. Clinical Trials (London, England), 7(1) :44–57, Feb. 2010. 127

[99] Methodology Committee of the Patient-Centered Outcomes Research Institute (PCORI). Methodological standards and patient-centeredness in comparative effectiveness research : The PCORI perspective. JAMA, 307(15) :1636–1640, Apr. 2012. 31, 101, 129

[100] S. Mook, M. Knauer, J. M. Bueno-de-Mesquita, V. P. Retel, J. Wesseling, S. C. Linn, L. J. Van't Veer, and E. J. Rutgers. Metastatic potential of T1 breast cancer can be predicted by the 70-gene MammaPrint signature. Annals of Surgical Oncology, 17(5) :1406–1413, May 2010. 38

[101] K. G. M. Moons, D. G. Altman, J. B. Reitsma, J. P. A. Ioannidis, P. Macaskill, E. W. Steyerberg, A. J. Vickers, D. F. Ransohoff, and G. S. Collins. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) : Explanation and elaboration. Annals of Internal Medicine, 162(1) :W1–73, Jan. 2015. 104

[102] O. Moranne, N. Maillard, C. Fafin, L. Thibaudin, E. Alamartine, and C. Mariat. Rate of renal graft function decline after one year is a strong predictor of all-cause mortality. American Journal of Transplantation : Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons, 13(3) :695–706, Mar. 2013. 47, 102

[103] A. M. O'Hare. Patient-Centered Care in Renal Medicine : Five Strategies to Meet the Challenge. American Journal of Kidney Diseases : The Official Journal of the National Kidney Foundation, 71(5) :732–736, May 2018. 101

[104] Optimising Analysis of Stroke Trials (OAST) Collaboration, P. M. W. Bath, L. J. Gray, T. Collier, S. Pocock, and J. Carpenter. Can we improve the statistical analysis of stroke trials ? Statistical reanalysis of functional outcomes in stroke trials. Stroke, 38(6) :1911–1915, June 2007. 127

[105] M. S. Pepe, H. Janes, G. Longton, W. Leisenring, and P. Newcomb. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. American Journal of Epidemiology, 159(9) :882–890, May 2004. 37

[106] J. C. Pinheiro, C. Liu, and Y. N. Wu. Efficient Algorithms for Robust Estimation in Linear Mixed-Effects Models Using the Multivariate t Distribution. Journal of Computational and Graphical Statistics, 10(2) :249–276, June 2001. 115, 117

[107] R. Pirracchio, M. Carone, M. R. Rigon, E. Caruana, A. Mebazaa, and S. Chevret. 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, 25(5) :1938–1954, Oct. 2016. 90

[108] F. K. Port, J. L. Bragg-Gresham, R. A. Metzger, D. M. Dykstra, B. W. Gillespie, E. W. Young, F. L. Delmonico, J. J. Wynn, R. M. Merion, R. A. Wolfe, and P. J. Held. Donor characteristics associated with reduced graft survival : An approach to expanding the pool of kidney donors. Transplantation, 74(9) :1281–1286, Nov. 2002. 90

[109] C. Proust-Lima and J. M. G. Taylor. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA : A joint modeling approach. Biostatistics (Oxford, England), 10(3) :535–549, July 2009. 114

[110] A.-H. Querard, Y. Foucher, C. Combescure, E. Dantan, D. Larmet, M. Lorent, L.-M. Pouteau, M. Giral, and F. Gillaizeau. Comparison of survival outcomes between Expanded Criteria Donor and Standard Criteria Donor kidney transplant recipients : A systematic review and meta-analysis. Transplant International : Official Journal of the European Society for Organ Transplantation, 29(4) :403–415, Apr. 2016. 24, 33

[111] A. H. Querard, F. Le Borgne, A. Dion, M. Giral, G. Mourad, V. Garrigue, L. Rostaing, N. Kamar, A. Loupy, C. Legendre, E. Morelon, F. Buron, Y. Foucher, and E. Dantan. Propensity score-based comparison of the graft failure risk between kidney transplant recipients of standard and expanded criteria donor grafts : Toward increasing the pool of marginal donors. American Journal of Transplantation : Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons, 18(5) :1151–1157, May 2018. 24, 33, 34, 89

[112] P. S. Rao, D. E. Schaubel, M. K. Guidinger, K. A. Andreoni, R. A. Wolfe, R. M. Merion, F. K. Port, and R. S. Sung. A comprehensive risk quantification score for deceased donor kidneys : The kidney donor risk index. Transplantation, 88(2) :231–236, July 2009. 34

[113] R. D. Riley, J. A. Hayden, E. W. Steyerberg, K. G. M. Moons, K. Abrams, P. A. Kyzas, N. Malats, A. Briggs, S. Schroter, D. G. Altman, H. Hemingway, and PROGRESS Group. Prognosis Research Strategy (PROGRESS) 2 : Prognostic factor research. PLoS medicine, 10(2) :e1001380, 2013. 124

[114] D. Rizopoulos. Joint Models for Longitudinal and Time-to-Event Data : With Applications in R. CRC Press, 2012. 20, 47, 62, 102, 114, 116, 132, 133

[115] D. Rizopoulos and P. Ghosh. A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. Statistics in Medicine, 30(12) :1366–1380, May 2011. 47, 102

[116] J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical Modelling, 7(9-12) :1393–1512, 1986. 135

[117] B. Roozenbeek, H. F. Lingsma, P. Perel, P. Edwards, I. Roberts, G. D. Murray, A. I. Maas, E. W. Steyerberg, IMPACT (International Mission on Prognosis and Clinical Trial Design in Traumatic Brain Injury) Study Group, and CRASH (Corticosteroid Randomisation After Significant Head Injury) Trial Collaborators. The added value of ordinal analysis in clinical trials : An example in traumatic brain injury. Critical Care (London, England), 15(3) :R127, 2011. 127

[118] B. Sautenet, A. Tong, J. R. Chapman, A. N. Warrens, D. Rosenbloom, G. Wong, J. Gill, K. Budde, L. Rostaing, L. Marson, M. A. Josephson, P. P. Reese, T. L. Pruett, N. Evangelidis, and J. C. Craig. Range and consistency of outcomes reported in randomized trials conducted in kidney transplant recipients : A systematic review. Transplantation, May 2018. 101

[119] Q. Shao and X. Zhou. A new parametric model for survival data with long-term survivors. Statistics in Medicine, 23(22) :3525–3543, Nov. 2004. 126

[120] J. J. Shuster. Median follow-up in clinical trials. Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology, 9(1) :191–192, Jan. 1991. 104

[121] J. M. Snowden, S. Rose, and K. M. Mortimer. Implementation of G-computation on a simulated data set : Demonstration of a causal inference technique. American Journal of Epidemiology, 173(7) :731–738, Apr. 2011. 135

[122] S. L. Spruance, J. E. Reid, M. Grace, and M. Samore. Hazard ratio in clinical trials. Antimicrobial Agents and Chemotherapy, 48(8) :2787–2792, Aug. 2004. 37

[123] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan. Assessing the performance of prediction models : A framework for traditional and novel measures. Epidemiology (Cambridge, Mass.), 21(1) :128–138, Jan. 2010. 30, 62

[124] E. W. Steyerberg, K. G. M. Moons, D. A. van der Windt, J. A. Hayden, P. Perel, S. Schroter, R. D. Riley, H. Hemingway, D. G. Altman, and PROGRESS Group. Prognosis Research Strategy (PROGRESS) 3 : Prognostic model research. PLoS medicine, 10(2) : e1001381, 2013. 124

[125] F. Subtil and M. Rabilloud. An enhancement of ROC curves made them clinically relevant for diagnostic-test comparison and optimal-threshold determination. Journal of Clinical Epidemiology, 68(7) :752–759, July 2015. 73

[126] B. C. Sutradhar and M. M. Ali. Estimation of the parameters of a regression model with a multivariate t error variable. Communications in Statistics - Theory and Methods, 15 (2) :429–450, Jan. 1986. 115

[127] F. Tagliaferri, C. Compagnone, M. Korsic, F. Servadei, and J. Kraus. A systematic review of brain injury epidemiology in Europe. Acta Neurochirurgica, 148(3) :255–268 ; discussion 268, Mar. 2006. 123

[128] J. M. G. Taylor, Y. Park, D. P. Ankerst, C. Proust-Lima, S. Williams, L. Kestin, K. Bae, T. Pickles, and H. Sandler. Real-time individual predictions of prostate cancer recurrence using joint models. Biometrics, 69(1) :206–213, Mar. 2013. 114, 115, 117

[129] G. M. Teasdale, L. E. Pettigrew, J. T. Wilson, G. Murray, and B. Jennett. Analyzing outcome of treatment of severe head injury : A review and update on advancing the use of the Glasgow Outcome Scale. Journal of Neurotrauma, 15(8) :587–597, Aug. 1998. 127

[130] A. Towse. Net Clinical Benefit : The Art and Science of Jointly Estimating Benefits and Risks of Medical Treatment. Value in Health, 13 :S30–S32, June 2010. 130

[131] G. Tripepi, K. J. Jager, F. W. Dekker, and C. Zoccali. Testing for causality and prognosis : Etiological and prognostic models. Kidney International, 74(12) :1512–1515, 2008. 30

[132] C. J. van Asch, M. J. Luitse, G. J. Rinkel, I. van der Tweel, A. Algra, and C. J. Klijn. Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time, according to age, sex, and ethnic origin : A systematic review and meta-analysis. The Lancet. Neurology, 9(2) :167–176, Feb. 2010. 123

[133] A. Wang, R. A. Nianogo, and O. A. Arah. G-computation of average treatment effects on the treated and the untreated. BMC medical research methodology, 17(1) :3, Jan. 2017. 135

[134] J. H. Ware. The limitations of risk factors as prognostic tools. The New England Journal of Medicine, 355(25) :2615–2617, Dec. 2006. 37

[135] M. C. Weinstein. Clinical Decision Analysis. Saunders (philadelphia) edition, 1980. 129

[136] M. C. Weinstein, G. Torrance, and A. McGuire. QALYs : The basics. Value in Health : The Journal of the International Society for Pharmacoeconomics and Outcomes Research, 12 Suppl 1 :S5–9, Mar. 2009. 32, 74, 129

[137] S. N. Wood. Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. Journal of the American Statistical Association, 99(467) :673–686, Sept. 2004. 120, 146

[138] L. Wu. Mixed Effects Models for Complex Data. Chapman and Hall/CRC, Nov. 2009. 115

[139] M. S. Wulfsohn and A. A. Tsiatis. A joint model for survival and longitudinal data measured with error. Biometrics, 53(1) :330–339, Mar. 1997. 47, 116, 132

[140] W. J. Youden. Index for rating diagnostic tests. Cancer, 3(1) :32–35, Jan. 1950. 73

[141] B. Yu, R. C. Tiwari, K. A. Cronin, and E. J. Feuer. Cure fraction estimation from the mixture cure models for grouped survival data. Statistics in Medicine, 23(11) :1733–1747, June 2004. 126

[142] R. C. Ziegelstein. Personomics. JAMA internal medicine, 175(6) :888–889, June 2015. 32, 74, 101