Prognosis of cancer patients : input of standard and joint frailty models

Audrey Mauguen

université de BORDEAUX

Thèse

présentée pour obtenir le grade de

docteur de l'université de Bordeaux

Ecole doctorale Sociétés, Politique, Santé Publique
Spécialité Santé Publique, option Biostatistique

Par Audrey Mauguen

Prognosis of cancer patients
Input of standard and joint frailty models

Pronostic en cancérologie
Apport des modèles à fragilité standards et conjoints

Sous la direction de Virginie Rondeau

Soutenue le 28 novembre 2014, devant les membres du jury :

Mme	Jacqmin-Gadda Hélène	DR, INSERM U897, Bordeaux	Présidente
M.	Putter Hein	Pr, Leiden UMC, Leiden	Rapporteur
M.	Rizopoulos Dimitris	Ass Pr, Erasmus UMC, Rotterdam	Rapporteur
Mme	Mathoulin-Pélissier Simone	Pr, INSERM CIC-EC7, Bordeaux	Examinateur
M.	Bonnetain Franck	Pr, CHU Besançon, Besançon	Examinateur
M.	Rachet Bernard	Dr, LSHTM, London	Examinateur
Mme	Rondeau Virginie	Dr, INSERM U897, Bordeaux	Directrice de thèse

Abstract

Research on cancer treatment has been evolving for last years in one main direction: personalised medicine. The treatment choice must be done according to the patients' and tumours' characteristics. This goal requires some biostatistical developments, in order to assess prognostic models and eventually propose the best one. In a first part, we consider the problem of assessing a prognostic score when multicentre data are used. We extended two concordance measures to clustered data in the context of shared frailty model. Both the between-cluster and the within-cluster levels are studied, and the impact of the cluster number and size on the performance of the measures is investigated. In a second part, we propose to improve the prediction of the risk of death accounting for the previous observed relapses. For that, we develop predictions from a joint model for a recurrent event and a terminal event. The proposed individual prediction is dynamic, both the time and the horizon of prediction can evolve, so that the prediction can be updated at each new event time. The prediction is developed on a French hospital series, and externally validated on population-based data from English and Dutch cancer registries. Its performances are compared to those of a landmarking approach. In a third part, we explore the use of the proposed prediction to reduce the clinical trial duration. The non-observed death times of the last included patients are imputed using the information of the patients with longer follow-up. We compared three methods to impute the data: a survival mean time, a time sampled from the parametric distribution and a time sampled from a non-parametric distribution of the survival times. The comparison is made in terms of parameters estimation (coefficient and standard-error), type-I error and power.

Key words: cancer; clinical trial; joint frailty model; prediction; recurrent event; shared frailty model; survival analysis; validation.

Résumé

La recherche sur le traitement des cancers a évolué durant les dernières années principalement dans une direction : la médecine personnalisée. Idéalement, le choix du traitement doit être basé sur les caractéristiques du patient et de sa tumeur. Cet objectif nécessite des développements biostatistiques, pour pouvoir évaluer les modèles pronostiques, et *in fine* proposer le meilleur. Dans une première partie, nous considérons le problème de l'évaluation d'un score pronostique dans le cadre de données multicentriques. Nous étendons deux mesures de concordance aux données groupées analysées par un modèle à fragilité partagée. Les deux niveaux inter et intra-groupe sont étudiés, et l'impact du nombre et de la taille des groupes sur les performances des mesures est analysé. Dans une deuxième partie, nous proposons d'améliorer la prédiction du risque de décès en tenant compte des rechutes précédemment observées. Pour cela nous développons une prédiction issue d'un modèle conjoint pour un événement récurrent et un événement terminal. Les prédictions individuelles proposées sont dynamiques, dans le sens où le temps et la fenêtre de prédiction peuvent varier, afin de pouvoir mettre à jour la prédiction lors de la survenue de nouveaux événements. Les prédictions sont développées sur une série hospitalière française, et une validation externe est faite sur des données de population générale issues de registres de cancer anglais et néerlandais. Leurs performances sont comparées à celles d'une prédiction issue d'une approche *landmark*. Dans une troisième partie, nous explorons l'utilisation de la prédiction proposée pour diminuer la durée des essais cliniques. Les temps de décès non observés des derniers patients inclus sont imputés en utilisant l'information des patients ayant un suivi plus long. Nous comparons trois méthodes d'imputation : un temps de survie moyen, un temps échantillonné dans une distribution paramétrique et un temps échantillonné dans une distribution non-paramétrique des temps de survie. Les méthodes sont comparées en termes d'estimation des paramètres (coefficient et écart-type), de risque de première espèce et de puissance.

Mots-clés : analyse de survie; cancer; essais cliniques; événement récurrent; modèle à fragilité partagée; modèle conjoint; prédiction; validation.

Remerciements

A ma directrice de thèse

A **Virginie**, un grand merci pour ces trois années à tes côtés. Tu m'as fait découvrir les modèles à fragilité et les mystères des modèles conjoints. Tu m'as surtout aidée à comprendre le métier de chercheuse et donné envie de continuer. Merci pour ta patience, ton enthousiasme et ton optimisme. Merci pour ta confiance et ton soutien. Enfin, merci de m'avoir donné l'opportunité de prendre goût à la vie en anglais.

Aux membres du jury

A **Hélène Jacqmin-Gadda**, merci de présider mon jury de thèse. Tes remarques toujours pertinentes m'ont permis de prendre du recul sur mon travail. C'est un plaisir quotidien de travailler dans ton équipe.

To **Hein Putter**, thank you for having accepted to review this thesis. Your work and your presentations are a model for me. I am honoured that you take part in my jury.

To **Dimitris Rizopoulos**, thank you for having accepted to review this thesis. Your work was a major inspiration to a part of this thesis. I am honoured that you take part in my jury.

A **Simone Mathoulin-Pélissier**, merci d'avoir accepté d'examiner cette thèse. Merci d'avoir apporté un regard médical et épidémiologiste plus concret sur mon travail. Les applications des méthodes statistiques sont pour moi essentielles.

A **Franck Bonnetain**, merci d'avoir accepté d'examiner cette thèse. Vos connaissances à la fois du cancer et des biostatistiques seront précieuses dans l'évaluation de ce travail.

A **Bernard Rachet**, un grand merci pour cette collaboration. Merci de m'avoir accueillie dans ton équipe à Londres et merci de m'avoir aidée à améliorer ma compréhension du cancer et ses enjeux.

Aux Bordelais

Un merci à l'ensemble de l'équipe de biostatistique de l'INSERM U897, en particulier à Daniel, pour m'avoir accueillie dans son équipe, et avoir créé une ambiance de travail aussi agréable. Merci à mes anciens professeurs pour leurs précieux enseignements, en particulier Pierre et Alioum. Un merci particulier à Cécile, toujours disponible pour partager ses meilleures lectures, mais surtout son expérience et répondre aux questions les plus basiques comme les plus farfelues, avec le sourire. Merci à Amadou et Alexandre, pour m'avoir aidée à comprendre et utiliser *frailtypack*, et aidée à debugger mes codes.

Merci aux coburalistes du bureau 45 : Linda, pour ta bonne humeur; Yassin, pour ta gentillesse constante; Mélanie, pour avoir la pêche en permanence; Julie, pour ton sourire et avoir partagé la montée de pression des dernières semaines – à deux, c'est mieux ! Nini pour ce vent de fraîcheur; Lingling et Agnieszka. Un merci particulier à Paulo : pour les conversations statistiques qui m'auront aidée à comprendre tellement de choses; merci pour les autres conversations, et les soirées improvisées qui sont souvent les meilleures. Cette thèse n'aurait jamais été la même sans toi.

Merci aux *student* et *ceux du troisième*, en particulier à Bobo et Didi pour avoir égayé mon quotidien; à Célia, Lucie, Mathilde, Emilie, Henri, Loïc, Ritchie, Perrine, Chloé, Aïssatou et les autres pour m'avoir donné trop d'occasions de ne pas dormir en semaine.

Merci aux autres (ex-)Ispediens : à Fanny et Hind, pour ces moments de pause qui me permettent vraiment de décompresser et pour les soirées à rigoler, à Aurore, à Clément, à Majid, Laura, Juan et Loulou pour la bonne humeur que vous apportez, à Reto pour les cours d'anglais et de culture écossaise improvisés.

Merci à Mamady Cisse, Clément Dubos et Gabriel Ferrand, étudiants en master 1, pour votre recherche bibliographique sur les critères de substitution.

Merci à Céline, Elodie, Michael, Alex et Sophie, pour les précieux moments passés avec vous, qui m'aident à m'échapper de l'Isped. Merci à Vanessa et Michel, pour les moments passés et les débats Paris vs Bordeaux.

To the Londoners

Many thanks to the Cancer Survival Group of the LSHTM for their kindness and the nice 6 months I spent there; in particular, thank you to Michel Coleman for having welcomed me in his team, and for favouring such a pleasant international atmosphere.

Aux Parisiens/Villejuifois

Merci à tous les membres du SBE de l'IGR, et en particulier à Marie-Cécile et Jean-Pierre pour m'avoir appris les bases du métier de biostatisticien, et avoir fait germer l'idée de me lancer dans une thèse. Merci à Charlotte, Monia et Vanessa pour m'avoir montré le chemin et pour le pain, vin, fromage. Merci à Romain pour m'avoir encouragée à commencer cette thèse. Merci à ASDJ de m'avoir fait redécouvrir Paris en touriste.

Aux Grenoblois

Merci à Salim Kobeissi et Alain Latour, pour vos enseignements et pour m'avoir fait découvrir les applications des statistiques dans le domaine de la santé, alors que j'étais tournée vers l'économie.

A mes proches

Aux "d'enfance": Anna, Jess et Ju. Merci d'être toujours là. Bientôt trente ans que je vous ai autour de moi, et j'espère que ce n'est que le début !

Merci mes sisters, Céline et Gaëlle, merci à Franck et Florian, et à Thibaud, Arthur et Alexis. Les moments passés avec vous sont toujours des perles qui donnent meilleur goût à la vie (mais pas aux piña colada). Enfin, à mes parents, merci de m'avoir donné la possibilité de faire mes études et de m'avoir laissée libre de mes choix. Et de m'offrir de si belles parenthèses quand j'en ai besoin : à ma mère et Jean-Pierre, merci pour le chant des cigales et la douceur provençale; à mon père merci pour les paysages riches en iode de la Bretagne.

Bande originale : Passenger (*All The Little Lights*), Regina Spektor (*What We Saw From The Cheap Seats*, *Far* et *Begin To Hope*), Angus & Julia Stone (*Down The Way*).

Scientific valorisation

Articles

Thesis publications

[1] Mauguen A, Collette S, Pignon J-P, and Rondeau V (2013). Concordance Measures in Shared Frailty Models: Application to Clustered Data in Cancer Prognosis. *Statistics in Medicine* 32(27):4803–20. doi:10.1002/sim.5852.

[2] Mauguen A, Rachet B, Mathoulin-Pélissier S, MacGrogan G, Laurent A, and Rondeau V (2013). Dynamic Prediction of Risk of Death Using History of Cancer Recurrences in Joint Frailty Models. *Statistics in Medicine* 32(30):5366–80. doi:10.1002/sim.5980.

[3] Mauguen A, Collette S, Pignon J-P, and Rondeau V (2014). Reply to 'Interpretation of concordance measures for clustered data'. *Statistics in Medicine* 33(4):717–18. [Reply to letter] doi:10.1002/sim.6022.

[4] Mauguen A, Rachet B, Mathoulin-Pélissier S, Lawrence G, Siesling S, MacGrogan G, Laurent A, and Rondeau V (2014). Death prediction after breast cancer relapses using joint models: validation on English and Dutch population-based data and comparison with landmark approach. *Submitted.*

[5] Mauguen A, Michiels S, and Rondeau V (2014). Predict treatment effects based on cancer relapses information: imputation using joint modelling. *In preparation.*

Related articles

[a] Mazroui Y, Mauguen A, Mathoulin-Pélissier S, MacGrogan G, Brouste V, and Rondeau V (2014). Time-varying coefficients in a multivariate frailty model: Application to breast cancer recurrences of several types and death. *In minor revision for Lifetime Data Analysis.*

[b] Rondeau V, Laurent A, Mauguen A, Berr C, and Helmer C (2014). Dynamic prediction models for clustered and interval-censored outcomes: investigating the intra-couple correlation in the risk of dementia. *Submitted.*

Presentation in congresses and seminars

Oral presentations in congresses

[i] Mauguen A, Collette S, Pignon J-P, and Rondeau V. Concordance measures in shared frailty models: application to clustered cancer data [French]. *EPICLIN 6 / 19èmes Journées de Statisticiens de CLCC*, May 2012, Lyon, France.

[ii] Mauguen A, Mathoulin-Pélissier S, and Rondeau V. Dynamic prediction for risk of death using history of cancer recurrences in joint frailty models. *International Society for Clinical Biostatistics ISCB 2012*, August 2012, Bergen, Norway.

[iii] Mazroui Y, Mauguen A (speaker), Mathoulin-Pélissier S, Brouste V, MacGrogan G, and Rondeau V. Multivariate frailty models for two types of recurrent events with a dependent terminal event with possibly time-varying coefficients: Application to breast cancer data. *International Society for Clinical Biostatistics ISCB 2012*, August 2012, Bergen, Norway.

[iv] Mauguen A, Rachet B, Mathoulin-Pélissier S, Lawrence G, Siesling S, MacGrogan G, Laurent A, and Rondeau V. Joint models for dynamic prediction of the risk of death: prediction of death after a breast cancer in France, England and the Netherlands [French]. *EPICLIN 8 / 21èmes Journées de Statisticiens de CLCC*, May 2014, Bordeaux, France.

[v] Mauguen A, Michiels S, and Rondeau V. Prediction of the Treatment Effect on Overall Survival Using an Intermediate Time-to-Event Endpoint Use of the joint frailty model. *International Biometric Conference IBC 2014*, July 2014, Florence, Italy.

Invited seminars

[i] Mauguen A, Rachet B, Mathoulin-Pélissier S, MacGrogan G, Laurent A, Rondeau V (speaker). Dynamic prognostic tools using joint models or recurrent and terminal events. *Workshop Dynamic prediction for repeated markers and repeated events*, October 2013, Bordeaux, France.

[ii] Mauguen A(co-speaker), Rachet B, Mathoulin-Pélissier S, Lawrence G, Siesling S, MacGrogan G, Laurent A, and Rondeau V (co-speaker). Dynamic prediction using joint models for recurrent and terminal events: Evolution after a breast cancer. *London School of Hygiene and Tropical Medicine - Centre for Statistical Methodology Seminar*, December 2013, London, UK.

Posters

[i] Mauguen A, Rachet B, Mathoulin-Pélissier S, MacGrogan G, Lawrence G, Walton J, Lagord C, and Rondeau V. External validation of dynamic prediction of the risk of death using relapses information. *International Society for Clinical Biostatistics ISCB 2013*, August 2013, Munich, Germany.

[ii] Mauguen A, Rachet B, Mathoulin-Pélissier S, MacGrogan G, Lawrence G, Walton J, Lagord C, and Rondeau V. External validation of dynamic prediction of the risk of death using relapses information. *Workshop Dynamic prediction for repeated markers and repeated events*, October 2013, Bordeaux, France.

[iii] Mauguen A, Rachet B, Mathoulin-Pélissier S, MacGrogan G, Lawrence G, Walton J, Lagord C, and Rondeau V. External validation of dynamic prediction of the risk of death using relapses information. *Workshop Statistical Analysis of Multi-Outcome Data*, June 2014, Cambridge, UK.

R package

Functions *Cmeasures* and *prediction* added to the  package *frailtypack*.

Research visit

For this thesis purpose, I spent 6 months as a visiting scientist in the Cancer Survival Group at the London School of Hygiene and Tropical Medicine (London, UK), to analyse

the data from the West Midlands registry and from the Comprehensive Cancer Centre The Netherlands (IKNL) as an external validation of our predictions.

Résumé substantiel en français

Introduction

Les challenges de la recherche en cancérologie

Le mot *cancer* recouvre en réalité plusieurs maladies, avec des caractéristiques différentes en fonction de la localisation ou de l'histologie, par exemple. Les tumeurs sont dues à une multiplication incontrôlée des cellules. Des facteurs de risque environnementaux du cancer sont aujourd'hui bien connus, tels que le tabac, qui est le carcinogène le mieux identifié (Boyle et al., 2008). D'autres facteurs sont encore en doute, comme les pesticides ou les ondes. Enfin, l'étiologie génétique des cancers est aussi à l'étude. En attendant que tous les cancers puissent être expliqués et prévenus, il y avait 14 millions de personnes diagnostiquées avec un cancer dans le monde en 2012 et 8 millions de décès associés (Ferlay et al., 2013). Cela représente une augmentation d'environ 11% de la prévalence et de 8% de la mortalité par rapport à 2008. Le nombre croissant de cas peut être expliqué par l'augmentation de la population, par son vieillissement et par une incidence croissante du cancer. D'après les projections, le nombre de nouveaux cas de cancer pourrait être de 20 à 26 millions en 2030 (Boyle et al., 2008). Parmi tous les cancers, le cancer du sein est le plus courant chez les femmes et représente 1.7 millions de nouveaux cas en 2012. Depuis 2008, son incidence a augmenté de plus de 20% et la mortalité associée de 14%. Dans ce contexte, une part importante de la recherche sur le cancer concerne le traitement des patients.

Le traitement du cancer inclut plusieurs options, telles que la chirurgie, la radiothérapie, la chimiothérapie et plus récemment les thérapies ciblées. Ces traitements peuvent être combinés, et parmi la quantité d'options disponibles, les cliniciens basent leurs choix en priorité sur la preuve de leur efficacité. Cependant, certains traitements destinés

à être très agressifs pour la tumeur sont aussi très agressifs pour les cellules normales, entraînant d'importants effets secondaires. C'est le cas de certaines chimiothérapies, mais aussi de la radiothérapie. Si d'autres options moins agressives sont disponibles, elles peuvent être préférées par les médecins comme première ligne de traitement pour les patients ayant un bon pronostic, c'est-à-dire de bonnes chances de survie. D'autre part, les thérapies ciblées ne ciblent que des tumeurs spécifiques et ne sont donc pas efficaces pour tous les patients. En conséquence, le choix du traitement du cancer doit être fait en fonction des caractéristiques du patient et de la tumeur. La recherche sur les traitements du cancer a donc évolué ces dernières années dans une direction principale : la médecine personnalisée. Le but ultime est de précisément identifier chaque type de cancer, pour proposer aux patients le traitement qui va spécifiquement cibler leur tumeur. La tendance est donc à la définition de sous-groupes et de sous-types de cancer, à trouver de nouveaux biomarqueurs permettant soit d'identifier les patients avec les meilleures chances de survie, soit de définir des types de tumeurs qui vont répondre différemment aux traitements. Dans le cancer du sein par exemple, deux classifications tumorales ont été récemment proposées (Badve et al., 2011). La première définit un groupe de tumeurs triple-négatives (pas de récepteurs à œstrogène ni progestérone, pas d'expression de HER-2). La seconde classification consiste en cinq sous-types moléculaires : luminal A, luminal B, HER2, basal et proche du sein normal. Les deux classifications se basent sur les spécificités de la tumeur et les biomarqueurs qu'elle exprime. Par ailleurs, les facteurs qui ont été montrés associés à l'évolution clinique des patientes avec un cancer du sein sont la taille de la tumeur, le type histologique, le grade histologique, la présence d'invasion vasculaire et la présence d'un envahissement ganglionnaire (IARC, 2002, chapitre 1). Ces facteurs sont utiles pour 1- décider du traitement le plus approprié, 2- planifier le dépistage du cancer du sein et 3- surveiller les changements de tendance dans l'incidence de la maladie. Les autres facteurs pronostiques connus incluent la présence de métastases, la fraction de croissance, la présence de récepteurs hormonaux ou de facteurs de croissance et, plus généralement, les autres caractéristiques moléculaires.

Le cancer est une maladie progressive. Après le diagnostic et le traitement de la tumeur initiale, des rechutes locales ou des métastases à distance peuvent survenir, parfois après peu de temps, parfois après des années. La survenue de tels événements liés à la maladie montre une réapparition ou une évolution de la tumeur. Un patient peut connaître successivement plusieurs événements après sa première ligne de traitement. En général, chaque rechute est suivie par une nouvelle ligne de traitement, choisie une nouvelle fois en fonction des caractéristiques du patient et de la tumeur, mais aussi en fonction de l'histoire de la maladie. Au final, la principale question d'intérêt lors du traitement

d'un patient est : combien de temps le patient va-t-il survivre à sa tumeur ? Quelles sont les chances qu'il soit en vie dans 5, 10 ou 20 ans ? Quelle va être l'amélioration de sa survie en utilisant un traitement plutôt qu'un autre, plutôt qu'aucun ? Augmenter la survie globale des patients reste l'objectif principal. Cependant, cela peut passer par l'allongement du temps passé sans rechute. C'est pourquoi la recherche est maintenant souvent tournée vers l'étude de ces événements intermédiaires.

Cette évolution de la recherche clinique doit s'accompagner d'une évolution parallèle des méthodes biostatistiques pour répondre aux nouveaux besoins. Le premier challenge soulevé par la recherche clinique en cancérologie est d'être capable de proposer le meilleur modèle pronostique aux médecins. Cela veut dire être capable de proposer un modèle approprié, qui étudie correctement le critère d'intérêt, incluant toutes les informations afin d'obtenir un outil utile, précis mais surtout facilement utilisable en pratique. Cela veut aussi dire que nous sommes capables d'évaluer les outils que nous proposons, de les comparer de façon juste pour finalement sélectionner le meilleur. La recherche pronostique est une part du champ des biostatistiques depuis de nombreuses années maintenant. Cependant, l'intérêt a été plus appuyé ces cinq dernières années, avec la publication de trois séries d'articles essayant de définir des recommandations à partir de l'importante littérature disponible jusque-là: *Prognosis and prognostic research* publié dans BMJ en 2009 (Moons et al., 2009; Royston et al., 2009; Altman et al., 2009; Moons et al., 2009); *Risk prediction models* publié dans Heart en 2012 (Moons et al., 2012, 2012); et *PROGRESS* publié dans BMJ et PLoS Medicine en 2013 (Hemingway et al., 2013; Riley et al., 2013; Steyerberg et al., 2013; Hingorani et al., 2013). Essentiellement, les principales étapes nécessaires à l'obtention d'un modèle pronostique pertinent sont les suivantes : 1) l'identification des facteurs pronostiques potentiels, 2) le choix d'un modèle adéquat, 3) la sélection des facteurs et l'estimation de leurs effets (développement du modèle), 4) application du modèle sur une nouvelle population et évaluation de ses performances (validation du modèle) et 5) mise à jour du modèle. L'étape 4 de validation est cruciale pour évaluer la possibilité d'utiliser le modèle en pratique, dans différentes populations (Altman and Royston, 2000; Konig et al., 2007). En recherche clinique et épidémiologique, la précision est recherchée, ce qui pousse à la conduite de larges études. Les sources d'information sont combinées, conduisant à des données groupées. Cela a pour l'instant bien été intégré dans les étapes de développement du modèle prédictif. Mais il y a un manque de méthodes de validation appropriées pour ce type de données.

La deuxième question posée est comment faire des prédictions qui tiennent compte non seulement des caractéristiques des patients à un temps donné, mais aussi de toute son histoire. Ce genre de prédiction est d'un intérêt particulier pour les patients ayant

un cancer, qui peuvent connaître plusieurs rechutes. Il est important de considérer ces rechutes qui peuvent être corrélées à la survie globale des patients. Il est également important que le patient se voie proposer le traitement le plus adapté pour lui chaque fois qu'il en a besoin, y compris après chaque rechute. C'est pourquoi l'information sur la rechute est importante. Pour prendre en compte cette information, le *landmarking* est une méthode simple et utile (Van Houwelingen, 2007; Van Houwelingen and Putter, 2011). Cependant, avec cette méthode, l'histoire du patient doit être simplifiée pour être incluse dans un modèle tel qu'un modèle de Cox. De précédents travaux ont utilisé les modèles conjoints pour données longitudinales et un temps de survie pour prédire un temps de survie en considérant l'évolution d'un biomarqueur longitudinal (Faucett et al., 2002; Proust-Lima and Taylor, 2009; Rizopoulos, 2011; McLain et al., 2012; Taylor et al., 2013). A notre connaissance, aucun travail similaire n'a été publié en utilisant un modèle conjoint pour un événement récurrent et un événement terminal.

La troisième question porte sur la durée des essais cliniques randomisés. En effet, pour définir quel traitement est le plus efficace, des essais cliniques doivent être conduits. Cependant, l'utilisation de la survie globale comme critère de jugement principal conduit à des essais durant plusieurs années avant de pouvoir conclure. Dans d'autres maladies que le cancer, le critère de jugement peut être trop cher ou trop complexe à mesurer. Pour répondre à ces problèmes, l'utilisation de critères de substitution s'est développée. Il s'agit de critères observables à plus court terme que le critère clinique d'intérêt, mais qui lui sont corrélés et sur lesquels l'effet observé du traitement peut renseigner sur l'effet du traitement sur le critère principal. Par exemple, la survie sans progression a été validée comme critère de substitution pour la survie globale (décès toutes causes) dans plusieurs cancers. En effet, les progressions du cancer sont observables à plus court terme que le décès, elles lui sont corrélées et l'efficacité du traitement sur les progressions peut se traduire par une efficacité sur la survie globale. La survie sans progression est donc un excellent candidat comme critère de substitution de la survie globale. Jusqu'à présent, les recherches ont principalement porté sur comment valider de tels critères de substitution. Cela a débuté avec Prentice (1989), qui a établi une liste de quatre critères de corrélation qui doivent être satisfaits pour qu'un critère soit validé comme critère de substitution. Ces critères ont été ensuite étudiés par Freedman et al. (1992) et Buyse and Molenberghs (1998). Dix ans plus tard, Buyse et al. (2000) ont proposé une approche méta-analytique à la validation de critères de substitution, qui est réduite à l'étude de deux corrélations : une corrélation entre les deux critères au niveau individuel, et une corrélation entre les effets du traitement sur les deux critères au niveau essais. Une fois qu'un critère est validé, il devrait être intégré à la pratique clinique. Considérant l'utilisation de

critères de substitution en pratique, deux directions ont été suivies. La première est d'utiliser les critères de substitution comme critère principal dans les essais, d'estimer l'effet du traitement sur eux, puis d'extrapoler l'effet du traitement sur la survie globale. C'est le concept d'*effet seuil substitutif*, proposé par Burzykowski and Buyse (2006) dans l'approche méta-analytique. La seconde méthode a été proposée par Faucett et al. (2002) dans la recherche sur le SIDA. La survie globale reste le critère de jugement principal de l'essai. L'information d'une variable auxiliaire –pour Faucett il s'agissait du niveau de CD4, pour nous il s'agit des rechutes de cancer– est utilisée pour imputer les temps de décès encore non observés. Ensuite, l'effet du traitement est estimé sur la survie globale en utilisant à la fois les temps de décès observés et imputés. Cette approche avait été proposée en utilisant un modèle conjoint pour données longitudinales et un temps de survie. Une approche similaire a été récemment proposée pour tenir compte d'événements intermédiaires tels que les rechutes en utilisant un modèle de guérison (Conlon et al., 2011).

Objectifs et plan de la thèse

Cette thèse s'inscrit dans le contexte de la médecine personnalisée dans la recherche en cancérologie. Comme vu au-dessus, plusieurs challenges ont été posés par la volonté de proposer le meilleur traitement à chaque patient. Notre objectif principal était d'évaluer l'apport potentiel de la modélisation conjointe pour un événement récurrent et un événement terminal. Cette modélisation semble le cadre idéal pour étudier le décès en tenant compte des événements intermédiaires de la maladie et des autres caractéristiques du patient.

Notre premier objectif était de proposer une mesure pour étudier la discrimination d'une prédiction proposée en présence de données groupées. Un deuxième objectif était de dériver des prédictions des modèles conjoints et d'évaluer leurs performances. Enfin, un troisième objectif était d'évaluer leur utilité dans le cadre des essais cliniques.

La première partie de la thèse présente les méthodes principales existantes en analyse de survie et en recherche pronostique. Elle contient des détails sur les méthodes classiques, mais aussi plus récentes, qui ont été utilisées dans nos développements.

La deuxième partie porte sur l'adaptation des mesures de concordance au modèle de survie à fragilité partagée. La concordance est une mesure de discrimination, largement utilisée, qui mesure si la prédiction proposée est cohérente avec les observations, c'est-à-dire si le risque prédit est en effet plus élevé pour les patients qui vont décéder plus tôt. Différentes estimations de cette quantité ont été proposées. Parmi elles, la plus commune

est le c-index, une mesure non-paramétrique basée sur le compte des paires de patients dites *concordantes*. Deux versions du c-index ont été proposées : une mesure brute (Harrell et al., 1982, 1996), dont on a montré qu'elle dépend de la censure, et une mesure pondérée (Uno et al., 2011). Une deuxième estimation proposée, appelée estimation de la probabilité de concordance, utilise les propriétés du modèle à risques proportionnels de Cox pour estimer la probabilité de concordance (Gönen and Heller, 2005). Les deux types d'estimation ont été initialement proposés pour des temps de survie non corrélés. Seul le c-index brut a été précédemment adapté pour temps de survie corrélés (Van Oirbeek and Lesaffre, 2010). Nous proposons ici une extension des deux autres mesures, qui sont indépendantes de la quantité de censure, aux temps de survie corrélés étudiés par des modèles à fragilité partagée. Cette extension tient compte des deux niveaux de données, intra-groupe et inter-groupe.

Dans une troisième partie, nous développons des prédictions individuelles dans le cadre de la modélisation conjointe. Les modèles conjoints pour un événement récurrent et un événement terminal semblent être un cadre idéal pour prédire le risque de décès en tenant compte de l'information sur les rechutes. Ils quantifient la corrélation entre les deux processus, qui peut alors être incluse dans la prédiction. Il est ensuite direct d'obtenir les probabilités de décès conditionnelles aux rechutes, de manière dynamique. Ces prédictions ont été d'abord développées sur une série hospitalière de patientes avec un cancer du sein. Dans un second temps, durant l'étape de validation externe sur des données de registres anglais et néerlandais, les prédictions issues du modèle conjoint ont été comparées à celles issues d'un modèle de Cox en *landmarking*.

Dans une quatrième partie, nous étudions l'utilisation de ces prédictions individuelles dans le cadre des essais cliniques. L'objectif est de réduire le temps de l'essai. Les informations observées à un temps t, en particulier les rechutes de cancer dans les deux bras de traitement mais aussi les caractéristiques des patients, sont utilisées pour prédire les temps de décès non observés des patients à un temps plus éloigné $t + w$. L'effet du traitement peut ensuite être estimé au temps $t+w$ basé sur un mélange de temps de survie observés et prédits. Nous étudions et comparons trois méthodes pour imputer les temps de décès manquants. La meilleure méthode est ensuite appliquée rétrospectivement sur des données de deux essais cliniques randomisés étudiant l'effet d'une chimiothérapie adjuvante chez des patientes avec cancer du sein.

Une cinquième partie montre l'intégration des méthodes proposées sous forme de deux fonctions dans le package  *frailtypack*. Dans une dernière partie, la thèse se termine par une discussion générale et des éléments de conclusion.

Extension des mesures de concordance aux modèles à fragilité partagée

Les modèles à fragilité partagée gagnent de l'intérêt dans les études pronostiques, en particulier dû à l'utilisation de plus en plus fréquente d'études multicentriques. Il y a cependant eu peu de recherches sur l'extension des outils pronostiques aux modèles à fragilité, et notamment l'extension des mesures de discrimination. De la même façon que proposé précédemment pour le c-index de Harrell, nous avons étendu deux différentes mesures de discrimination. La première est une mesure basée sur le modèle de Cox, appelée estimation de la probabilité de concordance. Elle est donc soumise aux hypothèses du modèle, mais présente l'avantage de considérer tous les patients dans l'estimation, censurés ou non. La deuxième est une mesure non-paramétrique, version pondérée du c-index de Harrell, dont la pondération corrige l'influence de la censure. Seules les paires comparables sont utilisées, mais l'exclusion des paires incomparables, dont la fréquence dépend de la fréquence de la censure, est prise en compte par un poids égal à l'inverse de la probabilité de censure. L'extension de ces deux mesures prend en compte l'appartenance à un groupe : une mesure intra-groupe (seuls les patients appartenant au même groupe sont comparés), et une mesure inter-groupe (seuls les patients appartenant à des groupes différents sont comparés) sont proposées. Une mesure globale, moyenne des deux précédentes, est également proposée. Une étude de simulation a été réalisée pour évaluer l'impact du nombre de groupes, de la taille des groupes et du pourcentage de censure sur le biais des estimations. Les résultats montrent que les deux mesures étudiées peuvent être étendues au modèle à fragilité tout en restant indépendantes du taux de censure, tant que les groupes ont une taille suffisante. Les mesures ont été appliquées sur deux jeux de données : une étude multicentrique étudiant l'effet d'un boost de radiothérapie sur l'apparition de fibrose chez des patientes ayant un cancer du sein, et une méta-analyse étudiant l'effet d'une chimiothérapie en adjonction d'une chirurgie ou radiothérapie sur les cancers oto-rhino-laryngologiques.

Prédiction individuelle du risque de décès après rechutes du cancer : utilisation du modèle conjoint

Dans cette partie, nous avons un double objectif : le premier est d'évaluer l'apport des modèles conjoints pour prédire le risque de décès en tenant compte des caractéristiques du patient, mais aussi de la survenue de rechutes. Le deuxième est d'évaluer l'intérêt de

prendre en compte les rechutes lors de la prédiction du risque de décès.

Dans l'étape de développement de la prédiction, nous proposons trois prédictions issues d'un modèle conjoint pour un événement récurrent et un événement terminal. Ces trois prédictions sont une probabilité de décès entre le temps de prédiction t et l'horizon de prédiction $t+w$, conditionnelle au fait que le patient est en vie au temps t, conditionnelle aux caractéristiques du patient et de la tumeur, et conditionnelle ou non à l'histoire de la maladie. La première prédiction $P^1(t, t+w; \xi)$ est conditionnelle à l'histoire exacte de la maladie : nombre de rechutes et temps des rechutes. Elle correspond à la probabilité de décès lorsque l'ensemble de l'histoire du patient est connu. Un exemple est : « la patiente a été diagnostiquée il y a 3 ans. Elle vient de faire une deuxième rechute. Quelles sont maintenant ses chances de survie dans les 5 ou 10 ans ? » La deuxième prédiction $P^2(t, t+w; \xi)$ ne considère qu'une histoire partielle de la maladie. A la différence de la prédiction P^1, la prédiction P^2 considère que d'autres rechutes ont pu survenir entre la dernière rechute observée et le temps de prédiction. L'intérêt de cette prédiction est de pouvoir réaliser des projections. Par exemple : « la patiente a été diagnostiquée il y a 3 ans. Elle vient de faire une deuxième rechute. Si elle est encore en vie dans 5 ans, quelles seront alors ses chances de survie à 10 ans ? » Enfin, la troisième prédiction proposée dans le cadre du modèle conjoint $P^3(t, t+w; \xi)$ est une prédiction moyenne, qui ne tient pas compte de l'histoire personnelle de la maladie. Les informations sur les rechutes sont utilisées dans l'estimation du modèle uniquement, mais pas dans la prédiction. Elle répond à la question : « Dans la population de femmes ayant un cancer du sein et ayant telles caractéristiques, quelles sont les chances de survie à 5 ou 10 ans ? Si elles sont en vie 5 ans après le diagnostic, quelles seront alors les chances de survie à 10 ans ? »

Notre application a porté sur une série hospitalière de patientes ayant un cancer du sein. Sur ces patientes, nous avons développé les prédictions. L'erreur de prédiction apparente, c'est-à-dire calculée sur les patientes ayant servi au développement des prédictions, suggérait un intérêt de la prise en compte de la rechute dans la prédiction, ainsi que de bonnes performances du modèle conjoint. Une erreur obtenue par validation croisée en 10 fois sur notre série montre des résultats plus mitigés, avec peu de différence entre les prédictions étudiées. Sur les patientes ayant rechuté au moins une fois, en revanche, la prédiction P^1 issue du modèle conjoint avait une erreur légèrement plus faible. Ces résultats suggéraient la nécessité d'une validation sur données externes.

Dans une seconde étape, nous avons validé les performances sur des données externes indépendantes. La prédiction P^2 donnant des résultats très proches de la prédiction P^1, elle n'a pas été incluse dans cette étape de validation externe. Les prédictions issues de notre modèle conjoint ont cette fois été comparées à des prédictions issues d'un modèle

de Cox dans une approche *landmark*, où le nombre précédent de rechutes a été inclus comme variable explicative. Le modèle tel que développé sur les données hospitalières françaises a été appliqué sur des données de registres anglais et néerlandais. Sur les deux populations, l'erreur de prédiction était diminuée lorsque l'information sur la rechute était considérée. Les approches conjointe et *landmark* donnaient des résultats très similaires lorsque suffisamment d'information était recueillie. L'approche conjointe présente l'avantage de pouvoir obtenir de façon directe des prédictions dynamiques lorsqu'un seul modèle a été estimé, lorsque l'approche *landmark* nécessite d'estimer un modèle à chaque temps de prédiction. Les calibrations étaient bonnes sur les deux populations, malgré la différence de population (série hospitalière *versus* registres en population générale).

Utilisation des prédictions pour réduire le temps des essais cliniques

Une fois les phases de développement et validation réalisées, il est important d'évaluer l'utilité des prédictions en pratique. Elles peuvent être utilisées directement pour informer les cliniciens, mais aussi utilisées dans d'autres contextes, tels que les essais cliniques. Suivant l'idée des marqueurs de substitution en essais cliniques, nous voulons utiliser l'information sur les rechutes de cancer pour conclure plus tôt sur l'effet du traitement sur la survie globale. Pour cela, les prédictions du risque de décès tenant compte des rechutes développées précédemment sont utilisées pour imputer les temps de décès non observés. Nous avons comparé trois méthodes d'imputation : l'imputation par la moyenne résiduelle restreinte du temps de survie, l'imputation par échantillonnage dans l'estimation paramétrique de la distribution de survie, et par échantillonnage dans l'estimation non-paramétrique de la distribution de survie. Avec les trois méthodes, l'effet du traitement a été estimé par imputation multiple. Les simulations ont montré que l'imputation par la moyenne était biaisée alors que l'imputation utilisant l'estimation non-paramétrique de la survie était non efficiente dû à un écart-type empirique élevé. L'imputation utilisant l'estimation paramétrique de la survie donnait de bons résultats en terme d'estimation et d'écart-type. Si ces résultats n'étaient pas associés à une amélioration de la puissance, ils permettaient de préserver l'erreur de première espèce et un bon taux de couverture.

Discussion

Dans ce travail de thèse, nous avions pour objectif de répondre à certaines questions posées par l'évolution de la recherche en cancérologie. Un objectif en particulier était

l'évaluation de l'apport des modèles conjoints dans la prédiction de la survie des patients.

Nous avons montré que l'approche des modèles conjoints peut être ajoutée aux options disponibles pour étudier l'impact des rechutes du cancer sur le risque de décès. Trois principaux travaux ont précédemment étudié l'impact d'événements intermédiaires sur le risque de décès dans le cancer du sein. Hatteville et al. (2002) ont proposé de prédire le risque de décès à 20 ans après la chirurgie d'un cancer du sein, en deux étapes. La première étape consistait en l'estimation du risque d'événement, utilisant notamment des variables dépendantes du temps. La seconde étape était le calcul de probabilités conditionnelles de décès à 20 ans, sachant la survenue d'événements intermédiaires avant le temps de prédiction t ($t < 10$ ans). Le résultat était que la probabilité de décès chute de 89% si aucun événement intermédiaire n'a été observé à 9% si à la fois une rechute loco-régionale et une métastase étaient survenues. Une alternative a été l'utilisation de modèle multi-état par Putter et al. (2006) pour estimer les probabilités de transition entre les différents types d'événements. Cette méthode permet d'obtenir des probabilités de transition vers le décès qui diffèrent en fonction de l'événement précédent. Par rapport à la méthode précédente, le modèle multi-état évite l'utilisation de variables internes dépendantes du temps dans un modèle de Cox. Enfin, Parast et al. (2011) ont proposé une estimation non-paramétrique du risque d'un événement à long terme en considérant la survenue d'un événement à court terme. Cependant, dans leur application sur le cancer du sein, Parast and Cai (2013) ont choisi d'adapter l'approche *landmark* proposée par Van Houwelingen (2007) avec une estimation par un modèle de Cox. Dans cet article, elles dérivent les mesures de capacités prédictives appropriées. Notre travail étend les travaux précédents en étant approprié pour les événements récurrents ayant plusieurs occurrences. Il permet aussi une compréhension entière des deux processus d'événements et de leur inter-dépendance. Cette corrélation peut être directement intégrée dans le calcul de prédiction. Bien que ces méthodes aient des approches et des objectifs légèrement différents, une comparaison rigoureuse serait intéressante. Il est attendu que les résultats vont probablement dépendre de la nature des données, en particulier de la force de la dépendance entre les événements mais aussi l'évolution de l'effet des covariables dans le temps.

Contrairement à l'approche par modèle multi-état, notre travail ne fait jusqu'à présent pas la distinction entre rechutes loco-régionales et métastases. Pourtant, ces deux types d'événements devraient être considérés différemment. Notamment, le risque de base devrait être différent pour chaque type d'événements. Dans cet objectif, un modèle conjoint multivarié a récemment été proposé par Mazroui et al. (2013). Il serait intéressant de développer les prédictions à partir de ce modèle multivarié et évaluer si considérer

séparément les deux types d'événements conduit à des prédictions plus exactes. Une deuxième approche peut également être d'utiliser l'information sur les rechutes loco-régionales pour prédire le risque de métastases. En effet, la survenue de métastases peut changer l'évolution de la maladie. Enfin, il peut également être intéressant d'étudier si la survie des patients diffère en cas de survenue de multiples cancers successifs.

Cette thèse présente des développements nouveaux pour les modèles conjoints avec événements récurrents. Cependant, la plupart de ces développements avaient déjà été proposés pour les modèles conjoints avec données longitudinales et un temps de survie. En effet, beaucoup de biomarqueurs mesurés sont des variables continues, telles que le compte de CD4 pour l'étude du VIH/SIDA et l'antigène prostatique spécifique pour le cancer de la prostate. Cependant, le développement pour données récurrentes avec un événement terminal peut concerner beaucoup d'applications : l'étude des ré-hospitalisations, des crises épileptiques, des crises d'asthme et toutes les maladies chroniques définies par des épisodes. L'événement terminal est très souvent le décès. Une perspective possible de ce travail est de combiner une information longitudinale, la taille de la tumeur par exemple, avec les rechutes pour étudier le risque de décès. Cela peut être fait avec un modèle conjoint à trois parties.

Le champ de cette thèse est la recherche pronostique et l'évaluation de l'apport des modèles à fragilité dans ce domaine. Nous n'avons donc pas évalué ni discuté le modèle en lui-même et son estimation. Cela a été étudié précédemment par Rondeau et al. (2007). Seules quelques simulations ont été réalisées, dans lesquelles l'estimation des paramètres était correcte.

Le modèle conjoint peut jouer un rôle non seulement dans l'utilisation de marqueurs de substitution, mais aussi dans leur validation. En effet, l'approche méta-analytique proposée par Buyse et al. (2000) est actuellement réalisée en deux étapes : une corrélation au niveau individuel basée sur des copules, et une corrélation au niveau essais estimée par régression. Ces deux étapes pourraient être combinées en une en utilisant un modèle conjoint avec un effet aléatoire au niveau individuel et un effet du traitement aléatoire au niveau essais. Une telle méthode pourrait conduire à des estimations plus précises des corrélations entre le critère de substitution et le vrai critère.

Finalement, une question reste en suspens : comment peut-on évaluer la discrimination d'une prédiction dans le cadre de données récurrentes ? Comme nous l'avons vu, le c-index n'est pas approprié. Nous avons donc choisi d'utiliser l'erreur de prédiction pour évaluer si nos prédictions sont proches de la réalité mais ce concept diffère quelque peu du concept de discrimination. En effet, l'erreur de prédiction estimée par le Brier score recouvre à la fois les concepts de discrimination et de calibration. Une erreur de

prédiction élevée peut donc être due à une faible discrimination, c'est-à-dire les covariables ne sont pas assez prédictives pour séparer les patients qui vont avoir des temps de survie différents, mais peut également être due à une mauvaise calibration, qui pourrait par exemple être expliquée par une mauvaise estimation du risque de base. Comme ces deux concepts n'ont pas les mêmes explications en cas de faibles résultats, ni les mêmes solutions, il serait intéressant de dissocier les deux.

Notations and abbreviations

Notations

Please, note that these notations were used throughout the thesis, but may differ a little in the published papers.

- $I[cond]$ is the indicator function, equals to 1 if *cond* is true, 0 otherwise
- $E[x]$ is the mean of the random variable x and $var[x]$ its variance
- v' denotes the transpose of the vector v
  $f''(.)$ denotes the second derivative of the function $f(.)$
- t^- denotes the time just before t
- $P(.)$ or $P[.]$ denotes the probability of an event
- To distinguish clustered data from recurrent event data, the following subscripts are used
  - $g = 1, \ldots, G$ denotes the group of individuals/patients
  - $i = 1, \ldots, N$ denotes the individual/patient
  - $j = 1, \ldots, n_i$ denotes the recurrent event index in one individual/patient
- T is a survival time and T_i is the survival of the individual i

  C_i is the censoring time of patient i

  $\tilde{T}_i$ is the observed survival time, $\tilde{T}_i = min(T_i, C_i)$
  T_{ij} is the observed time of occurrence of the event j of the individual i
  $\delta_i = I[\tilde{T}_i = T_i]$ and δ_{ij} are the associated event indicators

- In joint model

  T_{ij}^R is the time of the recurrence j of individual i

  T_i^D is the time of death of individual i

  δ_{ij}^R and δ_i^D are the associated event indicators

- $S(.)$ is the survival function related to the event of interest

  $\lambda(.)$ is the instantaneous hazard function

  $\Lambda(.)$ is the cumulative hazard function

  $f(.)$ et $F(.)$ are the density function and the cumulative distribution function, respectively

- $\lambda_0(.)$ is the baseline instantaneous hazard function

- In joint model

  λ_{ij}^R is the instantaneous hazard function for the recurrence j of individual i

  λ_i^D is the instantaneous death hazard function for the individual i

  $\lambda_0^R(.)$ is the baseline instantaneous recurrence hazard function

  $\lambda_0^D(.)$ is the baseline instantaneous death hazard function

- $g(.)$ and $G(.)$ are the density function and the cumulative distribution of the censoring time C

- Z_i is the covariate vector of the individual i

  Z_{ij} is the covariate vector associated to the recurrence j of the individual i

  Z_{ij}^R is the covariate vector associated to the risk of the recurrence j of the individual i

  Z_i^D is the covariate vector associated to the risk of death of the individual i

- β is the vector of covariate effects

- d_k is the number of events associated to time T_k

  m_i is the number of events associated to the cluster i

- $N_i(.)$ is the counting process of individual i

  $N_i^R(.)$ is the counting process of the recurrent event

Y_i is the at-risk process of individual i

$\alpha(t)$ is the intensity of the counting process

- u_i is the frailty of the individual i

  u_g the frailty of the group g

  $g(u)$ represents the density of the frailty

  θ is the variance of the frailty

- α is the flexibility parameter in the joint model

- $\xi = (\lambda_0(.), \beta, \theta)$ in shared frailty models and $\xi = (\lambda_{ij}^R(.), \lambda_i^D(.), \beta_1, \beta_2, \theta, \alpha)$ in joint models is the vector of model parameters

- $L(.)$ is the likelihood function of a variable or a model

  $LL(.)$ the log-likelihood

  $pLL(.)$ the penalised log-likelihood

- $H(.)$ is the Hessian matrix of the likelihood function

  $I(.)$ is the Information matrix of the likelihood function

- κ and κ_1, κ_2 are the smoothing parameters in the penalized likelihood

- CV_a is the approximated cross-validation score

Abbrevations

AIDS: Acquired ImmunoDeficiency Syndrome

BS: Brier score

CD4: Cluster of Differentiation 4

CRAN: Comprehensive R Archive Network

EM algorithm: Expectation-Maximisation algorithm

EORTC: European organisation for Research and Cancer Treatment

HER2: Human Epidermal Growth Factor Receptor 2

HIV: Human Immunodeficiency Virus

IPCW: Inverse Probability of Censoring Weighting

LCV: Likelihood Cross-Validation criterion

MACH-NC: Meta-Analysis of Chemotherapy in Head and Neck Cancer

SEER: Surveillance, Epidemiology, and End-Results

PFS: Progression-Free Survival

RCT: Randomised Clinical Trial

OS: Overall Survival

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Challenges in cancer research

The word *cancer* actually covers a lot of diseases, according to the localisation or the histology. Tumours are due to a uncontrolled division of the cells. Some environmental risks factors of cancer are now well-known, such as tobacco, which is the best identified human carcinogen (Boyle et al., 2008). Others are under investigation, such as pesticides or mobile phones. Finally, the genetic aetiology of cancers is also being studied. Awaiting that all cancers can be explained and prevented, there were 14 millions people diagnosed with cancer in 2012 worldwide, and 8 million dying of it (Ferlay et al., 2013). This represents an increase of around 11% in prevalence and 8% in mortality since 2008. The increasing number of cases can be explained by the population growth, the ageing of the population, and finally an increasing incidence of cancer. Using projections, the number of new cases could be between 20 and 26 millions in 2030 (Boyle et al., 2008). Among all cancers, the breast cancer is the most common in women and represented 1.7 million new cases in 2012. Since 2008, its incidence increased of more than 20 % and the mortality of 14%. In this context, an important part of the cancer research is about treating patients.

Treatment of cancer includes several options, such as surgery, radiotherapy, chemotherapy, and most recently targeted therapy. Treatments can be given in combination, and among the bunch of available options, clinicians make their a primarily based on treatments proven efficacy. However, some of those meaning to be very aggressive to the tumour are also very aggressive to the normal cells, leading to important adverse events. This is the case of chemotherapy, but also radiotherapy. If other, less aggressive, options are available, they may be preferred by clinicians as a first line of treatment for patients with good prognosis. Furthermore, treatments like targeted therapies aim at treating

specific types of tumour and they are thus not efficient for all patients. Therefore, the treatment choice in cancer must be done according to the patients' and tumours' characteristics. Research on cancer treatment has been evolving for last years in this one main direction: personalised medicine. The ultimate goal is to precisely identify every cancer types, in order to give the patient the treatment that will specifically target his tumour. The trend is thus to characterise sub-populations and cancer subtypes, to find new biomarkers that make it possible either to identify patients with the best survival chances, or to define different tumour types responding to different treatments. For example, in breast cancer, two subtype classifications were recently defined (Badve et al., 2011). The fist one identified a group of triple-negative tumour (neither oestrogen receptor, progesterone receptor nor HER2 expression). The second one consists of five molecular subtypes: luminal A, luminal B, normal breast-like, HER2, and basal-like. Both are based on the tumour features and the biomarkers it expresses. Otherwise, the classical factors shown to be related to clinical outcome of breast cancer patients are tumour size, histological type, histological grade, vascular invasion status and lymph node involvement status (IARC, 2002, chapter 1). These factors can be used to 1- decide the most appropriate treatment, 2-monitor breast cancer screening and 3- monitor the changing patterns of disease incidence. Other known prognostic factors include the presence of metastases, growth fraction, hormone and growth factor receptor status, and, more generally, other molecular characteristics.

Another specificity of cancer: it is a progressive disease. After the diagnosis and the treatment of the primary tumour, local relapses or distant metastases can appear, sometimes after a short time, sometimes after years. The occurrence of such disease events shows a reappearance or evolution of the tumour. A patient can undergo several events after the first line of treatment. In general, each relapse is followed by a new line of treatment, chosen according to the patient characteristics, but also his history regarding the disease. Ultimately, the main question of interest when treating a patient is: how long the patient will survive? What are his chance to be alive in five, ten or twenty years? How long will be his survival increased by using this treatment as compared to that one, as compared to none? Increasing the overall survival of patients remains the primary objective. However, it may be related to increasing the time without disease events. This is the reason why research is now often focusing on those intermediate events.

Such evolution of the clinical research implies a parallel evolution of the biostatistic methods to answer the new needs. The first challenge raised by cancer clinical research is to propose the best prognostic model to clinicians. This means to be able to propose an adequate model, studying the endpoint of interest the right way, including all the

information of interest to obtain a useful, accurate but also usable in practice prognostic tools. This also means that we are able to evaluate the tools that we proposed, to compare them fairly, to eventually select the best one. Prognostic research has been part of the biostatistic field for years now. However a special emphasis has been given in the last five years with the publication of three series of paper trying to define recommendations out of the amount of work done until then: *Prognosis and prognostic research* published in BMJ in 2009 (Moons et al., 2009; Royston et al., 2009; Altman et al., 2009; Moons et al., 2009); *Risk prediction models* published in Heart in 2012 (Moons et al., 2012, 2012); and *PROGRESS* published in BMJ and PLoS Medicine in 2013 (Hemingway et al., 2013; Riley et al., 2013; Steyerberg et al., 2013; Hingorani et al., 2013). Basically, the main steps required to obtain an accurate prediction model are as follows: 1) identification of potentially pertinent predictors, 2) choice of an adequate modelling method, 3) selection of predictors and estimation of their effects (model development), 4) application of the model on new population and assessment of the prediction ability (model validation) and 5) model revision. The validation step 4 is crucial to assess the possible use of the model in practice, and consider to use it in various populations (Altman and Royston, 2000; Konig et al., 2007). In clinical and epidemiological research, the accuracy is looked for, leading to big studies. The sources of information are combined, leading to clustered data. Clusters have to be accounted for when developing prediction model. The development methods exist for this kind of data, but there is a lack of appropriate validation methods.

The second question raised is how to do prediction considering not only the characteristics of the patients at a given time, but his whole history. This is of particular interest for cancer patients who may undergo disease events. It is important to take these events into account as they are correlated to the overall survival of the patient. It is also important that the patient is offered the most appropriate treatment each time he needs it, including after each relapse. For that, the information about his relapses are of importance. To take into account intermediate events in prognosis, landmarking is a useful and simple method (Van Houwelingen, 2007; Van Houwelingen and Putter, 2011). However, history of patient has to be simplified to be included in a model such as Cox model. In previous works, joint model for longitudinal data and a survival time were used to predict a survival time accounting for the evolution of a longitudinal biomarker (Faucett et al., 2002; Proust-Lima and Taylor, 2009; Rizopoulos, 2011; McLain et al., 2012; Taylor et al., 2013). To our knowledge, no similar work had been published using joint model for a recurrent event and a terminal event.

The third challenge is the randomized clinical trial (RCT) duration. Indeed, to define the most efficient treatment, RCTs have to be performed. However, using the

overall survival (OS) as primary endpoint, these trials can take years before reaching a conclusion. For some other diseases, the main endpoint may be too expensive or too complex to measure. As an answer to these problems, use of surrogate endpoints has been developed. They are endpoints that 1- are observable at shorter time that the main endpoint of interest, 2- are correlated to it and 3- on which the observed treatment effect may give some information about the treatment effect on the main endpoint. For example, the progression-free survival (PFS) has been shown to be a surrogate endpoint for OS in several cancers. Indeed, disease progressions are observable at a shorter time than death, they are correlated to it, and the efficacy of a treatment observed on the PFS may translate on the OS. PFS is therefore an excellent candidate surrogate for OS. The research has been mainly focusing on how to validate such criteria. It started with Prentice (1989), who stated a list of four correlation criteria to be fulfilled for an endpoint to be a valid surrogate endpoint, further investigated by Freedman et al. (1992) and Buyse and Molenberghs (1998). Ten years later, Buyse et al. (2000) proposed a meta-analytic approach to surrogate validation. This reduces to two correlations: correlation between the two endpoints at the individual level, and correlation between the two treatment effects at the trial level. Once a surrogate is validated, it should be integrated in practices. Considering the use of the surrogate endpoint in practice, two leads have been followed. The first one is to use the surrogate endpoint as the primary endpoint of trial, to estimate the treatment effect on it, and then to predict the treatment effect on the OS. This is the concept of *surrogate threshold effect*, proposed by Burzykowski and Buyse (2006) in the meta-analytic validation framework. The second method was proposed by Faucett et al. (2002) in AIDS research. The OS stays the primary endpoint of the trial. The information about an auxiliary variable – for Faucett it was the CD4 level, for us it is the disease events – is used to impute the not yet observed death times. Then, the treatment effect is estimated on OS, using both observed and imputed death times. This latter approach used joint modelling for longitudinal event and a survival time. A similar approach has been recently proposed to account for intermediate disease events using cure models (Conlon et al., 2011).

1.2 Thesis objective and outline

This thesis takes place in the personalised medicine context in cancer research. As seen in the previous section, several challenges were raised by the willing to get the best treatment for each patient. Our first objective was to propose a measure of discrimination when prediction was developed on clustered data. The second objective was to derive

some predictions from joint model model for a recurrent event and a terminal event and to assess its performances. This modelling framework seemed ideal to study the death accounting for intermediate disease events and other characteristics. As a third objective, we wanted to illustrate the use and usefulness of such prediction in RCT context.

The second part of this thesis explains the main existing methods that are used in survival and prognosis research. It contains details on classical methods as well as more recent ones, that were used in our methodological developments.

The third part is about the adaptation of the concordance measures to shared frailty survival models. Concordance is a discrimination measure, widely used, which assess whether the prediction is in coherence with the observation, i.e. whether the predicted risk is indeed higher for the patients who will die sooner. Different estimations of this quantity were previously proposed. Among them, the most commonly used is the c-index, a non-parametric measure based on the count of the pairs of patients said *concordant*. Two versions of it have been proposed: a crude one (Harrell et al., 1982, 1996), that has shown to be dependent of censoring, and a weighted one (Uno et al., 2011). A second proposed estimation, called concordance probability estimation, used the properties of the proportional hazards Cox model to estimate the probability of concordance (Gönen and Heller, 2005). Both measures were initially developed for non correlated survival times. Only the crude c-index has been previously adapted to correlated survival times (Van Oirbeek and Lesaffre, 2010). We have proposed an extension of the two other measures, to have both independence of the amount of censoring and correlated survival times studied through shared frailty models. This extension accounted for the two data levels, within-group and between-group.

In the fourth part, we have developed individual prediction in the framework of joint modelling. Joint models for a recurrent event and a terminal event seemed to be an ideal framework to predict the risk of death while accounting for the information on disease events. They quantify the correlation between the two processes, which can be included in the predictions. It is then direct to obtain some probabilities of death conditional on disease events, in a dynamic way. Those predictions were first developed on a hospital series of patients with breast cancer. In a second time, during an external validation step using Dutch and English registry datasets, the predictions from the joint model were compared to the ones from a landmark Cox model.

In the fifth part, we have studied the use of these individual predictions in randomized clinical trials. The objective was to reduce the trial duration. The information observed at a time t, especially disease events in each treatment group but also patient characteristics, were used to predict the not yet observed death times of patients at a

later time point. Based on a mixture of the observed and predicted events, the treatment effect was estimated. We studied and compared three different methods to impute the missing death times. The best method was then retrospectively applied on data from two randomized clinical trials studying the effect of an adjuvant chemotherapy in breast cancer patients.

A sixth part illustrates the two functions added in the  package *frailtypack* further to our developments. Finally, the thesis ends with a general discussion and concluding remarks.

Chapter 2

Theoretical background

2.1 From survival data to joint models

Patients' survival is one of the major clinical endpoints in cancer research, whether it be time to death or time to another event. In this thesis, we mainly focus on overall survival, defined as the time elapsed between the diagnosis and the death whatever the cause. There is a large literature on the basic concepts of survival analysis. First parts of this section are mainly based on books from Aalen et al. (2008) and Martinussen and Scheike (2006).

2.1.1 Survival time, censored data and Kaplan-Meier estimate

The survival time T is a continuous random variable. It represents the time elapsed between an origin and the occurrence of the event of interest. The specificity of this variable is that the event of interest, and thus the exact survival time, is not always observed. In that case, all we know is that it is greater than the observation time. The resulting right-censored data has to be accounted for in specific survival analyses. In addition to its density $f(t)$, the survival time distribution can be characterised by two concepts: the *survival function* and the *hazard function*.

The *survival function* $S(t)$ is the probability that the survival time exceed a time t: $S(t) = P(T > t)$. It equals one at the origin time (often $t = 0$), and tends to zero when t tends to infinity, except cases in which individuals can be cured. It is defined as

$$S(t) = 1 - \int_0^t f(u)du$$

The *hazard function* $\lambda(t)$ is the instantaneous event rate at time t. It represents the

rate of event at time t, given that the event did not occur before t. It is defined by

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

From it, is obtained the cumulative hazard function $\Lambda(t)$, defined by

$$\Lambda(t) = \int_0^t \lambda(u) \mathrm{d}u$$

Hazard and survival functions are related as follows: $\lambda(t) = \frac{-d \ln(S(t))}{dt} = \frac{f(t)}{S(t)}$ and $S(t) = e^{-\Lambda(t)}$.

Likelihood for right censored data We denote T_i the survival time of individual i, and C_i his censoring time. To each individual i is associated the couple $(\tilde{T}_i, \delta_i)$ where $\tilde{T}_i = min(T_i, C_i)$ is the maximum observation time for i and $\delta_i = I[\tilde{T}_i = T_i]$ the indicator of event. Let $g(t)$ denote the density function and $G(t)$ denote the cumulative distribution function of censoring times. The likelihood function for the survival time T is written as the product of the individual contributions of censoring and event times:

$$L(T) = \prod_{i=1}^{N} [(1 - G(\tilde{T}_i)) f(\tilde{T}_i)]^{\delta_i} [(1 - F(\tilde{T}_i)) g(\tilde{T}_i)]^{1-\delta_i}$$

Once we assume the independence of censoring and survival parameters, we have $[1 - G(\tilde{T}_i)]^{\delta_i}$ and $g(\tilde{T}_i)^{1-\delta_i}$ non informative for the likelihood. Thus we obtain:

$$L(T) = \prod_{i=1}^{N} \lambda(\tilde{T}_i)^{\delta_i} S(\tilde{T}_i)$$

Non parametric estimator The survival function can be estimated non parametrically by the Kaplan-Meier estimate. Let T_k be the ordered event times, and $Y(T_k)$ the number of individuals at risk at time T_k, i.e. not censored and not having undergone the event yet. The Kaplan-Meier estimate is written as:

$$\hat{S}(t) = \prod_{T_k \leq t} \left\{ 1 - \frac{d_k}{Y(T_k)} \right\}$$

where d_k is the number of events occurring at time T_k. The initial estimator had 1 instead of d_k but it was extended to account for ties. The corresponding variance estimator is:

$$\hat{\tau}^2 = \hat{S}(t)^2 \sum_{T_k \leq t} \frac{d_k}{Y(T_k)[Y(T_k) - d_k]}$$

2.1.2 The semi-parametric Cox model

To study the effect of covariates on survival time, different regression models can be used. Those models are not directly interested in the effect on the survival function but on the hazard function. We focus here on the proportional hazards models, which assume a constant effect of the covariates at every time t, and which particularly interest us for the following parts. Other models exist for non proportional hazards, such as the Aalen additive model (Aalen, 1980).

The most commonly used model for survival analyses is the Cox model (Cox, 1972). Let β denotes the effects of the covariates Z_i of individual i. The Cox model defines the hazard function of individual i by

$$\lambda(t|Z_i) = \lambda_0(t)\exp(\beta' Z_i)$$

where $\lambda_0(t)$ is the baseline instantaneous hazard of event, common to all patients, describing the shape of the hazard function. The hazard ratio $\exp(\beta' Z_i)$ is the over-risk of event specific to each individual depending on his covariates.

As initially proposed, $\lambda_0(t)$ is left unknown and the model parameters are estimated using a partial likelihood. Let T_k be the ordered event times, and $R(T_k)$ the set of individuals at risk at time T_k. Then the partial likelihood is written by:

$$L(\beta) = \prod_{T_k} \frac{e^{\beta' Z_k}}{\sum\limits_{l \in R(T_k)} e^{\beta' Z_l}}$$

2.1.3 Frailty models: introduction to counting processes and recurrent event analysis

Counting processes

Theoretical basis of the Cox model, and thus its extensions, can be found in the counting process theory. Using the counting process formulation, each individual i is characterized by a pair $(N_i(t), Y_i(t))$, where $N_i(t)$ is a counting process, describing the number of observed events in the interval $[0, t]$ for individual i, and $Y_i(t)$ is the at-risk process, that indicates if i is still under observation and at risk of an event at time t. Right-censored survival data is a special case, when we consider $N_i(t) = I[T_i \leq t, \delta_i = 1]$ and $Y_i(t) = I[T_i > t]$. Counting process theory generalizes easily to multiple recurrent events, by still holding the at-risk indicator $Y_i(t) = 1$ after the occurrence of an event, as long as the individual is under observation. The counting process jumps of one unit at each event time, and is constant in-between. The intensity of the process is defined by

$\alpha(t)\mathrm{d}t = P(\mathrm{d}N(t) = 1|past)$, with $\mathrm{d}N(t)$, the number of events on the period $[t, t + \mathrm{d}t)$, being binary. The intensity of the counting process is related to the hazard function of the survival time through $\alpha(t) = Y(t)\lambda(t)$.

Analysis of recurrent events

Analysis of recurrent events may be of interest in many applications, e.g. to explain repeated hospitalisation sojourns or epileptic seizures. In cancer, it is mainly the disease events, such as loco-regional relapses and distant metastases, that are studied as recurrent events. The specificity of these data is the non-independence between the observations. Indeed, we can assume that there is a dependence between different events of one individual. This dependence leads to heterogeneity between the individuals. To study these data, marginal approaches aim at a robust estimation of the parameters treating the heterogeneity as a nuisance, while conditional ones aim at measuring this heterogeneity.

When studying recurrent events, different time-scales can be used. The j^{th} occurrence of the event can be studied either using the time since the study start, or using time since the $j - 1^{st}$ occurrence (Figure 2.1). In the last case, called *gap time* and for which an illustration is the renewal process approach, the hazard of the event j is modified by the occurrence of the $j - 1^{st}$ event, as the time and thus the hazard function are reset to the origin. In the first case, called *calendar time* and for which an illustration is the Poisson process, the instantaneous hazard of the occurrence j is not impacted by the occurrence of previous events but only by the observation time. However, individuals can not be at risk of the j^{st} occurrence as long as they did not undergo the $j - 1^{st}$ occurrence.

Different marginal approaches have been proposed to study ordered multiple events, as described in Therneau and Grambsch (2000). First, the *Andersen-Gill approach (AG)* is very closed to the Cox model. The intensity of the process is defined by $\lambda(t|Z_i) = Y_i(t)\lambda_0(t)\exp(\beta' Z_i)$. The difference with the Cox model is that the at-risk process $Y_i(t)$ remains one when events occur as long as the patient is being observed. In this approach, the observations within a subject are assumed independent given the covariates. Second, the *marginal approach from Wei-Lin-Weissfeld (WLW)* is a stratified model. Each stratum correspond to an event rank j. The intensity for the j^{th} event of the i^{th} subject is $\lambda_{ij}(t|Z_i) = Y_{ij}(t)\lambda_{0j}(t)\exp(\beta_j' Z_i)$. In this approach, both the baseline hazard and the covariate effects are allowed to differ from one stratum to another. The at-risk process is also specific for each rank of event, and $Y_{ij}(t)$ equals one until the occurrence of the j^{th} event (or censoring). The third approach is the *model from*

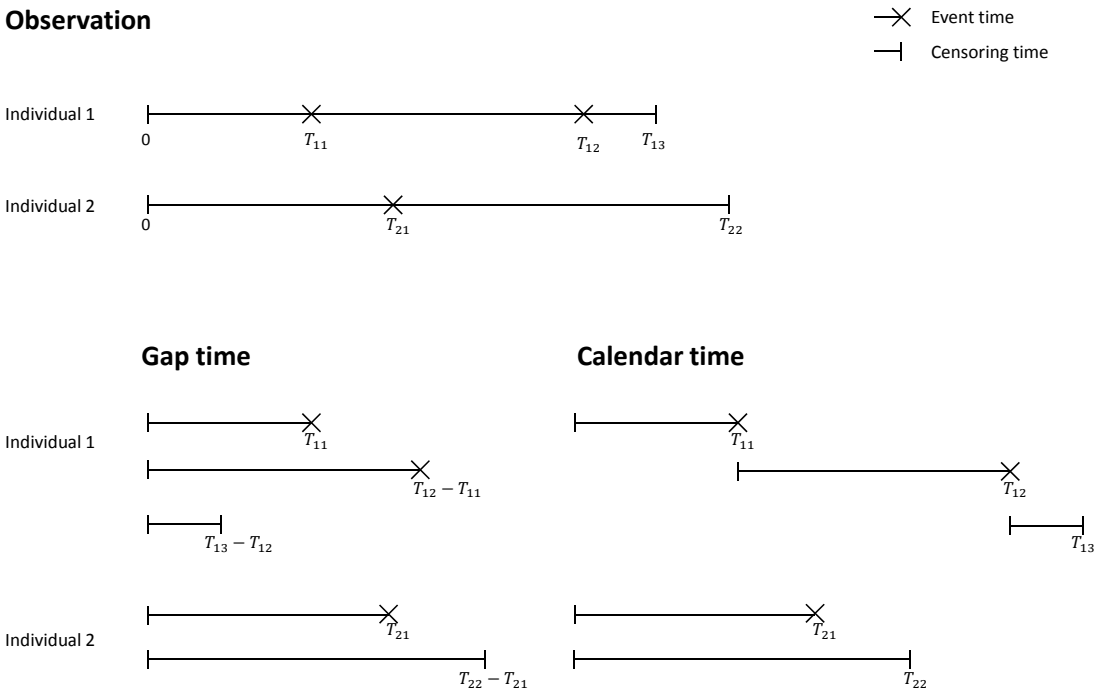

Figure 2.1: Illustration of two time scales to study recurrent events. The censoring can be either the end of the follow-up, or a loss to follow-up, or the occurrence of a concurrent event.

Prentice-Williams-Peterson (PWP). As in the WLW approach, the model is stratified on the rank. However, time-dependent strata are used, so that a subject is not at risk for the j^{th} event until the $j-1^{st}$ occurrence. The intensity is thus the same as in WLW, except that the at-risk process $Y_{ij}(t)$ is zero until the time of the j^{th} occurrence. Therneau and Grambsch (2000) preferred the use of the AG model, as the WLW gives biased estimations and the PWP is similar to AG when all the important covariates are included, but AG is best when an important covariate is omitted.

The shared frailty model

In this thesis, we are more interested in the use of conditional models, in particular *shared frailty models*. Shared frailty models are random effect model for survival data. The random effect, called *frailty*, describes the excess risk of a cluster. Typically, a cluster may be a patient when studying recurrent event, or a group of patients, such as

an hospital, a family or a clinical trial. The idea is that different clusters are associated to different frailties, due to some cluster-specific unobserved factors. This results in an heterogeneity between clusters. The clusters with the highest frailties will die earlier. The hazard function for the j^{th} recurrent event of the individual i is defined by:
$\lambda_{ij}(t|Z_{ij}, u_i) = u_i \lambda_0(t) \exp(\beta' Z_{ij})$
and for the i^{th} individual of the group g by:
$\lambda_{gi}(t|Z_{gi}, u_g) = u_g \lambda_0(t) \exp(\beta' Z_{gi})$

Cumulative hazard function conditional on covariates Z and conditional survival function both can be derived from the conditional hazard function following the same relation as in section 2.1.1 on page 8.

Methods of estimation proposed for these models were mainly Expectation-Maximisation (EM) algorithm. The EM algorithm consists in alternating between the two steps after parameter initialisation: the M-step treats the frailty term as fixed, and maximise the likelihood to estimate the fixed parameters (here, the fixed parameters include the variance of the frailty terms) as for a classical regression model; the E-step computes the expected values of the frailty terms given the current values of the fixed parameters (see Klein (1992) for example). A residual maximum likelihood method was described in Lam and Ip (2003) while Therneau and Grambsch (2000) proposed a penalised likelihood with the penalty being a function of the frailty term. The use of a some Bayesian approaches were also proposed (Clayton, 1991; Ducrocq and Casella, 1996). Ha and Lee (2005) compared two other approaches, the hierarchical likelihood (also called h-likelihood) and an approach using orthodox best linear unbiased predictor. The baseline hazard can be left unspecified, as in the Cox model, or estimated using a Weibull distribution or other parametric estimations. Rondeau et al. (2003) proposed to approximate the baseline hazard function using splines (see details about splines in section 2.5 of the present chapter). This led to the use of a full likelihood (as opposed to the partial likelihood of the Cox model) and to the addition in the likelihood of a penalization that is function of the baseline hazard. Box-Steffensmeier and De Boef (2006) compared different models to study recurrent events, studying both the heterogeneity in patients and the event dependence. They argue that in presence of heterogeneity, variance-corrected models, meaning non-random effect models such as the AG model presented earlier, only give a robust estimate of the variance but do not incorporate heterogeneity into the estimates of effects themselves, and thus remain biased. On the contrary, the frailty models better deal with this heterogeneity. In the presence of event dependence, however, they recommend a conditional frailty model.

Frailty distribution

The distribution of the random effect can be chosen among several distributions. The most common ones are the gamma distribution and the log-normal distribution. Other possibilities include positive stable, inverse Gaussian and more generally the family of power variance function. The choice of the frailty distribution has been extensively discussed by Hougaard (see Hougaard (1995) for an example). In practice, it has been shown that the fixed-effect parameter estimation is robust against the misspecification of the frailty distribution (see Pickles and Crouchley (1995) and Munda and Legrand (2014); see also Mazroui et al. (2012) for the joint frailty model). The gamma distribution is often preferred as its mathematical properties lead to close form of the likelihood.

The gamma-frailty model with spline approximation of the baseline hazard

In this thesis, we use the gamma-frailty model with baseline hazards estimated by splines as proposed in Rondeau et al. (2003). We use

$$u_i \sim \text{Gamma}(\frac{1}{\theta};\frac{1}{\theta}) \quad \text{and} \quad g(u_i) = \frac{u_i^{1/\theta-1}\exp(-u_i/\theta)}{\theta^{1/\theta}\Gamma(1/\theta)} \tag{2.1}$$

giving $E[u_i] = 1$ and $var[u_i] = \theta$. The $E[u_i] = 1$ is imposed to ensure the model to be identifiable (Andersen et al., 1997, chapter 9). Let $\xi = (\lambda_0(.), \beta, \theta)$ be the vector of model parameters. The marginal log-likelihood of the model is written as follows (given for a recurrent event, but indices can be changed accordingly for grouped data).

$$\begin{aligned}
LL(\xi) &= \sum_{i=1}^{N} \ln\left(\int_0^\infty L_i(\xi,u)g(u)\text{du}\right) \qquad (2.2)\\
&= \sum_{i=1}^{N} \ln\left(\int_0^\infty \prod_{j=1}^{n_i}\left[(\lambda_0(T_{ij})ue^{\beta' Z_{ij}})^{\delta_{ij}} e^{-\Lambda_0(T_{ij})ue^{\beta' Z_{ij}}}\right]\frac{u^{\frac{1}{\theta}-1}e^{-\frac{u}{\theta}}}{\theta^{\frac{1}{\theta}}\Gamma(\frac{1}{\theta})}\text{du}\right)\\
&= \sum_{i=1}^{N}\left\{\sum_{j=1}^{n_i}\delta_{ij}\{\beta' Z_{ij} + \ln(\lambda_0(T_{ij}))\}\right.\\
&\quad - (1/\theta + m_i)\ln\left[1+\theta\sum_{j=1}^{n_i}\Lambda_0(T_{ij})e^{\beta' Z_{ij}}\right]\\
&\quad \left. + I_{\{m_i\neq 0\}}\sum_{k=1}^{m_i}[\ln(1+\theta(m_i-k))]\right\}
\end{aligned}$$

where m_i is the number of events for the patient i.

If a parametric hazard function is preferred, the equation (2.2) can be directly maximized to obtain the estimator of the model parameters $\hat{\xi}$. If splines are used to estimate the baseline hazard function, a penalised likelihood is proposed by Rondeau et al. (2003), to ensure the smoothness of the estimation of the hazard function. The penalised likelihood is defined by:

$$pLL(\xi) = LL(\xi) - \kappa \int_0^\infty \lambda_0^{''2}(t)\mathrm{d}t \tag{2.3}$$

where $LL(\xi)$ is the full likelihood as defined in (2.2). The smoothing parameter κ allows to balance between being close to the data (low values) and the smoothness of the baseline hazard (high values). The estimates $\hat{\xi}$ are obtained by maximising $pLL(\xi)$. The penalized log-likelihood is maximised using a modified Marquardt algorithm, briefly presented in section 2.5. The variance-covariance matrix estimate is obtained by a direct or a sandwich estimator based on the inversion of the negative Hessian matrix of the penalised likelihood.

The choice of the smoothing parameter may be done either by assessing the smoothness of the baseline hazard graphically, or by cross-validation. The use of an approximate cross-validation score (CV_a) was proposed in Rondeau et al. (2003). The κ that maximise this CV_a will be chosen. The idea of such criterion is to maximize the sum of the N model likelihoods obtained when each individual is left-out once. In practice, an approximation is used to decrease the calculation time. The criterion has the form:

$$\widehat{CV_a}(\kappa) = \frac{1}{N} L_i(\hat{\xi}_s) - \frac{1}{N} trace([-\hat{H}(\hat{\xi}_s)]^{-1}\hat{I}(\hat{\xi}_s)) \tag{2.4}$$

where ξ_s are the parameters relating to the spline estimation of the baseline hazard, $L_i(\xi_s)$ is the individual contribution to the likelihood of the model, $H(\xi_s)$ is the Hessian matrix of the penalised likelihood and $I(\xi_s)$ is the information matrix of the penalised likelihood. When $trace([-\hat{H}(\hat{\xi}_s)]^{-1}\hat{I}(\hat{\xi}_s))$ is interpreted as an effective number of parameters, the CV_a criterion is equivalent to the Akaike information criterion.

2.1.4 Time-dependent covariates

Time-dependent covariates are repeated measures of a covariate that evolves with time. They can be of great interest to study a survival endpoint. For example, relapses are time-dependent covariates when studying their impact on the risk of death. However, a careful modelling is required. First, it is important to distinguish between exogenous and

endogenous covariates. Kalbfleisch and Prentice (2002, chapter 6.3) give us the following definitions:

- **External –or exogenous– covariate**: the future path of the covariate $Z_i(t), t > s$ is *not* affected by the occurrence of an event at time s. This condition can be defined by $P[Z_i(t)|Z_i(s), T \geq s] = P[Z_i(t)|Z_i(s), T = s], \forall s, t : 0 < s \leq t$. This condition is equivalent to $P[T_i \in [s, s + \mathrm{d}s]|Z_i(s), T \geq s] = P[T_i \in [s, s + \mathrm{d}s]|Z_i(t), T \geq s]$, meaning that the hazard at time s depends only on the covariate path up to time s^-.
  This definition includes all variables that vary in a predetermined way, like the age, and all the variables that are external to the individual, e.g. pollution level. It also includes all the fixed covariates, i.e. $Z_i(t) = Z_i, \forall t$.

- **Internal –or endogenous– covariate**: the future path of the covariate is affected by the occurrence of an event, thus the condition defined above does not hold. These covariates result from a stochastic process that is generated by the patient under study himself. Typically, it is a biomarker that is measured on the patient, and thus requires the patient to be alive. As a consequence, the measurement of $Z_i(s)$ requires $T > s$ and carries direct information on the survival time. For example, relapse at time s can be assessed at the condition that the patient is alive. It can be any value of a biological measure, such as the prostate-specific antigen or disease complications.

When studying a survival endpoint, the external covariates can be easily studied using the counting-process form of a Cox model, due to the condition defined above. The internal covariates however should be differently accounted for. Indeed, the hazard at time s for an external covariate depends on the covariate process until time s, but not further. This is not the case for the internal covariates. As a consequence, the associated hazard defined above is not related directly to a survival function and cannot be studied by a Cox model (Kalbfleisch and Prentice, 2002, chapter 6.3). Aalen et al. (2008, chapter 9) also illustrate the problem of internal covariate in causality, when another covariate has an effect on both the internal covariate and the survival time. We take the example of the treatment effect as covariate, the relapses as internal time-dependent covariate and the overall survival as outcome. The effect of the treatment on the survival is both direct and indirect through the relapses. Having both treatment effect and relapses in a Cox model can result in biased estimation of the treatment effect on the overall survival, as only its direct effect is estimated. Existing solutions to handle such data are

joint modelling, multi-state models and, for the prediction purpose, landmarking. Joint modelling and landmarking approaches are presented below in this chapter. The use of multi-state model is discussed later in section 4.2.2.

2.1.5 Joint modelling to study a recurrent event and a terminal event

After a cancer diagnosis, patients may undergo disease relapses such as loco-regional relapses or distant metastases. These events may reflect the tumoural activity, and in particular the tumour aggressiveness. They may be correlated to the death, and it would thus be of interest to take them into account when estimating the patients prognosis. However, in the most well-known prognostic models in oncology such as Adjuvant! (Ravdin et al., 2001), the previous disease events are ignored in the prediction. In this thesis, we are interested in the effect of disease relapses on the risk of death. Such relapses are time-dependent internal covariates as defined above. One adequate way to study their association with death is the joint model for a recurrent event and a terminal event. This model allows to quantify the correlation between the recurrent event and the terminal event processes. It thus makes it possible to study the impact of relapses on the risk of death, as well as covariate effects on both processes. It is also the appropriate framework for the unbiased study of the recurrent event process, for which the terminal event is a non-independent censoring. Indeed, the methods previously described make the assumption that the possible censoring is an independent censoring given the covariates. That means that the underlying event process is not affected by the presence of censoring, and the censoring process carries no information about the event risk. Especially, the instantaneous hazard of event at the time s should be the same whether the individual is censored or not (Aalen et al., 2008). In the case of relapses, one censoring is the death. However, we know that the risk of relapse is null once the patient is dead. Thus, the death is clearly not an independent censoring for the relapses. Death and relapses may be seen as *semicompeting risks* as discussed by Varadhan et al. (2014) and briefly by Andersen and Keiding (2012). Joint model is a way to take this non-independence into account in the estimations.

Liu et al. (2004) were the first to propose a shared frailty model to jointly estimate the risk of recurrent event and the risk of death. This model was extended by Huang and Liu (2007) to deal with gap times for recurrences, and by Rondeau et al. (2007) to give a smooth estimate of the two survival hazard functions and the possibility to have different frailty for recurrent event and death. In these three articles, introductions give a review of the previous methods proposed to jointly analyse a recurrent event and a

terminal event. Since then, works done on the subject include Ye et al. (2007) for a marginal approach, Zeng and Lin (2009) for the use of transformation models (including proportional hazard and proportional odds models), Huang et al. (2009) who included time-dependent covariates while leaving the frailty distribution unspecified, Zeng and Cai (2010) who proposed an additive rate model and Zhangsheng and Liu (2011) who allowed for a non-parametric covariate function. A review of the models proposed in a bayesian framework has been done by Sinha et al. (2008). Finally, addition of a second type of recurrent event was proposed by Zhao et al. (2012) and Mazroui et al. (2013) while addition of a longitudinal biomarker was proposed by Kim et al. (2012).

We use the model proposed by Rondeau et al. (2007). This model particularly interests us as it estimates the baseline hazard functions using splines, and we need this functions to be carefully estimated to do some predictions. In this model, the recurrent event time T_{ij}^R and the death time T_i^D are both assumed subject to censoring independent given the covariates at time C_i. The model can be written as follows.

$$\begin{cases} \lambda_{ij}^R(t|u_i, Z_{ij}^R) = u_i \lambda_0^R(t) \exp(\beta_1' Z_{ij}^R) = u_i \lambda_{ij}^R(t|Z_{ij}^R) \\ \lambda_i^D(t|u_i, Z_i^D) = u_i^\alpha \lambda_0^D(t) \exp(\beta_2' Z_i^D) = u_i^\alpha \lambda_i^D(t|Z_i^D) \end{cases}$$

where the subscripts R and D stands for *recurrence* and *death* respectively. The two processes are estimated jointly, and share the frailty effect u_i. The u_i are independent and gamma distributed, as in the shared frailty model (see equation (2.1) on page 13). The α brings some flexibility to the model. A null α shows an independence of the recurrent event and death processes, conditionally on covariates; $\alpha < 0$ shows a negative association: a higher risk of recurrent event is associated with a lower risk of death; $\alpha > 0$ shows a positive association: a higher risk of recurrent event is associated with a higher risk of death, with the case $\alpha = 1$ showing the same frailty for both events. The covariates may be different for recurrent event and death, and the covariate effects are assumed to be different.

As stated in Rondeau et al. (2007), the assumptions of the model are as follows. First, the studied times (recurrent event, death and censoring) are *continuous*, and recurrent event and death cannot occur at the same time. In practice, when they were recorded the same day, the recurrent event was considered to occur one day before. Second, the censoring is *independent* given the covariates for both recurrent event and death, meaning that the risk of recurrent event and the risk of death are not impacted by the censoring. In particular, some unobserved recurrent events can occur after C_i with the same probability than before C_i, while they cannot occur after T_i^D. Finally, it is assumed

that there is one common censoring. In practice however, especially in cancer study, the recurrent event process may not be under observation anymore (e.g. the patient is no more followed at the hospital), while the death is still under observation (the deaths are sought in a national record).

Let $\xi = (\lambda_{ij}^R(.), \lambda_i^D(.), \beta_1, \beta_2, \theta, \alpha)$. The log-likelihood of the model is written as

$$LL(\xi) = \sum_{i=1}^{N} \left\{ \sum_{j=1}^{n_i} \delta_{ij}^R \log \lambda_{ij}^R(T_{ij}^R) + \delta_i^D \log \lambda_i^D(T_i^D) - \log \Gamma(1/\theta) - \frac{1}{\theta} \log \theta \right. \tag{2.5}$$
$$\left. + \log \int_0^\infty u^{(N_i^R(T_i^D)+\alpha\delta_i^D+1/\theta-1)} \exp\left(-u \int_0^{T_i^D} \lambda_{ij}^R(t)\mathrm{d}t - u^\alpha \int_0^{T_i^D} \lambda_i^D(t)\mathrm{d}t - \frac{u}{\theta} \right) \mathrm{d}u \right\}$$

As done for the shared frailty model, the two baseline hazards functions are estimated using splines. In this case, the penalized log-likelihood defined below is maximized.

$$pLL(\xi) = LL(\xi) - \kappa_1 \int_0^\infty \lambda_0^{R''2}(t)\mathrm{d}t - \kappa_2 \int_0^\infty \lambda_0^{D''2}(t)\mathrm{d}t$$

κ_1 and κ_2 are two smoothing parameters, allowing to balance between the fit of the data and the smoothness of the baseline hazard functions: the hazard of recurrent event for κ_1 and the hazard of death for κ_2. The penalized log-likelihood is maximised using a modified Marquardt algorithm, and the integrals are approximated using a Gauss-Laguerre quadrature, both methods being briefly presented in section 2.5. As for the shared frailty model, the variance-covariance matrix estimate is obtained by a direct or a sandwich estimator based on the inversion of the negative Hessian matrix of the penalised likelihood.

For the shared frailty model, it was possible to chose the κ parameter using the CV_a criterion. No equivalent has been proposed to maximize two parameters at a time. The solution consists in choosing the κ_1 using the CV_a in a simple shared random effect model, and the κ_2 using the CV_a in a Cox model. Then, the selected values can be used in the joint model.

In this thesis, we use the joint model in a calendar time scale for the recurrent event. The choice between the gap time scale or the calendar time scale to study recurrent events is more of a clinical discussion. The time scale has almost no impact on the parameter estimation, as shown by simulations on the impact of the time scale choice on parameter estimation presented in chapter 6.

The goodness of fit of the joint model can be assessed by the same criteria used to choose the smoothing parameter, presented in equation 2.4 (page 14), as discussed in

Commenges et al. (2007). The goodness of fit measures have, however, to be distinguished from the prediction accuracy measures. Indeed, statistically significant variables do not necessarily add much to the predictive ability of a model (Simon and Altman, 1994). Specific measures have thus to be used, as presented in the section 2.3.

2.2 Prognostic models and individual prediction of event risk

2.2.1 Use of models for predictions

The models previously described estimate covariate effects. Based on these estimations, it is possible to derive probabilities of event at chosen times that take into account the individual's characteristics $P(T_i < t|Z_i)$. This is the patient's prognosis. Both the linear predictor, defined by $\hat{\beta}'Z_i$, and the baseline hazard are used to calculate such probabilities.

The question that arises when calculating the probability of event from a shared frailty model is how to account for the frailty terms. There are two possibilities:

1. the **probability conditional on the frailty** $P(T_i < t|Z_{gi}, u_g)$: in that case, the estimated value of the frailty can be obtained by posterior mean (as discussed in details in chapter 3.2) and is then directly used in the calculation. This probability is very specific to each group of patient. Two patients with the same covariate values will get different probabilities if they belong to different groups. This approach is straightforward for new patients from a group used to develop the model. The conditional probability becomes problematic when we are interested in the prognosis of patients from new groups, for which the frailty is not estimated. This question is developed in chapter 3.

2. the **marginal probability**: the probability of event $P(T_i < t|Z_{gi})$ is an average probability over the distribution of the frailty, defined as follows:

$$P(T_i < t|Z_{gi}) = \int_0^\infty P(T_i < t|Z_{gi}, u)g(u)\mathrm{d}u$$

where $g(.)$ is the density of the frailty term as defined in equation (2.1), page 13, for example. It corresponds, for a patient with characteristics Z_{gi}, to the probability of event among all patients with the same characteristics, whatever the group. It makes it straightforward to predict event for patient from a new group.

When proportional hazard models are used for the estimation, the covariate effect is assumed holding for all times. However, the follow-up of the patients is often limited in time, and past a time point, no information is known to support that hypothesis. This lack of information is also a concern for the estimation of the baseline hazard, especially when it is non-parametric. When we have no clue about what is happening past a time point and no data to support the estimation at that time, it seems appropriate to restrict attention on a specific period. For prediction purpose, we will define that period by $[0, \tau]$, where τ has to be chosen considering the available data. For example, it can be the last observed event time.

2.2.2 Dynamic predictions

Dynamic predictions are conditional probabilities of event or survival. The probability is calculated given the information gathered until a prediction time t, including the fact that the event did not occur before t. The predictions are calculated at one or several predictions horizons $t + w$ (w stands for *window of prediction*). The *dynamic* characteristic comes from the possibility to move both t and w, and to adjust for the available information. Thus, we want to predict the probability that the event occurs in a window $[t; t + w]$, given that the event did not occur before t. Basically, it is written as:

$$P(T_i < t + w | T_i > t, Z_{ij})$$

Following the Bayes theorem (Gelman et al., 2013), stating that $P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{P(A,B)}{P(B)}$, the conditional survival probability $S(t + w|t)$ is defined by:

$$S(t+w|t) = P(T > t+w|T > t) = \frac{P(T > t + w)P(T > t|T > t + w)}{P(T > t)} = \frac{S(t + w)}{S(t)}, \forall w \geq 0$$

as we have $P(T > t|T > t + w) = 1, \forall w \geq 0$.

Including covariate effects, we have $S(t + w|t, Z_{ij}) = \frac{S(t+w|Z_{ij})}{S(t|Z_{ij})}, \forall w \geq 0$.

It is also straightforward to obtain the cumulative probability of event:

$$P(T_i < t + w|T_i > t, Z_{ij}) = \frac{S(t|Z_{ij}) - S(t + w|Z_{ij})}{S(t|Z_{ij})} = 1 - S(t + w|t, Z_{ij})$$

The landmarking approach

The landmarking approach is a convenient answer to two time-related concerns in survival analysis: time-dependent covariates and time-dependent effects, in order to obtain dynamic predictions. As explained in Van Houwelingen and Putter (2011, chapter 7),

it leads to robust predictions, not sensitive to unchecked assumptions, while keeping the model as transparent as possible. Briefly, using the notations as above, a prediction model aiming at estimating the risk of event between t and $t + w$ will be fitted on the subpopulation of patients still at risk at time t, i.e. still alive and uncensored, and ignoring events occurring after $t + w$. The landmark model is defined as:

$$\lambda(u|t, Z_i) = \lambda_0(u|t) \exp(\beta'_{LM} Z_i), \quad \text{for} \quad t \leq u \leq t + w$$

This model is not supposed to be true for all times, but gives a simple approximation of the wanted estimations. Details can be found in Van Houwelingen (2007). Based on this model, the survival function can be estimated at the two times t and $t + w$, and dynamic predictions can be derived as presented just above. The landmark times are chosen in function of our interests. Specifically, they correspond to each time at which we want to do a prediction.

To avoid fitting a new model at each desired time of prediction, a *supermodel* can be used. The idea is to stacked the data used at each landmark time in one big dataset. Then the covariate effect as a smooth function of the prediction times $\beta_{LM}(t)$ is estimated. The model used, Cox model for example, can be stratified on the prediction times to allow different baseline hazards or can contain an interaction covariate effects/stratum t. In this model, the window w of prediction has to be fixed. The supermodel was not consider further in our work, but details can be found in Van Houwelingen and Putter (2011, chapter 7).

The use of joint modelling

Joint modelling approach already has been used for prediction purpose. When the impact of a continuous internal time-dependent covariate is of interest, the joint models for longitudinal data and survival data are used. For example, they have been used to predict the time to AIDS considering the CD4 counts (Faucett et al., 2002), to predict the risk of death considering the CD4 counts (Rizopoulos, 2011) and to study the risk of prostate cancer recurrence considering the prostate-specific antigen level (Proust-Lima and Taylor, 2009). To our knowledge, at the exception of the work developed in chapter 4, no prediction were directly derived from a joint model for recurrent event and a terminal event. Only multistate or landmarking approaches have been used to include intermediate event information in the death risk prediction (see Putter et al. (2006) and Parast et al. (2011) for an example of each approach).

2.3 Assessment of prognostic performances

2.3.1 A note on the importance of prediction validation

A model is estimated on the specific data structure it is developed on. This may result in over-fitting, which is a concern in prognosis research. The over-fitting arises when the model describes too specifically the studied data, becoming inaccurate on a different dataset. However, the goal of a prognostic model is precisely to be used on new patients. To assess and correct for the over-fitting, an *internal validation* –meaning performed of the dataset used to develop the model– can be performed, by bootstrap or cross-validation, for example. However, an *external validation* performed on a totally independent dataset is of great interest. It is the only method to assess adequately the generalizability of the prediction. It consists on fitting the model on the development dataset, and then applied it on patients from an independent population. Ideally, the validation population should be related to the development one, but sufficiently different. For example, studying the same disease in another country and/or in another time period.

To validate a model means assessing its prognostic performances. Different measures exist to do so. We present below the ones that we used.

2.3.2 Discrimination: concordance measures

One very common measure to assess the prediction quality in cancer research is the concordance, also called c-index. The concordance is defined as the probability that, for two patients randomly chosen, the one with the higher predicted risk of death will be the one with the shorter survival time. The concordance value goes from 0.5 to 1. 0.5 shows a random prediction, and 1 a perfect prediction (i.e. that perfectly ranks the survival times). In practice, values lower than 0.5 can occur, showing a prediction worst than a random one. In this case, using the opposite of the prediction will do better and give a value greater that 0.5. Harrell (2001) stated that a model with a value greater than 0.8 is of utility in prediction. However, having a threshold is not so obvious, especially in cancer research, as discussed in section 3.3. Contrary to the Brier score (developed in section 2.3.4), the c-index, as a rank measure, is not sensitive to the survival probability in the population.

In binary data, the concordance is equivalent to the area under the receiver operating characteristic curve, called area under the ROC curve or AUC, which is widely use both in diagnostic and prognostic research. The concordance is also related to the Kendall's τ

K via the relation $C = \frac{K}{2} + \frac{1}{2}$ (Korn and Simon, 1990; Harrell et al., 1996). But initially, the c-index is derived from the Wilcoxon-Mann-Whitney two-sample rank test. It has been defined in Harrell et al. (1996) by *the proportion of all usable patient pairs in which the predictions and outcomes are concordant.* The concept of *usable pair* defines a pair where it is possible to know who actually dies first. That is, one of the two patients has died, and the censoring time is equal or larger than the death time. This excludes pairs with the censoring occurring before death, and those with both events at the same time. Among the usable pairs, *concordant pairs* are those with prediction coherent with observation. That is, the predicted survival time is larger for the patient who lives longer. A pairs of patients with same predicted survival time counts for half concordant (random prediction). The c-index is thus defined by:

$$\hat{C} = Concordance = P\{\hat{P}_{i'} < \hat{P}_i | T_{i'} > T_i\}$$

In the situation of proportional hazards, the condition on the predicted probabilities $\hat{P}_{i'} < \hat{P}_i$ is equivalent to compare the linear predictors $\hat{\beta}' Z_{i'} > \hat{\beta}' Z_i$ for all times t. Thus, the c-index may be defined by:

$$\hat{C} = \frac{\sum_{i=1}^{N} \sum_{i'=1}^{N} I[T_i < T_{i'}] I[\hat{\beta}' Z_i > \hat{\beta}' Z_{i'}]}{\sum_{i=1}^{N} \sum_{i'=1}^{N} I[T_i < T_{i'}]}$$

Although Harrell (2001) stated that the c-index is relatively unaffected by the amount of censoring, it has since be observed that the amount of censoring actually impacts the c-index. Thus, two measures have been proposed to overcome this problem. The first one, non-parametric, is a IPCW (inverse probability of censoring weighting) correction proposed by Uno et al. (2011). The measure he proposed is in addition limited to a time horizon τ to avoid to use the unstable tail of the prediction. The proposed measure can be written as:

$$\hat{C}_\tau = \frac{\sum_{i=1}^{N} \sum_{i'=1}^{N} \delta_i \hat{G}(T_i)^{-2} I[T_i < T_{i'}, T_i < \tau] I[\hat{\beta}' Z_i > \hat{\beta}' Z_{i'}]}{\sum_{i=1}^{N} \sum_{i'=1}^{N} \delta_i \hat{G}(T_i)^{-2} I[T_i < T_{i'}, T_i < \tau]}$$

The second measure, proposed by Gönen and Heller (2005), is parametric and also holds in the case of proportional hazards models. It relies on the model assumptions and assumes the model holds for all times. It is written as:

$$\hat{K}_n(\hat{\beta}) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{i'=i+1}^{N} \left\{ \frac{I[\hat{\beta}'(Z_{i'} - Z_i) \leq 0]}{1 + \exp[\hat{\beta}'(Z_{i'} - Z_i)]} + \frac{I[\hat{\beta}'(Z_i - Z_{i'}) < 0]}{1 + \exp[\hat{\beta}'(Z_i - Z_{i'})]} \right\}$$

Only the initial c-index was extended to clustered data by Van Oirbeek and Lesaffre (2010). More details and discussion on these concordance measures can be found in chapter 3.

2.3.3 Calibration

The calibration is a measure of performance of the model at the population level. Generally speaking, if the predicted probability of event in a population is $p\%$, then we expect than $p\%$ of the population individuals will actually have the event. This is what the calibration measures. A model is said well calibrated if, when the population is divided in risk subgroups, the true survival probabilities do not differ from the predicted ones (Van Houwelingen and Putter, 2011). Let $g_1, \ldots, g_K$ be the K risk subgroups. The calibration definition implies $E[S_{\text{true}}(t_0|g_k)] = S_{\text{prediction}}(t_0|g_k)$. For survival data, an horizon t_0 has to be chosen.

A calibration plot is a good indicator of a prediction calibration. It consists in divided the population in risk groups, most of the times the deciles of the predicted probability of event, and simply draw the observed event probability versus the predicted one for each group. The observed event probability comes with a confidence interval to see if the predicted one does not significantly differ from it. For a well calibrated model, all the points should be close to the first bisector. Finally, the histogram describing the distribution of the predicted event probabilities can be added on the plot. When this histogram is done separately for whose who did or did not have the event, it may be a representation of the prediction discrimination.

Several calibration measures have been proposed, such as calibration-in-the-large and calibration slopes, proposed with a calibration test (Steyerberg, 2010). We however think that these measures do not add much to what can be read on the calibration plot. The calibration plot is a simple way to investigate calibration and we will focus on it.

In prognosis research, having a poorly calibrated model is less a concern than having a poor discrimination. Indeed, there is the possibility to recalibrate the model, for example by refitting the baseline hazard in the new population (Steyerberg, 2010). Calibration must however always be checked to propose an useful prognosis model.

2.3.4 Prediction error: the Brier score

Another way to assess a prediction accuracy is the prediction error. It measures the distance between the observation (the survival status of the patient at a given time t_0) and the prediction (the predicted probability of event at time t_0). Several prediction error, or

explained variation, measures have been proposed (see Korn and Simon (1990); Schemper and Henderson (2000); Lawless and Yuan (2010); Stare et al. (2011) for examples). We will focus on a widely used quadratic error: the Brier score (Brier, 1950) as extended by Graf et al. (1999) and Gerds and Schumacher (2006). At a given horizon t_0, the expected Brier Score is defined by:

$$\mathrm{BS}(t_0) = \mathbb{E}(\delta_{t_0} - \hat{P}(t_0|Z_i))^2$$

As stated by Harrell (2001), the Brier score has the nice property to be maximized when the predicted probabilities are equal to the true population probabilities. Measures having such properties are said to be *proper*. A perfect prediction is only possible if the prediction $\hat{P}(t_0|Z_i)$ equals 0 or 1.

The Brier score can be divided in two terms: a true variation $S(t_0|Z)(1-S(t_0|Z))$ and the model error due to model misspecification $(S(t_0|Z) - S_{\text{pred}}(t_0|Z))$ (Van Houwelingen and Putter, 2011). This can also be seen as discrimination and calibration part. As the Brier score depends on the true survival value in the population $S(t_0|Z)$, it cannot be used directly to compare the prediction error between two populations. It can, however, be used to compare two different predictions in one population. In a population of N individuals, it is estimated by:

$$\widehat{BS}(t_0) = \frac{1}{N}\sum_{i}^{N}(\delta_{i,t_0} - \hat{P}_i(t_0|Z))^2$$

To adequately study survival data, it is necessary to deal with the censored observations. Indeed, the status at t_0 is not known for all patients, and these patients are excluded from the BS calculation. We correct the estimation using a IPCW method, as proposed in Graf et al. (1999) and Gerds and Schumacher (2006). Let $\hat{G}_N(t)$ be the Kaplan-Meier estimator of the censoring distribution. The estimator of the Brier score becomes:

$$\widehat{BS}(t_0) = \frac{1}{N}\sum_{i}^{N}(\delta_{i,t_0} - \hat{P}_i(t_0|Z))^2 w_i(t_0, \hat{G}_N)$$

where

$$w_i(t_0, \hat{G}_N) = \frac{I(\tilde{T}_i \leq t_0)\delta_{i,t_0}}{\hat{G}_N(\tilde{T}_i)} + \frac{I(\tilde{T}_i > t_0)}{\hat{G}_N(t_0)}$$

By estimating the BS at each prediction time of interest, we obtain a prediction error curve.

When doing some dynamic predictions, we want to assess the prediction error at our prediction horizon $t + w$. However, the condition that the patients are still at risk at

the prediction time t has to be accounted for. We are thus using a modified IPCW as follows:

$$w_i(t+w, \hat{G}_N) = \frac{I(\tilde{T}_i \leq t+w)\delta_{i,t+w}}{\hat{G}_N(\tilde{T}_i)/\hat{G}_N(t)} + \frac{I(\tilde{T}_i > t+w)}{\hat{G}_N(t+w)/\hat{G}_N(t)}$$

The Brier score can be calculated on cross-validation predictions for internal validation of a prediction on the development data. Finally, to be able to assess if a prediction is accurate from one population to another, and thus use it for external validation purpose, we can use a normalised brier score, as proposed by Graf et al. (1999). This corresponds to a explained-variation-like measure.

$$R^2 = 1 - \frac{\widehat{BS}(t_0)}{\widehat{BS}_0(t_0)}$$

where $\widehat{BS}_0(t_0)$ is the prediction error from the Kaplan-Meier estimate of the population survival. This R^2 measures in what extent the prediction error can be reduced by considering the proposed prediction instead of a basic, average Kaplan-Meier.

2.3.5 Other measures of prognostic performance

Other measures of prediction accuracy, more closely related to Akaike's criterion or model likelihood have been proposed, like the Kullbach-Lieber measure (Van Houwelingen and Putter, 2011) and the expected prognostic observed cross-entropy (EPOCE) proposed for joint models (Commenges et al., 2012), but where not considered in this thesis.

Finally, measures based on patients' classification and reclassification have been proposed. However, predictions from a survival model are probabilities of event at a given time, going from 0 to 1. In his book, Harrell (2001) explains us why measures based on cut-off points should be avoided. One of his arguments is that it is possible to add an highly significant factor in the model and get the percentage of correct classification that actually decreases. Recently, measures such as the net reclassification index (NRI) and the integrated discrimination improvement (IDI) have gained in popularity. The drawbacks of such methods were also discussed in Hilden and Gerds (2014). We thus do not use them in this thesis.

2.4 Incomplete data and multiple imputation

This section presents the censored times from the missing data point of view.

2.4.1 Missing data

As described by Rubin and Schenker (1991), different mechanisms can induce missing data. Let Y be the recorded data matrix, $i = 1, \ldots, n$ denotes the individual (line) and $k = 1, \ldots, p$ the covariate (column). Y can be divided between Y_{mis} and Y_{obs}, its *missing* and *observed* components, respectively. Let R be the missing indicator matrix ($R_{ij} = 1$ if Y_{ij} is missing, 0 otherwise), and Ψ the vector of the unknowns parameters leading the missing data mechanism. This mechanism is defined by $P(R|Y, \Psi)$. Missing data mechanisms can be divided in three categories, as follows.

1. MCAR (*Missing completely at random*): the missing status does not depend on the data value, missing or observed; we have $P(R|Y, \Psi) = P(R|\Psi), \forall Y, \Psi$.

2. MAR (*Missing at random*): the missing status can be explained by observed values; we have $P(R|Y, \Psi) = P(R|Y_{obs}, \Psi), \forall Y_{mis}, \Psi$.

3. MNAR (*Missing not at random*): the missing status depends on missing values, even given the observed values. For example, when we are interested in the cancer relapses, patients having relapsed and having a more severe health status are more at risk not to come to follow-up appointments. They are more disposed to be lost to follow-up, and to have a missing relapse status.

The difference between MNAR and MAR is due to the quantity of recorded information. The more information, the lowest the probability that the missing status can not be explained by this information. When data are MAR, and the parameters explaining Y are distinct from the parameters Ψ explaining the missing data mechanism, then the mechanism is said *ignorable*. In this case, the complete case analysis gives a valid inference, even when the missing data mechanism is ignored. If the data are MAR but non ignorable, inference is valid when ignoring the missing data mechanism, but not fully efficient. In the other cases, the missing data mechanism has to be accounted for.

In randomised clinical trials, the survival status can be missing. Most of the time, it is not observed due to the end of the clinical trial. In this perspective, the censoring is considered administrative and independent of the death. However, the fact that the death time is observed or not is directly related to the risk of death of the patient: the higher the risk of death, the lower the chance that the vital status is missing. We consider that, in the context of clinical trials, all the prognosis factors are recorded, and that the patient survival time is due to these prognostic factors and the treatment received. As a consequence, the missing survival status may be explained by the recorded information,

and seems to respond to the MAR mechanism. Then, ignoring the non-observed event should lead to a valid, but not fully efficient, inference. This may be so even when using specific survival methods.

2.4.2 Analysis in presence of missing data

Mattei et al. (2012) described the division of the analysis methods for incomplete data in four categories:

1. procedures based on the complete data only; it includes the complete case analysis, where individuals with at least one missing covariate (whatever the covariate is) are excluded, and the available case analysis, where individuals with missing data for specific covariates only are excluded.

2. weighting procedures, where complete cases are analysed with a weight to correct the bias related to the exclusion of incomplete cases.

3. imputation procedures (simple, multiple or resampling) followed by a standard analysis of the imputed data.

4. model-based procedures.

The authors state that, generally speaking, only imputation or model-based procedures give valid inference. They suggest that the second ones are more complicated to use. We will thus focus on multiple imputations, method widely used in the literature. It consists in building M datasets based on different imputed values. Then the M estimations are combined following rules proposed by Rubin (1996). These rules are detailed in the paper of chapter 5.

2.4.3 A few words about pseudo-values

Andersen and Perme (2010) described the use of jackknife pseudo observations to analyse censored survival data. Each observation is replaced by its contribution to the estimator of interest. The idea is to have uncensored data to be able to use standard regression methods. Pseudo-observations were recently used for clustered time to event data (Logan et al., 2011) and dynamic prediction (Nicolaie et al., 2013), both in the presence of competing risks. We are more interested in their application to study the restricted mean survival time (Andersen et al., 2004; Royston and Parmar, 2011). Although we will not detail this method here, it is an alternative to the imputation procedure. We

however chose the imputation procedure instead, as we wanted to use a Cox model as a final analysis.

2.5 Notes about the model estimation procedures

2.5.1 Splines

In the survival model that we use, the baseline hazard is approximated by splines. Splines functions are piecewise polynomials used to fit curves (Harrell, 2001). The x axis is divided into intervals, whom endpoints are called *knots*. At each knot, a new polynomials is fitted, with the constrain that polynomials smoothly joint (their derivatives must be equals) at each knot. In our model, the k knots are placed equidistantly on the x axis between time 0 and the maximum observation time.

We use cubic M-splines (normalised splines), meaning that we use polynomials of order 4 ($\beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$). The spline function is defined by (Ramsay, 1988):

$$f = \sum a_i M_i(x|k,t)$$

where $M_i(x|k,t)$ are basis splines, positively defined by recurrence on the interval $[t_i, t_{i+k}]$ and 0 elsewhere. More details can be found in Ramsay (1988). M-splines are particularly appropriate to estimate the baseline hazard function as they are positive and their antiderivatives, which exist and are defined as I-splines, can be used to define the cumulative hazard function. The spline functions are very flexible, especially when the number of knots k is high. However, in our application, we expect baseline hazards to be smooth. That is why the parameter estimation is made through a likelihood that is penalized on the baseline hazard function approximated by splines (see paragraphs 2.1.3 on shared models and 2.1.5 on joint models).

2.5.2 Marquardt algorithm and integral approximation

For our models, the likelihood is maximized using a modified Marquardt algorithm (Marquardt, 1963). This iterative algorithm is a combination between a Newton-Raphson algorithm and a steepest descent algorithm (Rondeau et al., 2003). Three conditions have to be verified to get the convergence: one on the parameters to be estimated (difference between two successive iterations $< 10^{-4}$), on the log-likelihood (difference between two successive iterations $< 10^{-4}$) and on the gradient ($< 10^{-4}$).

Finally, the likelihood as defined by the equation (2.5), page 18, contains some integrals which have no close form. To estimate these integrals, defined on $\mathbb{R}^+$, we use a

Gauss-Laguerre quadrature. This method is based on approximating the integrand using polynomials functions (Krommer and Ueberhuber, 1998). The Gauss-Laguerre quadrature was also used to approximate the integrals in the prediction formulas in chapter 4.

Chapter 3

Extension of concordance measures to shared frailty models

3.1 Question and data

As developed in the chapter 2.3, the concordance measure is the main discrimination measure used in cancer research. However, two major drawbacks have been seen for the currently proposed measures: either they are not adapted to clustered data analysed by frailty models, or their estimations are impacted by the censoring. In this chapter, we propose to adapt two non-censoring dependent measures of concordance: the IPCW estimator initially proposed by Uno et al. (2011), and the concordance probability estimator proposed for proportional hazard model by Gönen and Heller (2005). The objective is to have some measure that are both independent of censoring and suitable for clustered data. Table 3.1 briefly describes how these measures consider the pairs of patients according to observation of event or censoring.

For this purpose, two illustrating datasets were used. The first one is data from a multicentre clinical trial from the European organisation for Research and Cancer Treatment (EORTC) studying the effect of adding a boost of radiotherapy in early breast cancer. However, we studied a secondary question of interest that was to characterise the time to the occurrence of fibrosis, potentially induced by radiotherapy, as in Collette et al. (2008). We were interested to assess the discrimination of the proposed prognostic score by a measure that take into account the centres. Indeed, as shown in Figure 3.1, there was an heterogeneity in survival across the 21 centres, for each treatment arm. It was thus of interest to take into account this heterogeneity when assessing the concordance. The number of events by center is given in appendix A.

Table 3.1: Comparison of concordance measures

Pairs	**Uno**	**Gönen and Heller**
2 censorings		
At the same time	Unused	Used
At different times	Unused	Used
1 censoring and 1 event		
Censoring before event	Unused	Used
Censoring and event at the same time	Used	Used
Event before censoring	Used	Used
2 events		
At the same time	Unused	Used
At different times	Used	Used
LP(lower time)>LP(higher time)	Concordant (counts 1/1)	Used
LP(lower time)<LP(higher time)	Discordant (counts 0/1)	Used
LP(lower time)=LP(higher time)	Tied on risk (counts 0.5/1)	Tied on risk (counts 1/2)

LP: linear predictors.
counts a/b means that the pair counts for a concordant pair and b total pair.

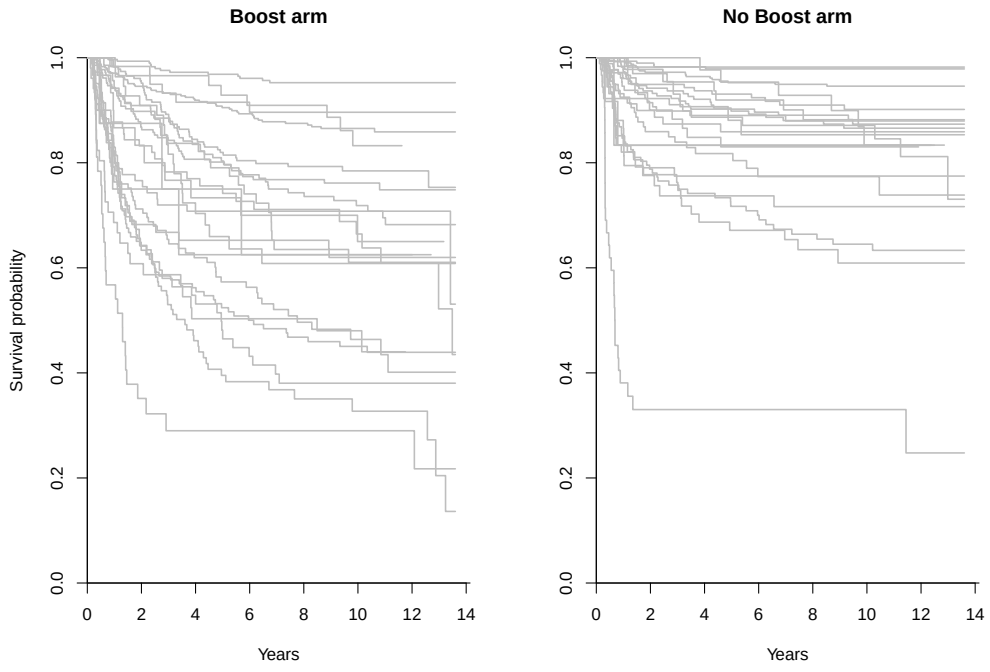

Figure 3.1: Fibrosis-free survival in the 21 institutions in EORTC data.

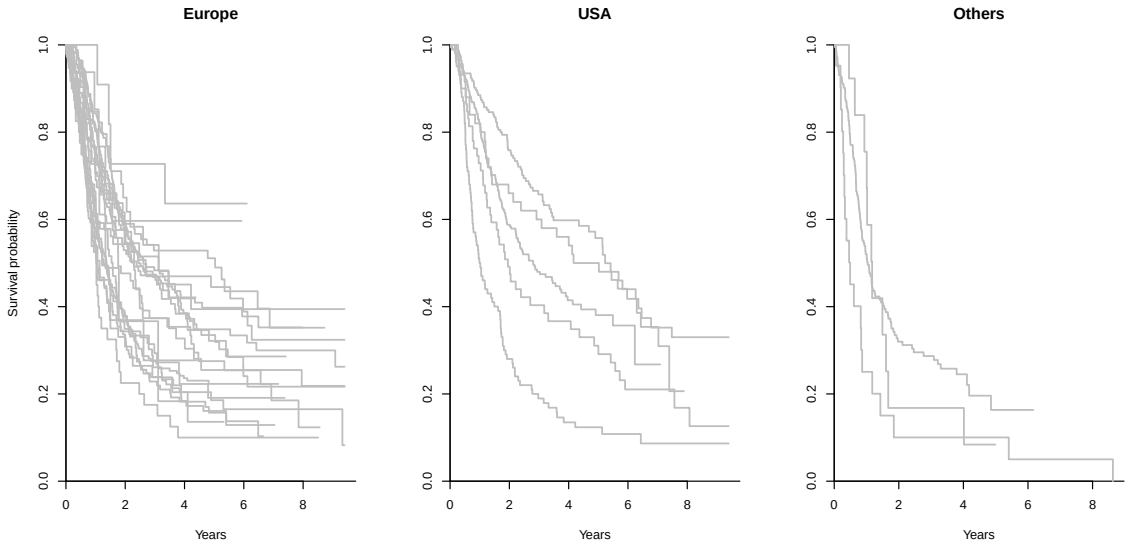

Figure 3.2: Survival in the 29 trials in MACH-NC data.

The second illustrating dataset is data from the meta-analysis of chemotherapy in head and neck cancer (MACH-NC) (Pignon et al., 2009). This meta-analysis assessed the efficacy of adding chemotherapy to locoregional treatment. We were interested in assessing the concordance of the well-known prognostic factors, and to investigate the heterogeneity between trials. Indeed, as shown in Figure 3.2, the survival differed across trials. This can impact the concordance measure value, and it is therefore of interest to have a measure that account for the trial membership. The trials were split according to the geographical area: Europe (N=21 trials) used to develop the prognostic model, and USA (N=5 trials) and others (N=3 trials) used for external validation purpose. Characteristics of the trials are detailed in appendix A.

This work has been published in Statistics in Medicine (Mauguen et al., 2013) and has been included in the  package *frailtypack* (see chapter 6).

3.2 Publication

Research Article

Statistics in Medicine

Received 31 October 2012, Accepted 24 April 2013 Published online 3 June 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5852

Concordance measures in shared frailty models: application to clustered data in cancer prognosis

Audrey Mauguen,[a,b*†] Sandra Collette,[c] Jean-Pierre Pignon[d] and Virginie Rondeau[a,b]

Frailty models are gaining interest in prognostic studies, especially because of the spread of multicenter studies. However, little research has been performed to extend prognostic tools to frailty models, including discrimination measures. As previously performed for the Harrell's c-index, we extended two different discrimination measures (the model-based concordance probability estimation of Gönen and Heller and the nonparametric Uno's c-index) to take into account cluster membership. We calculate measures at three levels: between-group, where only patients with different frailties are compared, within-group, where only patients sharing the same frailty are compared, and overall. We performed simulations to study the impact of group size and the number of groups on these measures. Results showed that the two measures can be extended to frailty models while remaining independent from censoring distribution, provided that the group size is sufficient. We apply the extended measures to two real datasets, a meta-analysis and a large multicenter trial. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: prognostic model; concordance; frailty model; clustered data; cancer

1. Introduction

In cancer research, knowing the prognosis of a patient is a key information for treatment choice. Thus, the development of accurate validated prognostic models has become a major issue, as well as the evaluation of the usefulness of a new marker predictive of treatment effect or survival [1, 2]. This assumes that prognostic model ability can be properly assessed and that prediction accuracy can be identified. In addition, the ability of each prognostic model needs to be validated on independent datasets. Assessment of the accuracy of a prognostic model differs from classical goodness-of-fit measures and from statistical significance of the prognostic variables because statistically significant variables do not necessarily add much to the predictive ability of a model [3–5].

Shared frailty models have been developed to take into account an existing but unmeasured heterogeneity between patients of different groups [6–8]. For example, characteristics such as genetic information or environmental exposure can be shared by patients of a group (e.g., family, hospital) and can influence time to the studied event. Frailty models are an extension of proportional hazards survival models. Dependence is produced by sharing an unobserved variable that is treated as a random effect, the frailty. Patients of the same group share the same frailty, and frailties are independent between the groups. The assumption is that individuals with high frailty will have the event first. These models are useful in prognostic research, where some studies include patients from several centers, countries, or

[a] *Univ. Bordeaux ISPED, Centre INSERM U897-Epidémiologie-Biostatistique, F-33000 Bordeaux, France*
[b] *INSERM, ISPED, Centre INSERM U897-Epidémiologie-Biostatistique, F-33000 Bordeaux, France*
[c] *EORTC HeadQuarters, B-1200 Bruxelles, Belgique*
[d] *Service de Biostatistique et d'Epidémiologie, Institut de Cancérologie Gustave-Roussy, F-94805 Villejuif, France*
**Correspondence to: Audrey Mauguen, INSERM U897 - Equipe de biostatistique, ISPED, Université Bordeaux Segalen, 146 rue Leo Saignat, 33076 BORDEAUX CEDEX, France.*
[†] *E-mail: audrey.mauguen@isped.u-bordeaux2.fr*

randomized trials, resulting in clustered data structure. Thus, there is a need to adapt prognostic tools to this kind of model.

To assess the ability of a prognostic model, two dimensions are usually used: discrimination and calibration [9]. We focus here on discrimination. Discrimination is the ability of a model to separate patients with good outcome from those with poor outcome. Concordance is one measure of this ability and is widely used when assessing new prognostic models. Concordance is defined as the probability that, between two patients randomly chosen, the one who has the shorter predicted survival time will be the one with the shorter observed survival time. The comparison of the predicted times can be substituted by the comparison of the estimated survival probabilities at each time t, when we are in a proportional hazards models framework. In proportional hazards models, it is also equivalent to the probability that, between two patients randomly chosen, the one who has the higher predicted risk of event will be the one who has the shorter observed survival time. Let T_1 and T_2 be the two observed survival times, x_1 and x_2 the vectors of covariates, and $\hat{\beta}$ the covariate effects estimated by a proportional hazards model. The concordance can thus be defined by

$$Concordance = \mathbb{P}\left\{T_2 > T_1 | \hat{\beta}' x_2 < \hat{\beta}' x_1\right\}$$

or by

$$Concordance = \mathbb{P}\left\{\hat{\beta}' x_2 < \hat{\beta}' x_1 | T_2 > T_1\right\}$$

As detailed in [10] and [11], the estimation of the concordance can be restricted to the interval $[0, \tau]$, where τ is chosen to avoid considering the tail part of the estimated survival function of T, which can be unstable. The concordance value varies from 0 to 1, 1 showing a perfect concordance, 0.5 showing random prediction, and 0 showing perfect discordance. When values are below 0.5, considering the opposite values of the prediction will result in values above 0.5 (see [12] for more details). Thus, it is generally admitted that concordance is between 0.5 and 1.

Concordance values can be used for two different objectives. First, they can be used to compare the prognostic ability of two models on one given population, to help choose the best prognostic model. Second, they can be used to validate prognostic models and therefore to compare the prognostic ability of a model between different populations. These populations may not be comparable, especially with regard to censoring distribution. It follows that a desirable property of a good concordance measure is its independence from censoring distribution. That is, the expected value of the concordance should be the same in all the populations compared. However, to date, only an extension of the Harrell's c-index [9] has been proposed for frailty models [13]. This measure is not independent of the amount of censoring, as it cannot classify patients who did not undergo the event at a known date [10, 12]. Its calculation is restricted to the pairs of patients deemed usable, which implies that the patient with the shorter observed time actually undergoes the event (not censored), without any correction to take these exclusions into account.

Our objective using the methodology employed by Van Oirbeek and Lesaffre [13] is to adapt two types of concordance measures that were shown to be independent of censoring distribution. The first is the concordance probability estimator proposed by Gönen and Heller [12], which is a model-based measure valid for the proportional hazards models. The second is the correction of the Harrell's c-index proposed by Uno, a nonparametric measure [10]. We aimed to assess the impact of group size and the number of groups on concordance values. We also wanted to assess the influence of the censoring rate on concordance estimations. The performance of frailty model estimations has been extensively studied elsewhere, so we did not focus on this subject [14–16]. Section 2 of this article briefly describes the shared frailty models, the estimation of the frailty terms, and the extension of the two concordance measures to frailty models. In section 3, we evaluate the two extended measures in a set of simulations. In section 4, we apply the proposed approaches to two datasets with clustered data. Finally, section 5 presents the discussion and conclusion.

2. Methods

2.1. Shared frailty models

Shared frailty models are proportional hazards models that are extended to take into account correlations between observations, a scenario that occurs especially in clustered data or when studying recurrent

events [6]. The correlation is accounted for by the means of a random effect term, called frailty, which is shared by all the observations of one group. The model defines the risk of event at time t by

$$h_{ij}(t) = h_0(t)u_i \exp\left(\beta' x_{ij}\right)$$

where i indicates the group ($i = 1, \ldots, G$) and j indicates the subject ($j = 1, \ldots, n_i$). $h_0(t)$ is the baseline risk of event, that is, the risk of event for patients with all x equal to zero and a frailty u_i equal to one. In our model, the frailty terms are independent gamma distributed, with a mean 1 and variance θ. That is, $u_i \sim \Gamma\left(\frac{1}{\theta}; \frac{1}{\theta}\right)$.

We use a semiparametric penalized likelihood approach to estimate the different parameters: the regression coefficients β, the variance of the random effects θ, and the baseline hazard function $h_0(.)$. In most situations, it is reasonable to expect a smooth baseline hazard function, the piecewise constant modeling for the hazard function often being unrealistic. To introduce such a priori knowledge, we penalize the likelihood by a term that has large values for rough functions. The estimator $\hat{h}_0(.)$ cannot be calculated explicitly but can be approximated on the basis of splines. Splines are piecewise polynomial functions that are combined linearly to approximate a function on an interval [17]. For more details on the estimation procedure, see [14, 18].

An important issue when using frailty models in prognostic studies is the prediction of the frailties at an individual level. The $\hat{u}_i s$ are obtained from the posterior distribution of the $u_i s$ conditional on the observed data, knowing the estimated values of the regression parameters [19]. More specifically, from the Bayes theorem, the conditional distribution $f_U(u_i|data)$ is equal to

$$f_U\left(u_i|h_0(.), \theta, \beta\right) = \frac{f_i\left(h_0(.), \theta, \beta|u_i\right) f_U(u_i)}{f_{\text{marg},i}\left(h_0(.), \theta, \beta\right)}$$

where $f_i(h_0(.), \theta, \beta|u_i)$ is the conditional density of the group i and $f_{\text{marg},i}(h_0(.), \theta, \beta)$ is the marginal density. The marginal density is equal to the conditional one integrated over the distribution of the frailty effects. That is,

$$f_{\text{marg},i}\left(h_0(.), \theta, \beta\right) = \int_0^\infty f_i\left(h_0(.), \theta, \beta|u_i\right) f_U(u_i) \mathrm{d}u_i$$

As $u_i \sim \Gamma\left(\frac{1}{\theta}; \frac{1}{\theta}\right)$, we have

$$f_U(u_i) = \frac{u^{1/\theta - 1} \exp(-u/\theta)}{\theta^{1/\theta}\Gamma(1/\theta)}$$

and $E(u_i) = 1$ and $\text{var}(u_i) = \theta$. Finally, we obtain

$$f_U(u_i|h_0(.), \theta, \beta) = \frac{u_i^{m_i + 1/\theta - 1} \exp\left(-u_i\left(1/\theta + \sum_{j=1}^{n_i} H_{ij}\left(t_{ij}\right)\right)\right)\left(1/\theta + \sum_{j=1}^{n_i} H_{ij}\left(t_{ij}\right)\right)^{1/\theta + m_i}}{\Gamma(m_i + 1/\theta)}$$

where m_i is the number of events in the group i and $H_{ij}(t_{ij})$ is the cumulative risk of events at the observation time t_{ij}. This expression corresponds to a gamma density with parameters $(1/\theta + m_i)$ and $\left(1/\theta + \sum_{j=1}^{n_i} H_{ij}\left(t_{ij}\right)\right)$, and therefore, the expected value of the frailty is given by

$$\hat{u}_i = \frac{(1/\theta + m_i)}{\left(1/\theta + \sum_{j=1}^{n_i} H_{ij}\left(t_{ij}\right)\right)} \tag{1}$$

and the corresponding variance estimates is $\frac{(1/\theta + m_i)}{\left(1/\theta + \sum_{j=1}^{n_i} H_{ij}\left(t_{ij}\right)\right)^2}$. Consequently, if H_{ij} is well estimated by the model, $\sum_{j=1}^{n_i} H_{ij}(t_{ij})$ must be close to m_i, and the mean of the predicted $\hat{u}_i$ must be close to 1. We note $\hat{u}$ the vector of the predicted frailties.

Shared frailty models and linear predictors were computed using the R package *frailtypack* [20], in which concordance values were included.

2.2. *Extension of Uno's c-index to frailty models*

Harrell's c-index was initially developed in the framework of Cox models [21]. It can be defined as the probability that a pair of patients randomly chosen is concordant, that is, the patient with the higher predicted risk of event is the one with the actual shorter survival time. It is calculated as the number of concordant pairs over the number of comparable pairs. Not all the pairs are comparable because if one patient of the pair is censored, the exact time of event is unknown, and we cannot always conclude which patient had the event first. The pairs that are not comparable are excluded from the calculation of Harrell's c-index. The consequence of such exclusions is that the value of this c-index is dependent on the censoring distribution in the studied population [10].

To overcome this weakness, Uno *et al*. proposed to correct this index by using an inverse probability weighting technique that weights each pair comparison by the probability that the two observations are not censored [10]. They used a truncated version of the concordance in the interval $[0, \tau]$. Let $j = 1, \ldots, n$ and $j' = 1, \ldots, n$ denoted two subjects, and $\hat{T}_c(.)$ be the Kaplan–Meier estimate of the censoring distribution in the population. T_j indicates the survival time of patient j and x_j is his covariate vector. Δ_j is the indicator of event for patient j ($\Delta_j = 1$ if T_j is an event time, 0 if T_j is a censoring time). Uno's c-index is defined as follows:

$$\hat{C}_\tau = \frac{\sum_{j=1}^{n} \sum_{j'=1}^{n} \Delta_j \left\{\hat{T}_c(T_j)\right\}^{-2} I\left[T_j < T_{j'}, T_j < \tau\right] I\left[\hat{\beta}' x_j > \hat{\beta}' x_{j'}\right]}{\sum_{j=1}^{n} \sum_{j'=1}^{n} \Delta_j \left\{\hat{T}_c\left(T_j\right)\right\}^{-2} I\left[T_j < T_{j'}, T_j < \tau\right]}$$

where $I[.]$ represents an indicator function.

The c-index can be calculated for a frailty model by including the estimation of the random effect in the predicted risk. However, as proposed by Van Oirbeek and Lesaffre for the Harrell's c-index, a further development is to take group membership into account [13]. Three measures can then be defined. The first is a within-group measure, where the concordance is based on pairs of patients of the same group, that is sharing the same frailty. The second is a between-group measure, where only patients from different groups are compared. Finally, a third overall measure is a weighted mean of the previous two. Van Oirbeek and Lesaffre proposed to calculate these three measures both conditionally on the random effect and marginally. Both take into account the heterogeneity of the data in estimating the parameters, but the conditional one adds the random effect estimates directly in the risk prediction, whereas the marginal one does not. We focus here on conditional measures only, which are of greater interest in frailty models. Note that the within measure is the same at both levels.

Following the same idea, we propose three measures of Uno's c-index for the frailty models framework.

First, we defined a within-group estimator, in which only observations sharing the same frailty are compared. An estimation of concordance is made for each group and averaged to obtain one within-group estimator as follows. Let $i = 1, \ldots, G$ defines the group, and ij the subject j of the group i ($j = 1, \ldots, n_i$). In the following, we denote by ij and ij' two patients from the same group i and by ij and $i'j'$ two patients from two different groups i and i' ($i \neq i'$). The within-group Uno's c-index $\hat{C}_{\tau,W}$ is defined by

$$\hat{C}_{\tau,W} = \frac{1}{G} \sum_{i=1}^{G} \left[\frac{\sum_{j=1}^{n_i} \sum_{j'=1}^{n_i} \Delta_{ij} \left\{\hat{T}_c(T_{ij})\right\}^{-2} I\left(T_{ij} < T_{ij'}, T_{ij} < \tau\right) I\left(\hat{\beta}' x_{ij} > \hat{\beta}' x_{ij'}\right)}{\sum_{j=1}^{n_i} \sum_{j'=1}^{n_i} \Delta_{ij} \{\hat{T}_c(T_{ij})\}^{-2} I\left(T_{ij} < T_{ij'}, T_{ij} < \tau\right)} \right]$$

where $\hat{T}_c(.)$ is still estimated on the whole population. The frailty terms are not included directly in the calculation of the within-group concordance as they are the same for the compared patients in each pair, but they make it possible to compute a more accurate estimation of the β parameters. This concordance measures the discriminatory power of only the observed covariates.

Then, we defined a between-group estimator, implying only comparisons between patients of different groups. This estimator includes the estimated frailty terms, which are different for two compared patients. Here, we take patients in the whole population, and i and i' indicate the

membership group of patients j and j', respectively. For this measure, we select two patients who do not belong to the same group ($i \neq i'$). The between-group Uno's c-index $\hat{C}_{\tau,B}$ is defined by

$$\hat{C}_{\tau,B} = \frac{\sum_{i=1}^{G}\sum_{j=1}^{n_i}\left\{\sum_{\substack{i'=1\\ i'\neq i}}^{G}\sum_{j'=1}^{n_{i'}}\Delta_{ij}\{\hat{T}_c(T_{ij})\}^{-2}I(T_{ij}<T_{i'j'},T_{ij}<\tau)I(\hat{\beta}'x_{ij}+ln(\hat{u}_i)>\hat{\beta}'x_{i'j'}+ln(\hat{u}_{i'}))\right\}}{\sum_{i=1}^{G}\sum_{j=1}^{n_i}\left\{\sum_{\substack{i'=1\\ i'\neq i}}^{G}\sum_{j'=1}^{n_{i'}}\Delta_{ij}\{\hat{T}_c(T_{ij})\}^{-2}I(T_{ij}<T_{i'j'},T_{ij}<\tau)\right\}}$$

where n is the number of subjects in the whole population and $\hat{u}_i$ and $\hat{u}_{i'}$ are the estimation of the frailty parameters of groups i and i', respectively. This concordance measures the discriminatory power of both measured covariates and frailty terms.

To sum up these two information levels, an overall measure $\hat{C}_{\tau,O}$ is calculated which is equal to a mean of the two previous indices and is weighted by the number of within-group and between-group pairs.

$$\hat{C}_{\tau,O} = \frac{n_{\text{comp},W}}{n_{\text{comp},T}}\hat{C}_{\tau,W} + \frac{n_{\text{comp},B}}{n_{\text{comp},T}}\hat{C}_{\tau,B}$$

with $n_{\text{comp},W}$, $n_{\text{comp},B}$, and $n_{\text{comp},T}$ being the number of comparable pairs (as defined at the beginning of this section) within-group, between-group, and overall, respectively. This overall measure is not meant to compare directly the discriminatory power on different populations, as it is dependent on the structure of the data and the relative weight of $\hat{C}_{\tau,B}$ and $\hat{C}_{\tau,W}$ for a given model. However, it can be used to compare different models in one population. It is also a simple way to obtain an overall view of a model discrimination in all the patients of a given population. It is the value closest to the initial concordance measure.

2.3. *Extension of the Gönen and Heller's concordance probability estimation to frailty models*

Gönen and Heller proposed the concordance probability estimation for Cox proportional hazards models [12]. By being related to the importance of the $\beta' x_j$ values, it is a straight measure of the separation between patients' predicted risks. Gönen and Heller showed in their paper how its interpretation is equivalent to the c-index in the framework of Cox proportional hazards models. The concordance is defined by

$$K(\beta) = \mathbb{P}\left(T_{j'} > T_j | \beta' x_j \geq \beta' x_{j'}\right)$$

Using

$$\mathbb{P}\left\{T(\beta' x_{j'}) > T(\beta' x_j)\right\} = \frac{1}{1+\exp\left[\beta'(x_{j'}-x_j)\right]}$$

where $T\left(\beta' x_j\right)$ is the survival time corresponding to the linear combination $\beta' x_j$, we obtain

$$K(\beta) = \frac{\int\int\left[1+\exp\left[\beta'\left(x_{j'}-x_j\right)\right]\right]^{-1} \mathrm{d}F\left(\beta' x_j\right)\mathrm{d}F\left(\beta' x_{j'}\right)}{\int\int \mathrm{d}F\left(\beta' x_j\right)\mathrm{d}F\left(\beta' x_{j'}\right)}$$

The concordance probability estimation is finally estimated by

$$\hat{K}_n(\hat{\beta}) = \frac{2}{n(n-1)}\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}\left\{\frac{I\left[\hat{\beta}'(x_{j'}-x_j)\leq 0\right]}{1+\exp\left[\hat{\beta}'(x_{j'}-x_j)\right]} + \frac{I\left[\hat{\beta}'(x_j-x_{j'})<0\right]}{1+\exp\left[\hat{\beta}'(x_j-x_{j'})\right]}\right\}$$

where $I[.]$ represents an indicator function. As previously discussed, x_j is the covariates vector of the patient j, and $\hat{\beta}$ is the vector of covariate effect estimated with a Cox model.

This measure was shown to be robust to censoring because the effect of the observed times is mediated through the partial likelihood estimator $\hat{\beta}$ and the effect of censoring on the bias of $\hat{\beta}$ is negligible. The value of $\hat{\beta}' x_j$ is known for all j, and the comparison $(x_j - x_{j'})$ can always be made. Thus, all pairs of subjects are taken into account in the concordance estimation. The way this estimation handles the ties

on prediction is similar to Harrell's and Uno's c-indexes: They accounted for 0.5 in the concordance, meaning that for these patients, the prognostic model is unable to determine which one will live longer.

We extend the concordance probability estimation for application in shared frailty models in a way similar to the previous measure.

The within-group Gönen and Heller's $\hat{K}_W(\hat{\beta})$ is defined as follows:

$$\hat{K}_W(\hat{\beta}) = \frac{1}{G}\sum_{i=1}^{G}\left[\frac{2}{n_i(n_i-1)}\sum_{j=1}^{n_i-1}\sum_{j'=j+1}^{n_i}\left\{\frac{I\left[\left(\hat{\beta}'(x_{ij'}-x_{ij})\right)\leq 0\right]}{1+\exp\left[\hat{\beta}'(x_{ij'}-x_{ij})\right]}+\frac{I\left[\left(\hat{\beta}'(x_{ij}-x_{ij'})\right)<0\right]}{1+\exp\left[\hat{\beta}'(x_{ij}-x_{ij'})\right]}\right\}\right]$$

As explained previously, the frailty terms are not included directly in the calculation of the concordance within-group. The between-group Gönen and Heller's $\hat{K}_B\left(\hat{\beta},\hat{u}\right)$ is defined as follows:

$$\hat{K}_B\left(\hat{\beta},\hat{u}\right) = \frac{1}{n_B}\sum_{i=1}^{G-1}\sum_{j=1}^{n_i}\left[\sum_{i'=1+1}^{G}\sum_{j'=1}^{n_{i'}}\left\{\frac{I\left[\left(\log\left(\frac{\hat{u}_{i'}}{\hat{u}_i}\right)+\hat{\beta}'(x_{i'j'}-x_{ij})\right)\leq 0\right]}{1+\frac{\hat{u}_{i'}}{\hat{u}_i}\exp\left[\hat{\beta}'(x_{i'j'}-x_{ij})\right]}+\frac{I\left[\left(\log\left(\frac{\hat{u}_i}{\hat{u}_{i'}}\right)+\hat{\beta}'(x_{ij}-x_{i'j'})\right)<0\right]}{1+\frac{\hat{u}_i}{\hat{u}_{i'}}\exp\left[\hat{\beta}'(x_{ij}-x_{i'j'})\right]}\right\}\right]$$

where n_B is the total number of between-group pairs, $n_B = \sum_{i=1}^{G-1} n_i\left(n-\sum_{k=1}^{i} n_k\right)$. Here again, all patients are considered whatever their membership group, but only those belonging to two different groups are compared ($i \neq i'$). Given $\hat{\beta}$, the value of $\hat{K}_B\left(\hat{\beta},\hat{u}\right)$ is directly related to the values of the frailty terms $\hat{u}_i$ and $\hat{u}_{i'}$, obtained with the equation (1).

Finally, the overall estimator $\hat{K}_O\left(\hat{\beta}\right)$ is defined as follows:

$$\hat{K}_O\left(\hat{\beta}\right) = \frac{n_W}{n_T}\hat{K}_W\left(\hat{\beta}\right)+\frac{n_B}{n_T}\hat{K}_B\left(\hat{\beta},\hat{u}\right)$$

with n_W, n_B, and n_T being the number of pairs between-group, within-group, and overall, respectively. Here, all pairs of patients in the population are used.

2.4. *Variability of the concordance estimators*

The standard error of the concordance measures was obtained by $B = 500$ bootstrap replicates. We wanted to focus on nonparametric estimation of the variance. Field *et al*. suggested that using a simple cluster bootstrap, in which the clusters are randomly chosen with replacement, gives consistent nonparametric estimation of the two first moments of the variable of interest when $G \to \infty$ [22]. They also suggested that two-stage and reverse bootstraps, in which both clusters and individual in clusters are resampled, generate excess variation, and produce consistent estimation of the variance of the variable of interest provided that both G and $n_i \to \infty$. Thus, we chose to perform a simple cluster bootstrap.

Let $\widehat{CM}$ be our concordance measure $\left(\hat{C}_{\tau,W}, \hat{C}_{\tau,B}, \hat{C}_{\tau,O}, \hat{K}_W(\hat{\beta}), \hat{K}_B\left(\hat{\beta},\hat{u}\right), \text{or } \hat{K}_O(\hat{\beta})\right)$ estimated on the original dataset. We kept and applied the estimated coefficients on the bootstrap population. Let $\widehat{CM}^b$ be the concordance measure estimated on the bth bootstrap dataset and $\widehat{CM}^B = \frac{1}{B}\sum_{b=1}^{B}\widehat{CM}^b$ be the bootstrap mean of the concordance measure. We define the variance of $\widehat{CM}$ by

$$\hat{\text{var}}\left(\widehat{CM}\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{CM}^b-\widehat{CM}^B\right)^2$$

3. Simulation study

3.1. Setting

We were interested in the effect of the group size, the number of groups, and the effect of the percentage of censoring on the estimation of the concordance. In this aim, we simulated datasets firstly with a fixed number of 40 groups and varying group sizes: 2, 10, and 50 patients per group; and secondly, with a varying number of groups: 10, 40, and 100 groups of fixed size ($n = 40$ patients per group). This resulted in 80 to 4000 patients in each dataset. For each setting, we generated 1000 datasets.

The $u_i s$ were randomly chosen from a gamma distribution with a variance $\theta = 0.8$. Given u_i, independent survival times $T^*_{i1}, \ldots, T^*_{in_i}$ were generated from a Weibull distribution [23]. Right-censoring times $T_{c,ij}$ were generated from a uniform distribution $[0, c]$ where c was selected to target a percentage of censoring around 0%, 20%, 50%, and 70%. In a fifth setting, the target of the percentage of censoring was different in each group, to assess the effect of a group-specific censoring on the concordance value. The c was fixed for each cluster, increasing uniformly from $c_{\min}$ for the first cluster to $c_{\max}$ for the last one. The use of a uniform distribution rather than a fixed one allowed us to assess Uno's c-index with weights different from 1. To calculate the Uno's c-index and ensure that $\mathbb{P}(T_{c,ij} > \tau) > 0$, we set $\tau = 0.9c$. For each patient, we considered the observed time as $T_{ij} = \min(T^*_{ij}, T_{c,ij})$ and the indicator of event was $\delta_{ij} = I[T^*_{ij} \leqslant T_{c,ij}]$. Considering that, in a prognostic model, all covariates can be aggregated in a single linear predictor, we simulated one continuous covariate x_{ij} following a uniform distribution $U(0, 1)$ with a size effect $\beta = 3$. Considering the time needed to compute bootstrap standard error at each of the 1000 simulation steps, we only present the empirical standard error for the concordance measure for each setting.

We calculated reference values, considered as 'true' values, by simulating one dataset of 400,000 observations and computing measures of concordance using the true value of β, the generated u_i, and the generated survival times without censoring. For the fixed group number setting, we maintained the number of groups at 40, and the group size was 10,000. For the fixed group size setting, we maintained the group size at 40, and the number of groups was 10,000. We computed true values only for between-group and within-group measures, considering that the overall measure can not be compared between the simulation sets and the true value set. Bias was defined as the difference between the estimated measure and the true; thus, a positive value means that we overestimated the true value of concordance. Relative bias was defined by the bias divided by the true value.

3.2. Simulation results

Table I summarizes simulation results. For a low sample size ($n_i = 2$ patients per group), estimations were moderately biased for both concordance types, especially in the between-group level. The value of the bias increased with the amount of censoring. The presence of bias in the concordance estimation may partly be due to the lack of information, which results in less accurate parameter estimation. For all other settings, the bias was low (relative bias <5%), and even null at the within-group level, when both n_i and G were sufficiently large. Note that the bias decreased as the group size increased, showing the impact of the group size on the concordance value, whereas the bias was more stable for all number of groups when the group size was 40 patients per group. This suggests that when the group size is sufficient, the number of groups has little impact on the concordance value. In the settings where the bias was not null, it was increased by the amount of censoring, but the influence of censoring seemed to be controlled by the group size. As expected, the standard error of the measures was decreased when the number of groups G or the sample size n_i increased, but it increased when the censoring rate increased. The Gönen and Heller approach tended to underestimate the true concordance value, with relative biases from −0.044 to 0.009, whereas Uno's c-index tended to overestimate it, with relative biases from 0.000 to 0.090. Considering the within-group Uno's c-index, we calculated the weights on the basis of the censoring distribution in the whole population. When we computed the weights based on the censoring distribution in each group separately, we obtained very similar results in these settings (data not shown). Finally, simulation of group-dependent censoring had quite no impact on the concordance results.

Concordance was always greater at the between-group level, which includes the frailty estimation, than at the within-group level, which does not include it. For instance, in the absence of censoring, when the number of groups was 100 ($n_i = 40$), the true value of within-group concordances, which takes into account only the observed covariates, was 0.710 and 0.709 for Gönen and Heller's measure and for

Table I. Simulation study: mean, standard error (se), and relative bias (Rbias) for between and within concordance measures for varying group size (n_i), number of groups (G) and censoring rate, and mean of the model estimations of β and u_i among 1000 datasets. In bold are the true values of concordance measures for each setting.

C(%)	mean($\hat{\beta}$)	mean($\hat{u}_i$)	Gönen & Heller's $\hat{K}(\hat{\beta})$						Uno's $\hat{C}_\tau$					
				Between			Within			Between			Within	
			Mean	(se)	Rbias	Mean	(se)	Rbias	Mean	(se)	Rbias	Mean	(se)	Rbias
$G = \mathbf{40}$			**0.768**			**0.710**			**0.768**			**0.710**		
$n_i = 2$														
3.2	2.93	1.01	0.736	0.044	−0.041	0.703	0.041	−0.009	0.802	0.047	0.045	0.712	0.074	0.003
20.3	3.13	1.01	0.746	0.044	−0.029	0.712	0.043	0.004	0.814	0.049	0.060	0.716	0.078	0.008
50.2	3.08	1.00	0.734	0.053	−0.044	0.706	0.059	−0.005	0.826	0.067	0.076	0.716	0.101	0.009
69.9	3.40	1.00	0.735	0.061	−0.043	0.716	0.071	0.009	0.837	0.088	0.090	0.727	0.123	0.024
35.2*	3.18	1.01	0.742	0.046	−0.034	0.714	0.046	0.006	0.816	0.054	0.063	0.718	0.085	0.012
$n_i = 10$														
3.3	3.01	1.02	0.769	0.016	0.001	0.709	0.013	0.000	0.784	0.016	0.021	0.710	0.016	0.000
20.1	3.03	1.03	0.768	0.017	0.000	0.711	0.014	0.001	0.786	0.016	0.023	0.711	0.020	0.002
50.0	3.00	0.99	0.761	0.020	−0.009	0.709	0.022	−0.001	0.798	0.019	0.039	0.722	0.027	0.017
70.0	2.99	0.99	0.751	0.022	−0.022	0.707	0.027	−0.004	0.811	0.025	0.055	0.726	0.035	0.023
35.1*	3.03	1.02	0.765	0.018	−0.004	0.711	0.016	0.002	0.789	0.017	0.027	0.714	0.024	0.005
$n_i = 50$														
3.4	3.00	1.04	0.773	0.012	0.007	0.710	0.005	0.000	0.776	0.012	0.011	0.710	0.006	0.000
20.1	3.03	1.05	0.773	0.012	0.006	0.711	0.006	0.002	0.776	0.012	0.011	0.711	0.007	0.002
49.8	3.01	1.01	0.771	0.013	0.003	0.710	0.007	0.000	0.781	0.013	0.017	0.716	0.011	0.009
70.2	3.01	1.00	0.768	0.013	0.000	0.710	0.009	0.000	0.791	0.014	0.030	0.721	0.015	0.016
35.1*	3.03	1.05	0.773	0.013	0.006	0.711	0.006	0.002	0.778	0.013	0.012	0.711	0.009	0.002

$n_i = 40$		0.774			0.710			0.774			0.709		
$G = 10$													
3.3	3.02	1.09	0.773	0.026	−0.002	0.711	0.012	0.001	0.777	0.026	0.004	0.710	0.014	0.001
20.0	3.03	1.10	0.772	0.026	−0.003	0.711	0.013	0.002	0.776	0.026	0.003	0.711	0.016	0.002
50.0	2.98	0.98	0.768	0.028	−0.008	0.708	0.018	−0.003	0.782	0.029	0.011	0.715	0.025	0.008
70.1	3.01	0.96	0.767	0.028	−0.010	0.709	0.022	0.000	0.794	0.031	0.026	0.722	0.033	0.018
36.9*	3.03	1.07	0.771	0.026	−0.004	0.711	0.015	0.002	0.778	0.027	0.006	0.711	0.022	0.003
$G = 40$														
3.3	3.00	1.03	0.772	0.013	−0.003	0.710	0.006	0.000	0.776	0.013	0.002	0.710	0.007	0.000
20.1	3.02	1.05	0.773	0.013	−0.002	0.711	0.006	0.001	0.777	0.013	0.004	0.711	0.008	0.002
50.0	3.00	1.00	0.771	0.014	−0.005	0.710	0.009	0.000	0.783	0.014	0.011	0.716	0.012	0.010
70.0	2.99	0.99	0.765	0.014	−0.012	0.709	0.011	−0.001	0.791	0.015	0.022	0.721	0.016	0.016
35.0*	3.02	1.04	0.771	0.013	−0.004	0.710	0.007	0.001	0.778	0.013	0.005	0.711	0.010	0.003
$G = 100$														
3.3	3.00	1.02	0.772	0.007	−0.003	0.710	0.004	0.000	0.776	0.007	0.002	0.710	0.004	0.001
20.0	3.02	1.03	0.772	0.008	−0.002	0.711	0.004	0.001	0.777	0.008	0.004	0.711	0.005	0.002
50.0	3.00	1.00	0.770	0.008	−0.006	0.710	0.005	0.000	0.782	0.008	0.010	0.716	0.007	0.009
70.0	3.01	1.00	0.766	0.008	−0.011	0.710	0.006	0.001	0.792	0.009	0.023	0.721	0.011	0.017
34.6*	3.02	1.03	0.772	0.008	−0.004	0.711	0.004	0.001	0.778	0.008	0.005	0.711	0.007	0.003

*Group dependent censoring.

Uno's measure, respectively, and the value of between-group concordances, which take into account the observed covariates and frailty effect, was 0.774 for both measures. This result suggests the interest of taking frailty estimation into account to obtain more accurate predictions.

4. Application

These applications aim at assessing the discriminatory ability of survival prognostic models with clustered data. We will illustrate measures of concordance on two datasets using the R package frailtypack [20]. The first is a large multicenter trial studying the effect of a radiotherapy boost on fibrosis onset in patients with breast cancer. The second example is a meta-analysis in which patients are grouped in randomized trials, where the effect of chemotherapy in head and neck cancer is studied. The two examples are defined by small to large groups for the multicenter trial and by several large groups for the meta-analysis.

In these applications, the goodness of fit of the frailty and Cox models is given by an approximate likelihood cross-validation criterion (LCV) [24]. In the case of parametric approach, LCV is approximately equivalent to Akaike's criterion. Lower values of LCV indicate a better fitting model.

4.1. Effect of a boost of radiotherapy on the occurrence of fibrosis in patients with early breast cancer

The new approach for c-index has been computerized in a large multicenter trial from the European Organisation for Research and Treatment of Cancer (EORTC): the EORTC 22881-10882. In this trial, 5318 patients who received microscopically complete excision of a breast tumor and axillary dissection, followed by whole breast irradiation (WBI) of 50 Grays (Gy) in 5 weeks, were randomized between no extra irradiation and a boost dose of 16 Gy to the original tumor bed. In a separate stratum of the trial, 251 patients with a microscopically incomplete excision were randomized from 1989 to 1996 to receive a boost dose of 10 versus 26 Gy. Randomization occurred after surgery, and patients were stratified for age, menopausal status, presence of extensive ductal carcinoma in situ, clinical tumor size, nodal status, and institute where they received treatment. The final analysis of the trial showed that a 16-Gy boost improved the local control but increased the risk of moderate or severe fibrosis [25]. Following those results, risk factors of fibrosis subsequent to breast-conserving radiotherapy treatment with or without a boost of radiotherapy in patients with early breast cancer were studied in a prognostic study [26]. This analysis was conducted on a subset of patients ($n = 5178$) from the EORTC 22881-10882 trial with complete resection, no major deviations from eligibility criteria, no wrong randomized treatment received, and sufficient data regarding baseline characteristics.

Factors found to be associated with a higher risk of moderate or severe fibrosis in the boost arm were adjuvant tamoxifen, menopausal status, hematoma or edema after surgery, radiation quality, maximum irradiation dose, concomitant chemotherapy, irradiation boost technique, and energy of electron boost. In the no boost arm, prognostic factors of fibrosis were the maximum irradiation dose and the addition of concomitant chemotherapy. The concordance of these two prognostic models, calculated through a Harrell's c-index, was 0.66 in the development set (70% of the patients) and 0.62 in the validation set (30% of the patients) for the boost arm, and 0.65 and 0.59, respectively, in the no boost arm, when the data clustering was not taken into account. We repeated these analyses by adding a frailty on the institute of treatment, to see if the concordance, and therefore the prediction, can be improved by considering the heterogeneity between institutes. For the purpose of this application, we kept all patients in one dataset and kept only patients with complete information. In this overall population, the Harrell's c-index for the Cox model was 0.65 on the boost arm and 0.63 in the no boost arm.

We analyzed 4829 patients with complete information, 2424 in the boost arm and 2405 in the no boost arm. The median follow-up was 10 years. Patients have been randomized in nine countries and a total of 31 participating institutes. The principal source of heterogeneity is the radiotherapy treatment received, which is associated with the institute of treatment. Thus, the random effect was on the institutes. However, some institutes included very few patients, and we aggregate some institutes from the same country (this was performed for three institutes from Belgium, three from France, two from Netherlands, three from Germany, two from Israel, and three from Spain; but none from Switzerland, UK, and Australia). This results in 21 analyzed institutes, with 14 to 816 patients per institute (median=141). This range was 8 to 410 patients per institute in the boost arm (median=70) and 6 to 406 patients per institute in the no boost arm (median=73). We calculated the time to fibrosis from the day of randomization to the day moderate or severe fibrosis was first reported. We censored patients alive without moderate or severe

fibrosis at the last follow-up. We censored patients with salvage mastectomy or death before the occurrence of fibrosis at the time of first event (mastectomy or death). Overall, 997 patients (21%) developed fibrosis: 676 in the boost arm and 321 in the no boost arm, bringing the censoring rate close to 80%.

Table II presents the prognostic frailty model for each treatment arm. In the boost arm, menopausal status and treatment by tamoxifen, as well as radiation quality and maximum WBI dose, were no longer significantly associated with the risk of fibrosis. By comparison with the Cox model, radiation quality and administration of tamoxifen were not associated with the risk of fibrosis once the heterogeneity between institutes was taken into account. In the no boost arm, the estimated effect of WBI dose and chemotherapy was significantly associated with the risk of fibrosis but lower when the heterogeneity between institute was taken into account. For both arms, the variance of the frailty parameter was significantly different from zero (one-sided Wald test $= 0.52/0.16 = 3.25 > 1.64$ for boost arm and $= 0.61/0.20 = 3.05 > 1.64$ for no boost arm), meaning that heterogeneity in risk of fibrosis exists between the institutes.

Table III presents results of the different concordance measures. We set τ=13.6 years for the Uno's c-index in both arms, which corresponds to the time of the last observed event. The overall c-indexes were higher than the previously published ones, with a value of 0.749 in the boost arm and 0.740 in the no boost arm. Higher values for between-institute concordance than for within-institute ones indicate that taking into account the institute in the prediction may be of some interest. Figure 1 represents the value of the both concordance measures at the within level for each group. The vertical grey lines represent the probability of being censored before τ in each institute (that is, $1 - \mathbb{P}(C > \tau)$). The value of the Gönen and Heller's measure was quite stable over the institutes and does not seem to depend on the censoring rate nor on the number of events. Considering the Uno's value, there was a higher heterogeneity, and some outliers. These extreme values tended to correspond to the institutes with only few number of events. This was observed in both boost arm and no boost arm. These findings should be validated on an external population.

4.2. Effect of chemotherapy in head and neck cancer in the update of the meta-analysis of chemotherapy in head and neck cancer

Head and neck carcinomas (oral cavity, oropharynx, hypopharynx, and larynx) are frequent tumors for which surgery and/or radiotherapy are the standard locoregional treatments. In the absence of a large randomized trial, the most reliable way to evaluate chemotherapy effect is a meta-analysis based on updated individual patient data. The meta-analysis of chemotherapy in head and neck cancer (MACH-NC) compared locoregional treatment (radiotherapy and/or surgery) with locoregional treatment plus chemotherapy [27]. The update of this meta-analysis added trials performed between 1994 and 2000 to those performed between 1965 and 1993. It was based on 108 comparisons and 17,493 patients, with a median follow-up of 5.6 years. The results were in favor of chemotherapy, with a benefit seen on overall survival [27]. The benefit was higher with concomitant chemotherapy compared with induction or adjuvant chemotherapy. We chose to focus on this timing and on recent trials (1994–2000) for the purpose of our application. The heterogeneity between the underlying risks in trials and the heterogeneity of treatment effect between trials was previously investigated by using correlated random effects [28]. Heterogeneity in treatment effect was no longer significant when the concomitant trials were considered separately. However, previous research work did not investigate the impact of adding random effects in the model on the quality and accuracy of the prediction. Moreover, prediction in a meta-analysis framework is still a matter of research [29].

The patients were divided in three groups of trials, according to their country of randomization: Europe, USA, and others (Argentina, India, Pakistan, Malaysia, Bulgaria, and Turkey). To preclude the effect of the randomized treatment, we kept only patients from the control arm. We excluded one trial, having randomized in both Europe and USA, from this application. The Europe group was the group used to develop the prognostic model. We then validated the model on the USA and others populations. We analyzed 2541 patients (1652 Europe, 622 USA, and 267 others) from 29 trials (21, 5, and 3, respectively). A total of 1652 patients (66%) died during the follow-up (1110, 379, and 163, respectively), with a number of deaths per trial ranging from 4 to 146, a median of 47 deaths per trial. We focused on the prognosis of overall survival, which is defined as the time from randomization to death from any cause. We censored patients alive at the date of last contact.

We can do some conditional prediction on new groups in two fashions. The first one is to predict frailty term of the new group using the posterior distribution of the frailties, but given the regression parameters

Table II. Prognostic frailty model of the risk of fibrosis in the boost versus no boost trial (EORTC 22881-10882) in each arm.

Arm	Factors	Shared frailty model Hazard ratio	95% CI	*p*-value	Cox model Hazard ratio	95% CI	*p*-value
Boost	Age (by 10 years)	1.13	[1.00–1.28]	0.06	1.16	[0.98–1.36]	0.08
($N = 2424$)	Tamoxifen						
	Post-menopausal versus premenopausal	0.83	[0.64–1.09]	0.19	0.77	[0.56–1.07]	0.12
	Tamoxifen versus no tamoxifen	1.20	[0.97–1.48]	0.08	1.39	[1.14–1.71]	0.001
	Hematoma	1.67	[1.35–2.06]	< 0.0001	1.85	[1.51–2.27]	< 0.0001
	Edema	1.61	[1.14–2.29]	0.007	1.79	[1.27–2.51]	0.0008
	Radiation quality			0.32			< 0.0001
	$Cobalt^{60}$ versus X-ray=6 MV	1.17	[0.89–1.54]		1.14	[0.91–1.43]	
	X-ray < 6 versus X-ray=6 MV	1.11	[0.80–1.53]		1.64	[1.32–2.03]	
	X-ray > 6 versus X-ray=6 MV	1.28	[0.96–1.71]		0.63	[0.47–0.84]	
	Maximum WBI dose	1.03	[0.99–1.08]	0.12	1.12	[0.98–1.28]	0.08
	Chemotherapy during radiotherapy	2.09	[1.52–2.88]	< 0.0001	2.15	[1.58–2.93]	< 0.0001
	Irradiation technique boost			< 0.0001			0.0001
	$Cobalt^{60}$ versus X-ray	0.59	[0.40–0.87]		0.91	[0.64–1.29]	
	Electron beam versus X-ray	0.37	[0.23–0.59]		0.38	[0.24–0.60]	
	Interstitial versus X-ray	1.14	[0.79–1.66]		1.03	[0.74–1.43]	
	If electron beam, energy (MV)	1.07	[1.03–1.10]	0.0007	1.07	[1.04–1.11]	< 0.0001
	Frailty term	$\theta = 0.52$	(se=0.16)				
	LCV	2.67			2.72		
No boost	Age (by 10 years)	1.04	[0.92–1.18]	0.49	1.06	[0.94–1.19]	0.37
($N = 2405$)	Maximum WBI dose	1.13	[1.01–1.26]	0.04	1.22	[1.20–1.25]	< 0.0001
	Chemotherapy during radiotherapy	1.60	[1.01–2.51]	0.04	2.47	[1.65–3.70]	< 0.0001
	Frailty term	$\theta = 0.61$	(se=0.20)				
	LCV	1.37			1.39		

CI, confidence interval; WBI, whole breast irradiation; MV, mega volt; θ, variance of the frailty parameter; se, standard error of the variance parameter; LCV, likelihood cross-validation criterion.

Table III. Concordance values at the between, within, and overall levels, for the prognostic frailty models of the boost versus no boost trial (EORTC 22881-10882) in each randomization arm.

Model	Gönen and Heller's $\hat{K}(\hat{\beta})$			Uno's $\hat{C}^*_{13}$		
	Between	Within	Overall	Between	Within	Overall
Boost	0.736	0.590	0.724	0.760	0.619	0.749
(standard error)	(0.031)	(0.002)	(0.028)	(0.034)	(0.020)	(0.032)
No boost	0.720	0.562	0.707	0.759	0.525	0.740
(standard error)	(0.030)	(0.003)	(0.028)	(0.033)	(0.030)	(0.030)

*$\tau = 13.6$ years, corresponding to the time of the last observed event.

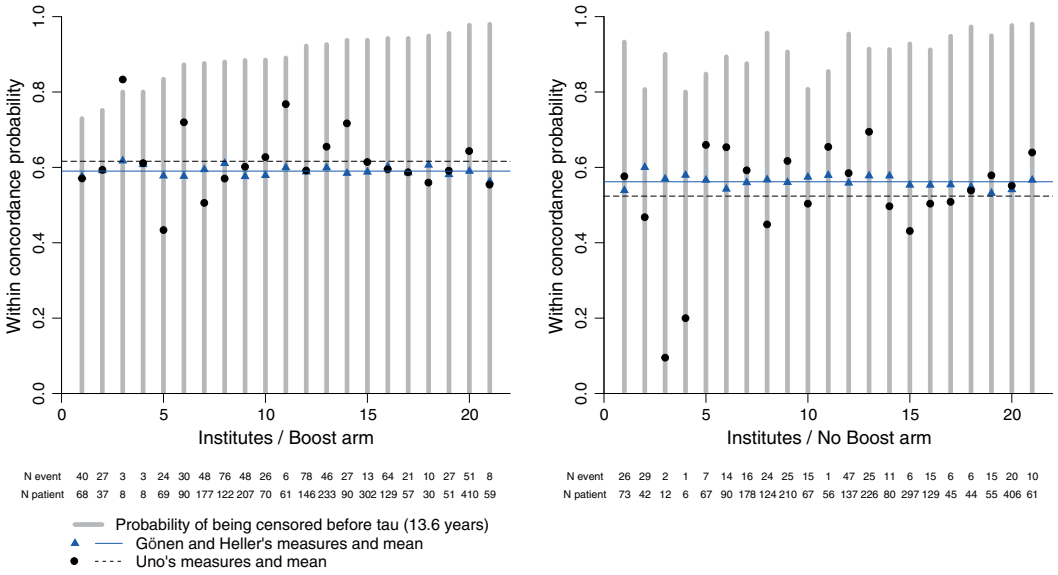

Figure 1. Values of the Gönen and Heller's and Uno's within-group concordance probabilities by institute in the boost arm and no boost arm in the EORTC 22881-10882 trial, and probability of being censored before τ (vertical grey lines).

and the baseline hazard function estimated from the development population. More precisely, on the development population, the following parameters are estimated by the model:

- the covariate effect vector $\hat{\beta}_{\text{development}}$
- the estimated variance of the random effects $\hat{\theta}_{\text{development}}$
- the cumulative baseline hazard function $\hat{H}_{0,\text{development}}(t)$

On the basis of this estimation, it is possible to predict the value of the frailty terms extending the equation (1). Let l denotes a patient of a new group k ($l = 1,\ldots,n_k$), m_k the observed number of deaths in the group k, and t_{kl} the times of events. We then have $H_{kl}(t_{kl}) = \hat{H}_{0,\text{development}}(t_{kl})\exp(\hat{\beta}'_{\text{development}}x_{kl})$ and deduce the predicted $\hat{u}_{k,\text{validation}}$:

$$\hat{u}_{k,\text{validation}} = \frac{\left(1/\hat{\theta}_{\text{development}} + m_k\right)}{\left(1/\hat{\theta}_{\text{development}} + \sum_{l=1}^{n_k} H_{kl}(t_{kl})\right)} \tag{2}$$

The second method is to consider that all the patients had a frailty equals to the posterior mean of the frailties estimated from the development population. That is,

$$\hat{v}_{k,\text{validation}} = \frac{1}{G}\sum_{i=1}^{G}\hat{u}_{i,\text{development}} \tag{3}$$

for all group i from the development population. For a patient from the validation population, the linear predictor was then defined by $\hat{\beta}_{\text{development}} x_{kl} + \ln(\hat{u}_{k,\text{validation}})$ or by $\hat{\beta}_{\text{development}} x_{kl} + \ln(\hat{v}_{k,\text{validation}})$. To make prediction for a patient from a development group (i), we directly use the predicted frailty $\hat{u}_i$ (as in equation (1)).

Table IV presents the results of a Cox model and a shared frailty model, with a random effect on the trial, developed on the Europe population. Parameter estimation was similar between the two models. The variance of the frailty term (θ) was significantly different from zero, indicating that there was a significant heterogeneity between the trials. Considering the discrimination ability, we observed some differences. As expected, concordance values were very similar at the within-trial level, for both measures. The between-trial measures were higher with the frailty model than with the Cox model, showing the interest of taking the frailty associated to the trial into account.

With both Gönen and Heller's and Uno's measures, the between and overall measures indicate a better discriminatory ability for the shared frailty model. The model is also able to discriminate

Table IV. Cox and frailty prognostic models for the overall survival in the control arm from the meta-analysis of chemotherapy in head and neck cancer ($N = 1652$) and corresponding concordance values (standard error).

	Frailty model			Cox model		
Factors	Hazard ratio	95% CI	p-value	Hazard ratio	95% CI	p-value
Sex (female versus male)	0.72	[0.60–0.85]	0.0002	0.72	[0.61–0.86]	0.0002
Age			0.03			0.02
51–60 versus $\leqslant 50$	1.21	[1.04–1.40]		1.22	[1.05–1.41]	
$\geqslant 61$ versus $\leqslant 50$	1.16	[1.00–1.35]		1.19	[1.02–1.38]	
Stage			< 0.0001			< 0.0001
III versus I+II	1.24	[0.94–1.64]		1.22	[0.93–1.60]	
IV versus I+II	1.90	[1.46–2.46]		1.85	[1.45–2.37]	
Site (larynx versus other)	0.66	[0.55–0.80]	< 0.0001	0.63	[0.52–0.76]	< 0.0001
LCV	1.42			1.43		
Frailty term	$\theta = 0.08$	(se $= 0.04$)				
Concordance values	Between	Within	Overall	Between	Within	Overall
On development population (Europe)						
Gönen and Heller's $\hat{K}(\hat{\beta})$	0.616 (0.006)	0.578 (0.005)	0.613 (0.006)	0.589 (0.008)	0.579 (0.005)	0.589 (0.007)
Uno's $\hat{C}^*_9$	0.638 (0.011)	0.615 (0.019)	0.636 (0.011)	0.600 (0.013)	0.620 (0.021)	0.602 (0.012)
Validation on USA population						
G&H - predicted $\hat{u}_{k,\text{validation}}$	0.644 (0.027)	0.570 (0.005)	0.625 (0.020)	0.616 (0.022)	0.571 (0.006)	0.604 (0.017)
Uno's - predicted $\hat{u}_{k,\text{validation}}$	0.624 (0.044)	0.488 (0.022)	0.589 (0.035)	0.577 (0.046)	0.486 (0.022)	0.553 (0.038)
G&H - mean $\hat{v}_{k,\text{validation}}$	0.612 (0.021)	0.570 (0.005)	0.601 (0.016)			
Uno's - mean $\hat{v}_{k,\text{validation}}$	0.579 (0.045)	0.488 (0.023)	0.555 (0.038)			
Validation on 'others' population						
G&H - predicted $\hat{u}_{k,\text{validation}}$	0.590 (0.011)	0.569 (0.011)	0.575 (0.010)	0.579 (0.007)	0.571 (0.009)	0.573 (0.008)
Uno's - predicted $\hat{u}_{k,\text{validation}}$	0.614 (0.053)	0.559 (0.028)	0.574 (0.039)	0.590 (0.026)	0.558 (0.027)	0.567 (0.023)
G&H - mean $\hat{v}_{k,\text{validation}}$	0.579 (0.008)	0.569 (0.010)	0.572 (0.009)			
Uno's - mean $\hat{v}_{k,\text{validation}}$	0.594 (0.028)	0.559 (0.028)	0.568 (0.023)			

CI, confidence interval; θ, variance of the frailty parameter; se, standard error of the variance parameter; LCV, likelihood cross-validation criterion; G&H, Gönen and Heller.
*$\tau = 9.4$ years, corresponding to the time of the last observed event.

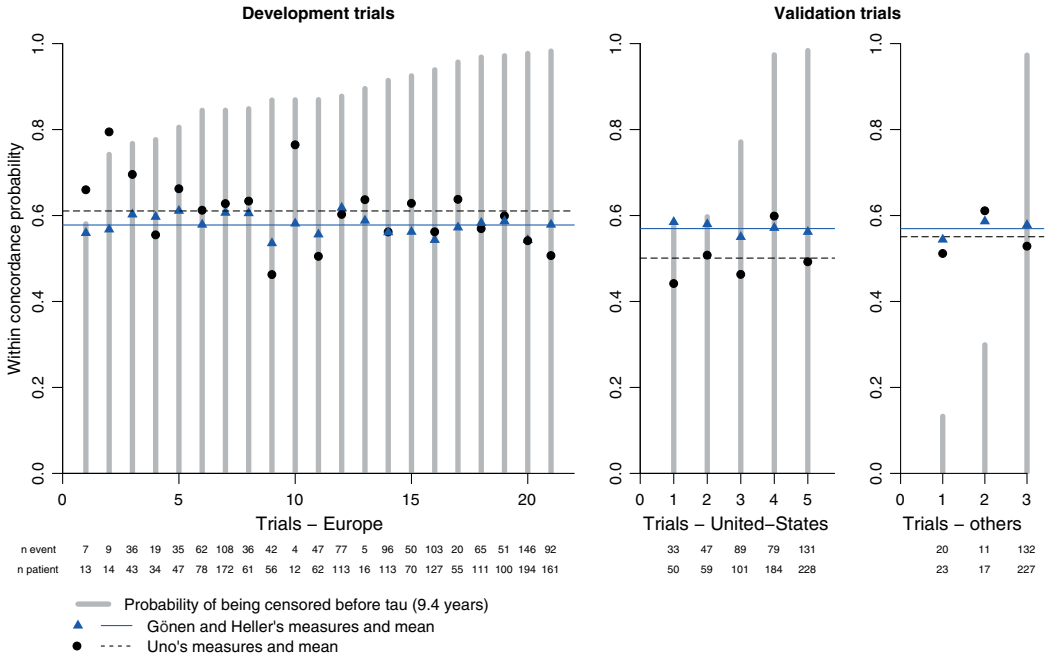

Figure 2. Values of the Gönen and Heller's and Uno's within-group concordance probabilities by trial in the Europe, USA, and others trials from the meta-analysis of chemotherapy in head and neck cancer, and probability of being censored before τ (vertical grey lines).

survival of patients in new population. The concordance values on the USA population was similar than on Europe. However, as expected, concordance values were lower on the others population (57% with both measures). The results of the between-trial concordance were higher when the frailty terms were newly predicted on new groups ($\hat{u}_k$), instead of considering a mean value for all groups ($\hat{v}_k$). The within-group concordance was not impacted by the choice of the frailty used, as the frailty terms are not taking into account. As illustrated on Figure 2, there was more heterogeneity in the within-group Uno's measure than in the Gönen and Heller's one, in both development and validation level. This was also seen with the higher standard errors for the Uno's measure in Table IV. Between the populations also, the average Uno's measure was quite different, whereas the Gönen and Heller's one was stable. Considering these results, we can conclude that the proposed model, using the information about sex, age, stage, and tumor site, was able to well classify the patients regarding their survival times in 60% of cases. This suggests that other factors may influence the prognostic of patients with head-and-neck cancer. External validation results suggest that this model developed on Europe patients may be applied on a USA population. We can also apply it on a population from other countries, if we accept a little loss of discrimination ability. This external validation confirms that the model performs as expected in new but similar patients [30].

5. Discussion

The extension of concordance measures to the frailty models framework is of interest in cancer prognostic studies. The approach was previously adopted for the well-known Harrell's c-index by Van Oirbeek and Lesaffre [13]. We propose here to extend two other measures that do not depend on the censoring distribution in the study population. The chosen measures are two different types of concordance. One is a nonparametric measure, correction of the Harrell's c-index [10]. The other is a model-based measure, which supposes that the risks are proportional and that the model holds for all times [12]. The principal difference between these two measures is that the first can be applied to all prognostic models, whereas the second considers all the pairs of patients available and does not exclude pairs of patients because of presence of censoring.

The extension of concordance measures to frailty models makes it possible to take into account different levels of information: first the between-group information, then within-group information, and finally overall information. Application results suggest that concordance values may vary from one

group to another, especially using Uno's measure. Values from small groups, which are more prone to be extreme, can greatly affect the average concordance measure. The question of the predictive ability of a prognostic index among groups was previously discussed, and the impact of such variation on the generalizability of the prognostic index was investigated [31]. As compared with the Van Oirbeek and Lesaffre approach [13], we chose to focus on the conditional concordance. The conditional approach takes into account all the available information including the observed covariates effect and the frailty estimation. This leads to more accurate prediction. To make conditional prediction on patients from new groups, it is necessary to predict value of the frailty for this new group. For that, there are two possibilities: using frailty prediction (equation(2)) or using a mean value estimated on the development population (equation(3)). The second method gives less accurate estimation, as it gives all groups the same frailty. This is equivalent to use some marginal prediction. Making conditional prediction on new groups illustrates the ability of the proposed method to externally validate the prediction model.

Simulations show that for a large cluster size ($n_i \geqslant 40$), the influence of the censoring rate on the concordance measures proposed is limited, with a low bias even at 70% of censoring and whatever the number of groups. However, when the information is restricted by the group size (n_i=2 or 10 patients per group), the censoring rate influences the estimation of the concordance. This suggests that the performance of these concordance measures is maintained in the framework of frailty models provided that the group size is sufficient. However, frailty models not only handle grouped data but also handle repeated data. The feature of repeated data, such as recurrent events observed several times in a given patient, is few observations in a large number of patients. In this case, frailty is shared between events in a patient. In our simulation, the setting with $G = 40$ and $n_i = 2$ was similar to this type of data, and the performance of the concordance estimator was poor. This raises the question of the accuracy of such measures for recurrent event models but also the extent of its usefulness. For instance, how can one interpret a within-patient concordance in which times to different events of a patient are compared, while ignoring the succession of these events? This question has to be addressed in further research.

When computing Uno's c-index, a choice has to be made of setting the value of τ. As recently suggested by Stare *et al*. [32], we set τ at the time of the last observed event. We also performed sensitivity analyses where τ equals the median of the censoring distribution (10 years for the EORTC trial application and 6 years for the MACH-NC application). Concordance values were slightly lower in the EORTC trial (between-country concordances equal 0.743 and 0.745 for the boost and no boost arm, respectively, versus 0.760 and 0.759, respectively). In the MACH-NC application, results were very similar. This is coherent with the results obtained by Uno *et al*. where a small-to-moderate impact of the τ on the concordance values was seen [10]. We do not need such a parameter to compute Gönen and Heller's concordance, as observed events are not directly used to compute the concordance but are used only in the estimation of the model. In this situation, the censoring distribution has little impact on the estimations. Moreover, we only focus here on independent censoring. It is possible to consider covariate-dependent censoring when computing the c-index. Such dependence can be accounted for by replacing the Kaplan–Meier estimation of the censoring distribution by an adequate modeling. This was, for example, investigated recently in [33]. This is easily possible in the presence of frailty. However, this means to add a semiparametric estimation in the measure computation. This question has to be addressed specifically for each application, and the presence of clustered data and use of a shared frailty model do not prevent from dependent censoring.

Concordance measures encounter the problem of ties on prediction. Tied pairs account for 0.5 in the concordance measure. Thus, the more ties there are on prediction, the worse is the concordance. This is coherent with the fact that a model leading to many ties is not able to discriminate patients. We can improve this lack of discrimination by the use of frailty in the prognostic model, which diminishes the number of ties and improves prediction. Another proposition to handle ties is to exclude them [34]. We performed sensitivity analyses for our meta-analysis application, excluding pairs of patients tied on prediction (almost no ties were seen on the multicenter trial application, because of the presence of the continuous variable age). This resulted in a slight improvement in the within-group Gönen and Heller's concordance (0.589 versus 0.578 in shared frailty model). The between-group measure remained unchanged, as all patients compared had different frailties. Consequently, we only slightly modified the overall measure by the exclusion of ties (0.614 versus 0.613).

The types of grouped data that are mainly found in cancer studies are meta-analyses and multicenter clinical trials. These two types of studies share a large sample size and may be international. However, meta-analyses may have more heterogeneity owing to different protocols and to the presence of different study times. This issue of time is important, especially for prediction making, and can be considered

by adjusting on period of randomization. A third type of grouped data is kinship, to study for example genetic factors. The feature of such data is small group sizes, for which the proposed concordance measures seem less accurate.

This paper does not include the possibility of stratified frailty models. The presence of stratification modifies the concordance formulation. Indeed, in the current approach, when two patients were compared, we considered that the baseline hazards function was the same for both. However, this is not the case when the model is stratified. Moreover, we assume that we can consider information at a given time t, whatever t is. This may not no longer be exact with different baseline hazards. Intuitively, the first solution that we suggest is to compute separate concordance for each stratum. Thus, we compare only patients in the same stratum, that is, having the same baseline hazards.

Finally, Gönen and Heller studied the asymptotic distribution of their estimator, which enables them to propose an analytic variance [12]. Stare *et al*. [35] proposed a similar development for computing variance, stating that a bootstrap variance estimate may also be used. The extension of this variance for between-group and within-group estimates including the variance of the frailty estimation is a matter for further development.

Acknowledgements

We thank the Institut National du Cancer for funding this research. We thank the European Organisation for Research and Treatment of Cancer for sharing the data of the boost versus no boost trial and especially Laurence Collette for her insightful comments. We thank the meta-analysis of chemotherapy in head and neck cancer (MACH-NC) trialists who agreed to share and update their data and the following institutions for funding the investigators' meeting or the meta-analysis project: Association pour la Recherche sur le Cancer, Programme Hospitalier de Recherche Clinique, Ligue Nationale Contre le Cancer, and Sanofi-Aventis (unrestricted grants).

References

1. Moons K, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how?. *BMJ* 2009; **338**(1):b375–b375.
2. Steyerberg E, Vickers A, Cook N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; **21**(1):128–138.
3. Korn E, Simon R. Measures of explained variation for survival data. *Statistics in Medicine* 1990; **9**(5):487–503 (en).
4. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *British Journal of Cancer* 1994; **69**(6):979–985.
5. Pencina M, D' Agostino Sr R, D' Agostino Jr R, Vasan R. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine* 2008; **27**(2):157–172 (en).
6. Duchateau L, Janssen P. *The Frailty Model*. Springer: New York, 2008.
7. Wienke A. Frailty models. *MPIDR Working Paper WP 2003-032, Max Planck Institute for Demographic Research, Rostock Germany* 2003.
8. Cook R, Lawless J. *The Statistical Analysis of Recurrent Events*. Springer: New York, 2007 (en).
9. Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; **15**(4):361–387.
10. Uno H, Cai T, Pencina M, D'Agostino R, Wei L. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* 2011; **30**(10):1105–1117 (en).
11. Pencina MJ, D'Agostino Sr. RB, Song L. Quantifying discrimination of framingham risk functions with different survival c statistics. *Statistics in Medicine* 2012; **31**(15):1543–1553 (en).
12. Gonen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 2005; **92**(4):965–970.
13. Van Oirbeek R, Lesaffre E. An application of harrell's c-index to PH frailty models. *Statistics in Medicine* 2010; **29**(30):3160–3171 (en).
14. Rondeau V, Commenges D, Joly P. Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime Data Analysis* 2003; **9**(2):139–153.
15. Ripatti S, Palmgren J. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* 2000; **56**(4):1016–1022 (en).
16. Box-Steffensmeier J, De Boef S. Repeated events survival models: the conditional frailty model. *Statistics in Medicine* 2006; **25**(20):3518–3533 (en).
17. Ramsay JO. Monotone regression splines in action. *Statistical Science* 1988; **3**(4):425–441.
18. Rondeau V. Statistical models for recurrent events and death: application to cancer events. *Mathematical and Computer Modelling* 2010; **52**(7-8):949–955.
19. Louis T. Using empirical bayes methods in biopharmaceutical research. *Statistics in Medicine* 1991; **10**(6):811–829.
20. Rondeau V, Mazroui Y, Gonzalez J. FRAILTYPACK: an R package for the analysis of correlated survival data with frailty models using the penalized likelihood estimation. *Journal of Statistical Software*; **In press**.

21. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA: The Journal of the American Medical Association* 1982; **247**(18):2543–2546.
22. Field C, Welsh A. Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2007; **69**(3):369–390 (en).
23. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* 2005; **24**(11):1713–1723 (en).
24. Rondeau V, Pignon JP, Michiels S. A joint model for the dependence between clustered times to tumour progression and deaths: a meta-analysis of chemotherapy in head and neck cancer. *Statistical Methods in Medical Research* 2011. [In press].
25. Bartelink H, Horiot JC, Poortmans P, et al. Impact of a higher radiation dose on local control and survival in breast-conserving therapy of early breast cancer: 10-year results of the randomized boost versus no boost EORTC 22881-10882 trial. *Journal of Clinical Oncology* 2007; **25**(22):3259–3265.
26. Collette S, Collette L, Budiharto T, et al. Predictors of the risk of fibrosis at 10 years after breast conserving therapy for early breast cancer-a study based on the EORTC trial 22881-10882 'boost versus no boost'. *European Journal of Cancer* 2008; **44**(17):2587–2599.
27. Pignon JP, Le Maitre A, Maillard E, et al. Meta-analysis of chemotherapy in head and neck cancer (MACH-NC): an update on 93 randomised trials and 17,346 patients. *Radiotherapy and Oncology* 2009; **92**(1):4–14.
28. Rondeau V, Michiels S, Liquet B, Pignon JP. Investigating trial and treatment heterogeneity in an individual patient data meta-analysis of survival data by means of the penalized maximum likelihood approach. *Statistics in Medicine* 2008; **27**(11):1894–1910.
29. Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statistics in Medicine* 2013 (en).
30. Altman DG, Royston P. What do we mean by validating a prognostic model?. *Statistics in Medicine* 2000; **19**(4):453–473.
31. Legrand C, Duchateau L, Janssen P, Ducrocq V, Sylvester R. Validation of prognostic indices using the frailty model. *Lifetime Data Analysis* March 2009; **15**(1):59–78 (en).
32. Stare J, Maucort-Boulch D, Kejzar N. On using simulations to study explained variation in survival analysis. *33rd annual conference of the International Society for Clinical Biostatistics*. Abstract book: pages 60 (C22.5). 19-23 August 2012, Bergen, Norway [Oral presentation].
33. Gerds TA, Kattan MW, Schumacher M, Yu C. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine* 2012 (en).
34. Yan G, Greene T. Investigating the effects of ties on measures of concordance. *Statistics in Medicine* 2008; **27**(21):4190–4206.
35. Stare J, Perme M, Henderson R. A measure of explained variation for event history data. *Biometrics* 2011; **67**(3):750–759 (en).

3.3 Additional remarks

The steps of calculation of the Gönen and Heller measure in a frailty framework are detailed in appendix B.

3.3.1 On the use of an analytic variance

Derive from the U-statistics theory, the concordance value is associated to an analytical variance estimation. In their paper, Gönen and Heller (2005) derive this variance. In the frailty model framework, we found the thing complicated by the consideration of the frailty variance. Indeed, Gönen and Heller proposed the following development. Knowing that the first order Taylor series expansion of the concordance estimator is given by

$$K_n(\hat{\beta}) = K_n(\beta_0) + \left\{\frac{\partial K_n(\beta)}{\partial \beta}\right\}'\Big|_{\beta=\beta_0} (\hat{\beta} - \beta_0) + o_p(1)$$

We have

$$var\{K_n(\hat{\beta})\} = var\{K_n(\beta_0)\} + \left\{\frac{\partial K_n(\beta)}{\partial \beta}\right\}'\Big|_{\beta=\beta_0} var(\hat{\beta}) \left\{\frac{\partial K_n(\beta)}{\partial \beta}\right\}\Big|_{\beta=\beta_0}$$

Where

$$var\{K_n(\beta_0)\} = \frac{4}{\{n(n-1)\}^2} \sum_j \sum_{j'} \sum_{k \neq j'} \{v_{j'j} + v_{jj'} - K_n(\hat{\beta})\}\{v_{kj} + v_{jk} - K_n(\hat{\beta})\}$$

with $v_{j'j} = \frac{I[(\hat{\beta}'(x_{j'} - x_j)) \leq 0]}{1 + e^{\hat{\beta}'(x_{j'} - x_j)}}$

Considering the extension to frailty models, we have

$$K_n(\hat{\beta}, \hat{u}) = K_n(\beta_0, u_0) + \left\{\frac{\partial K_n(\beta, u)}{\partial \beta}\right\}'\Big|_{\beta=\beta_0, u=u_0} (\hat{\beta} - \beta_0) + \left\{\frac{\partial K_n(\beta, u)}{\partial u}\right\}'\Big|_{\beta=\beta_0, u=u_0} (\hat{u} - u_0) + o_p(1)$$

Assuming β and u independent, we obtain

$$\begin{aligned} var\{K_n(\hat{\beta}, \hat{u})\} = var\{K_n(\beta_0, u_0)\} &+ \left\{\frac{\partial K_n(\beta, u)}{\partial \beta}\right\}'\Big|_{\beta=\beta_0, u=u_0} var(\hat{\beta}) \left\{\frac{\partial K_n(\beta, u)}{\partial \beta}\right\}\Big|_{\beta=\beta_0, u=u_0} \\ &+ \left\{\frac{\partial K_n(\beta, u)}{\partial u}\right\}'\Big|_{\beta=\beta_0, u=u_0} var(\hat{u}) \left\{\frac{\partial K_n(\beta, u)}{\partial u}\right\}\Big|_{\beta=\beta_0, u=u_0} \end{aligned}$$

However, due to the relatively high value of the variance of the frailty effect, the overall resulting variance was really high in the application that we made. This development requires some other theoretical statistics knowledge. Therefore, we consider that it should be matter of further research, that we think is beyond the scope of this thesis.

3.3.2 On the interpretation of the concordance and the overall measure

Following our publication, in their letter to the editor, van Klaveren et al. (2014) discussed the interpretation of concordance measures in clustered data, and especially insisted on the utility of the within-group measure, to the detriment of the between-group and overall measures. We agree that more importance can be given on within-group concordance when we are interested in decision at a group level, for example patient care in an hospital. However, we insist on the fact that there is an interest to compute both within-group and between-group measures, as comparing their values can give us some information about the between-group heterogeneity. If the between-group concordance is greatly higher than the within-group one in a study, then we should have a thought on why such difference in the prognosis of patient is seen among centres. There are probably some important prognostic factors associated to the group that are not measured. This question is, to my mind, as important as evaluating the measured clinical prognostic factors, and is crucial to fully understand and determine the best care for the patients. Moreover, the within-group concordance is a mean, and this measure can vary from one group to another as investigated in our article. Variation of the hazard ratio of the prognostic index has also been investigated in Legrand et al. (2009). Finally, the way of combining the within-cluster concordance was subsequently further investigated in an article by van Klaveren et al. (2014). Both the letter (van Klaveren et al., 2014) and our response (Mauguen et al., 2014) are given in Appendix C.

Finally, similarly to thoughts about the best way of combining different within-group values into one measure, we may improve the averaging of within- and between-group measures to get the overall measure. We use the proposition of Van Oirbeek and Lesaffre (2010) of a weighted average, the weights being based on the number of used pairs. For example, for the Uno's index, we used: (see the article for the notations)

$$\hat{C}_{\tau,O} = \frac{n_{comp,W}}{n_{comp,T}}\hat{C}_{\tau,W} + \frac{n_{comp,B}}{n_{comp,T}}\hat{C}_{\tau,B}$$

Another approach could be to use weights similar to those proposed by Uno, based on the censoring probability. That is to replace the counts of usable pairs by the denominator of the Uno's measures (see pages 4806 and 4807 of the previous article; Δ_{ij} is the event indicator and $\hat{T}_c(.)$ is the Kaplan-Meier estimator of the censoring times):

- $n_{comp,W}$ by $\sum_{i=1}^{G} \sum_{j=1}^{n_i} \sum_{j'=1}^{n_i} \Delta_{ij}\{\hat{T}_c(T_{ij})\}^{-2} I(T_{ij} < T_{ij'}, T_{ij} < \tau)$

- $n_{comp,B}$ by $\sum_{i=1}^{G}\sum_{j=1}^{n_i}\left\{\sum_{\substack{i'=1\\i'\neq i}}^{G}\sum_{j'=1}^{n_{i'}}\Delta_{ij}\{\hat{T}_c(T_{ij})\}^{-2}I(T_{ij}<T_{i'j'},T_{ij}<\tau)\right\}$
- $n_{comp,T}$ by the sum of the two previous quantities

3.3.3 On the use of randomized clinical trial data to develop prognostic index

The inclusion of treatment in prognosis can be a matter of debate. If one of the main objectives of predicting patients' risk is the choice of the treatment, it seems odd to consider the treatment to develop the prognostic index. However, data from RCTs are of high quality, considering data collection and patients' follow-up. Therefore, they are of great interest in prognostic research. When using them, the treatment status should not be ignored when developing the prognosis index. It should be included if given before the starting point of the prognosis study. In the series of papers *Prognosis and prognostic research*, Moons et al. (2009); Royston et al. (2009); Altman et al. (2009); Moons et al. (2009) discussed that point. They concluded that data from RCTs can be used as long as treatment is included, if effective. There is also the possibility to keep only patients from the control arm, if enough data are available (meta-analysis context, for example). As an example, the PREDICT model for operated invasive breast cancer includes adjuvant therapy (chemotherapy, endocrine therapy) in the prognostic factors (Wishart et al., 2010).

3.3.4 On the value of the concordance in cancer research

In our applications, it appeared that the estimated concordance values can be seen as rather low. The within-group concordance value barely exceeded 0.50 in both applications. This issue has been discussed in the letter by Van Zee et al. (2011). They comment that the most well-known prognostic indices such as the *Gail model*, *Adjuvant!* and *Oncotype Dx* show concordance values below 0.70. This is coherent with the results we found in our applications. This statement raises two hypotheses: either the concordance is not a suitable indicator of the prognostic ability of cancer prognosis models, or the currently proposed prognosis models are not good enough to consider the complexity of this disease. Considering the first hypothesis, it is still matter of research to determine which values of concordance, or which difference in concordance, may be considered as sufficient to claim that a model has a good discrimination. Moreover, one of the known disadvantage of these measures is that they are not much sensible to the addition of prognostic factors (see Harrell (2001) or Pencina et al. (2008) for example). Considering

the second hypothesis, we believe that cancer is a progressive disease, and that a dynamic prediction accounting for its evolution could lead to more accurate predictions of the risk of death. We therefore thought about developing prediction models that can account for the cancer intermediate events, which are key information for the prognosis. This is the object of the chapter 4.

Chapter 4

Individual prediction of the risk of death after cancer relapses: Development and validation using joint modelling

4.1 Question and data

As discussed in the previous chapter, the existing prognosis models in cancer all have a rather small concordance index value, meaning limited discrimination ability. Thus, there is a need to propose some new prognostic models to try to obtain more accurate predictions. One possible way to improve prediction in this framework is to have some dynamic predictions that account for the cancer events. Indeed, cancer are often characterised by disease events. Patients may undergo loco-regional relapses or distant metastases which reflect an evolution of the disease. This evolution may result in a change in the risk of death, which must be updated. Basically, as illustrated on Figure 4.1, we want to predict the survival probability between the prediction time t and the prediction horizon $t + w$, accounting for the relapses occurring before t.

This chapter presents the dynamic probabilities calculation, as well as the development and the external validation of such prediction model. It was illustrated on breast cancer data. The development phase was done on a French hospital series (section 4.2), and the validation on two population-based datasets (section 4.3). The comparison between the three datasets is done in the validation article (see section 4.3.1) and other

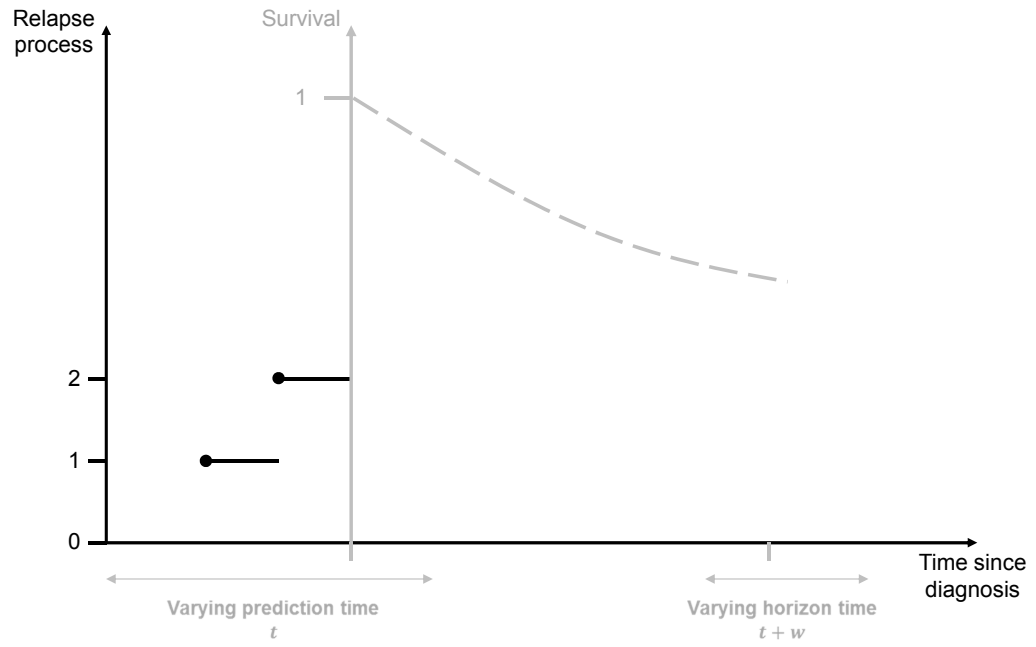

Figure 4.1: Illustration of the dynamic prediction. The grey part is the survival part to be predicted. The black part are the observed relapses to be used in prediction.

details can be found in the validation section 4.3.

The development part has been published in Statistics in Medicine (Mauguen et al., 2013) and probabilities of death from a joint model can now be conveniently calculated using the R package *frailtypack* (see chapter 6). The validation part is submitted for publication. Detailed calculation of the conditional probabilities that are not in the article are given in Appendix D.

4.2 Development step

4.2.1 Publication

Statistics
in Medicine

Special Issue Paper

Received 30 November 2012, Accepted 27 August 2013 Published online 13 September 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5980

Dynamic prediction of risk of death using history of cancer recurrences in joint frailty models

Audrey Mauguen,[a,b*†] Bernard Rachet,[c] Simone Mathoulin-Pélissier,[a,d,e] Gaetan MacGrogan,[d] Alexandre Laurent[a,b] and Virginie Rondeau[a,b]

Evaluating the prognosis of patients according to their demographic, biological, or disease characteristics is a major issue, as it may be used for guiding treatment decisions. In cancer studies, typically, more than one endpoint can be observed before death. Patients may undergo several types of events, such as local recurrences and distant metastases, with death as the terminal event. Accuracy of clinical decisions may be improved when the history of these different events is considered. Thus, it may be useful to dynamically predict patients' risk of death using recurrence history. As previously applied within the framework of joint models for longitudinal and time to event data, we propose a dynamic prediction tool based on joint frailty models. Joint modeling accounts for the dependence between recurrent events and death, by the introduction of a random effect shared by the two processes. We estimate the probability of death between the prediction time t and a horizon $t+w$, conditional on information available at time t. Prediction can be updated with the occurrence of a new event. We proposed and compared three prediction settings, taking into account three different information levels. The proposed tools are applied to patients diagnosed with a primary invasive breast cancer and treated with breast-conserving surgery, followed for more than 10 years in a French comprehensive cancer center. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: prediction; joint model; recurrence history; frailty; cancer

1. Introduction

Predicting risk of death is of great interest for patient care and medical choices. Indeed, knowing patients' prognosis may help determine their treatments. Particularly, interest lies in predicting the probability of death in a specific time window, given the history of the patient before the time of prediction t. Traditionally, history of the patient before t is a synthesis of available information at time t. In this case, the previous dynamic of the disease is forgotten, leading to a loss of information. Evolution of external covariates in the sense of Kalbfleisch and Prentice [1], such as treatment received or environmental exposures, can be considered simply by using time-dependent covariates in a survival model such as the Cox model. However, evolution of internal covariates, whose future values depend on the event of interest, for example, repeated measurements taken on a studied subject needs specific modeling. Prediction of death taking into account not only the last known value of a biomarker but its whole trajectory before t was recently proposed on the basis of joint modeling of longitudinal data and survival time [2, 3]. However, in some diseases such as cancer, disease relapses may also have an impact on the instantaneous risk of

[a]*INSERM, ISPED, Centre INSERM U897-Epidémiologie-Biostatistique, F-33000 Bordeaux, France*
[b]*Univ. Bordeaux, ISPED, Centre INSERM U897-Epidémiologie-Biostatistique, F-33000 Bordeaux, France*
[c]*Cancer Research UK Cancer Survival Group, Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, UK-WC1E7HT London, U.K.*
[d]*Unité de recherche et d'épidemiologie cliniques, Institut Bergonié, F-33000 Bordeaux, France*
[e]*INSERM CIC-EC7, F-33000 Bordeaux, France*
**Correspondence to: Audrey Mauguen, INSERM U897 - Equipe de biostatistique, ISPED, Université Bordeaux Segalen, 146 rue Leo Saignat, 33076 Bordeaux Cedex, France.*
[†]*E-mail: audrey.mauguen@isped.u-bordeaux2.fr*

death. Particularly, the number of recurrences and the time of occurrences may be a predictor of the risk of death.

Recurrent events may be observed in many clinical studies, as well as in industrial or social research [4]. In the presence of such data of interest, two censoring types may occur: an independent censoring, which does not prevent further events from occurring; or a terminal event, which stops the recurrent event process. Interest has increased over these last years for methods studying this kind of data, with a growing literature. Specific survival joint models were developed to take a terminal event into account and to quantify the correlation between, on one hand, the successive recurrent events of a patient and the recurrent and terminal events, on the other hand, through a shared random effect [4–8]. More recent extensions include an additive model [9], possibility of time-dependent covariates [10], non-linear effect of the covariates [11], multivariate recurrent event data [12, 13], and the addition of longitudinal data [14].

Prediction using frailty models is relatively underdeveloped. In the framework of joint models for longitudinal and time-to-event data, Proust-Lima *et al.* [2] and Rizopoulos [3] developed dynamic prediction tools to predict the risk of death given the trajectory of a longitudinal biomarker. In the first approach, the prediction was validated through an error of prediction based on loss function. In the second approach, the prediction was validated in the framework of the area under the receiver operating characteristic curve methodology. Recently, Li *et al.* [15] proposed a cure frailty model in order to predict individual long-term smoking cessation (or smoking abstinence). The area under the receiver operating characteristic curve was also used to measure the discrimination capacity of the proposed tool. At this time and to our knowledge, no prediction tool has been proposed to take the previous occurrences of a recurrent event into account.

This work was motivated by a prospective study including patients with primary invasive breast cancer, treated at Institut Bergonié, a French comprehensive cancer center. Following the diagnosis of breast cancer and a surgical treatment, patients might experience several events, including loco-regional relapses and distant metastases. We are interested in the evolution of their risk of death with the onset of these disease events. In this aim, we propose a dynamic prediction tool to estimate the probability of death between t and $t+w$, knowing the number and times of previous recurrences, in the framework of joint survival models. The estimated probability can be updated following a new disease relapse. Three prediction settings were developed. The first one considers the exact recurrence history of the patient before time t. The second one considers the observed recurrences but considers that others may arise before the prediction time. The last one does not consider the patient recurrence history in the prediction but only in the parameter estimation.

Section 2 of this paper presents the joint modeling for recurrent events and a terminal event. Section 3 presents the three proposed dynamic prediction tools for the risk of death and standard-error estimation. A measure of error of prediction is presented in Section 3 . An application on the motivating dataset in breast cancer is presented in Section 4 . Finally, Section 5 contains a discussion and concluding remarks.

2. Joint survival modeling for recurrent events and a terminal event

2.1. Joint gamma frailty model for recurrent events and a terminal event

A joint model for recurrent events and a terminal event was previously detailed [5, 7]. We denote for subject i ($i = 1,\dots,N$), X_{ij} the j^{th} recurrent time ($j = 1,\dots,n_i$), C_i the censoring time (not by death), and D_i the death time. $T_{ij}^R = min(X_{ij}, C_i, D_i)$ corresponds to each follow-up time, and δ_{ij}^R is a binary indicator for recurrent events, which is 0 if the observation is censored or if the subject died and 1 if X_{ij} is observed ($\delta_{ij}^R = I[T_{ij}^R = X_{ij}]$ where $I[.]$ denotes indicator function). Similarly, we note T_i^D the last follow-up time for subject i, which is either a time of censoring or a time of death ($T_i^D = min(C_i, D_i)$) and $\delta_i^D = I[T_i^D = D_i]$. We actually observe the sequence $(T_{ij}^R, \delta_{ij}^R, T_i^D, \delta_i^D)$. Finally, we denote by Z_{ij}^R and Z_i^D the vectors of covariates associated with the risk of recurrent events and death, respectively. Both death and recurrent times are in the calendar timescale, that is, measured by the time elapsed since the origin of the study. However, a patient is considered at risk of a j^{th} recurrence only after the $(j-1)^{st}$ recurrence.

In the calendar timescale, the joint frailty model for recurrent events and the death is

$$\begin{cases} \lambda_{ij}^R(t|u_i) = u_i \lambda_0^R(t) \exp\left(\beta_1' Z_{ij}^R\right) = u_i \lambda_{ij}^R(t) \\ \lambda_i^D(t|u_i) = u_i^\alpha \lambda_0^D(t) \exp\left(\beta_2' Z_i^D\right) = u_i^\alpha \lambda_i^D(t) \end{cases} \quad (1)$$

where $\lambda_0^R(.)$ is the baseline risk of event, irrespective of event rank, and $\lambda_0^D(.)$ the baseline risk of death. The effects of explanatory variables β_1 and β_2 are assumed to be different for the risk of recurrent events and the risk of death. The two processes are linked by the patient-specific frailty effect u_i. The frailty is assumed to follow a gamma distribution with variance θ and, without loss of generality, a mean equal 1. That is

$$u_i \sim \text{Gamma}\left(\frac{1}{\theta};\frac{1}{\theta}\right) \quad \text{and} \quad g(u_i) = \frac{u^{1/\theta-1}\exp(-u/\theta)}{\theta^{1/\theta}\Gamma(1/\theta)} \tag{2}$$

The frailty effects u_i are assumed independent. The between-subject heterogeneity is considered significant if the variance of the frailty θ differs from 0. The presence of the α term allows more flexibility in the model. When $\alpha = 1$, the frailty has an identical effect on the risk of recurrent events and on the risk of terminal event. When $\alpha > 0$, the recurrent events rate and the terminal event rate are positively associated. Finally, $\alpha = 0$ would show that $\lambda_i^D(t|u_i)$ does not depend on u_i and thus that the terminal event process does not depend on the recurrent events process. The interpretation of α makes sense only when the variance θ is statistically different from zero.

2.2. *Penalized likelihood estimation*

Rondeau *et al.* [7] proposed the inference of the joint model that is based on the semiparametric penalized likelihood approach. We denote the parameters vector by $\xi = \left(\lambda_0^R(.), \lambda_0^D(.), \boldsymbol{\beta}, \alpha, \theta\right)$. The full log-likelihood in the calendar timescale is the following expression:

$$\begin{aligned} l(\xi) = \sum_{i=1}^{N}\Bigg\{ & \sum_{j=1}^{n_i}\delta_{ij}^R \log\lambda_{ij}^R\left(T_{ij}^R\right) + \delta_i^D\log\lambda_i^D\left(T_i^D\right) - \log\Gamma(1/\theta) - \frac{1}{\theta}\log\theta \\ & + \log\int_0^\infty u^{\left(N_i^R(T_i^D)+\alpha\delta_i^D+1/\theta-1\right)}\exp\left(-u\int_0^{T_i^D}\lambda_{ij}^R(t)\mathrm{d}t - u^\alpha\int_0^{T_i^D}\lambda_i^D(t)\mathrm{d}t - \frac{u}{\theta}\right)\mathrm{d}u\Bigg\} \end{aligned} \tag{3}$$

where $N_i^R(t)$ is the observed number of recurrent events at time t.

The baseline hazard functions ($\lambda_0^R(.)$ for recurrent events and $\lambda_0^D(.)$ for death) are approximated using splines. To constraint smooth functions, we penalize the likelihood by a term that has large values for rough functions. The penalized log-likelihood is as follow:

$$lpl(\xi) = l(\xi) - \kappa_1\int_0^\infty\left\{\left(\lambda_0^R\right)''(t)\right\}^2\mathrm{d}t - \kappa_2\int_0^\infty\left\{\left(\lambda_0^D\right)''(t)\right\}^2\mathrm{d}t \tag{4}$$

with κ_1 and κ_2 two positive smoothing parameters. They control the trade-off between the data fit and the smoothness of the functions. These two coefficients are chosen using an approximate likelihood cross-validation criterion [16]. This cross-validation can only be made for one parameter at a time. Consequently, in joint modeling framework, the estimation of the smoothing parameters is made in two steps. As a first step, two separate models are fitted: one shared frailty model for the risk of recurrent event and one Cox model for the risk of death. The two obtained cross-validated values of κ_1 and κ_2 can then be used in the joint model. The penalized log-likelihood is maximized using a modified robust Marquardt algorithm ([17], see also [7]). The integrals are approximated using a Gauss–Laguerre quadrature.

3. Dynamic marginal prediction of risk of death

We propose three different prediction tools that take into account the recurrence history of the patient at different levels. They are illustrated in Figure 1. In the first setting, we consider the exact recurrence history of the patient, that is, J recurrences at the times they occurred and considering that no more than J recurrences occurred before t. In this case, we consider that the patient may have a $J+1^{st}$ recurrence only after the prediction time t. An example is a patient was diagnosed 5 years ago. During these 5 years, he underwent two recurrences. We want to know his probability of death during the next 5 years (that is, during 5 and 10 years after diagnosis) considering these two recurrences. In a second setting, we consider the history of the patient as having at least the J observed recurrences before the prediction time t, whatever happens between the J^{th} event and the prediction time t. This makes projections possible.

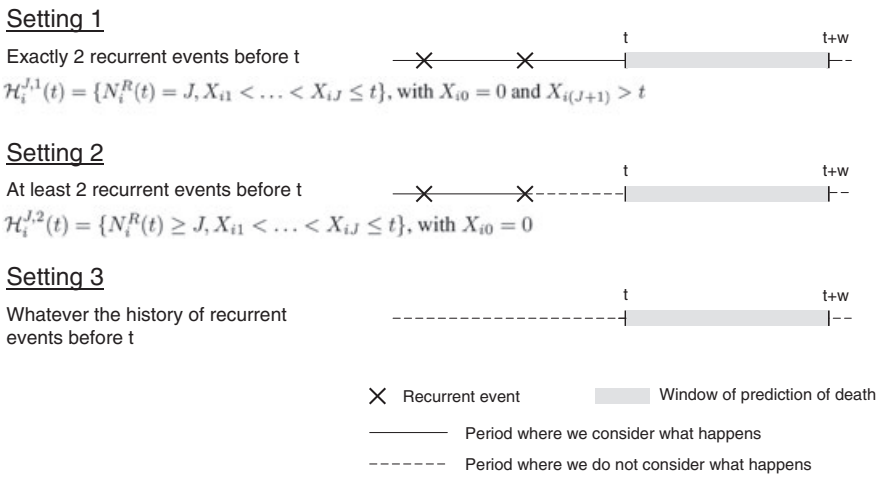

Figure 1. Three settings to take into account patient history of recurrent events in prediction. Illustration with two recurrent events.

An example is a patient just had his second recurrence 3 years after his diagnosis. We want to make projection for this patient. A question could be if this patient is still alive at 5 years (whatever happens until then), what is his probability of death from then to 10 years considering the recurrences he already had? For example, if he is still alive at 5 years, would his risk of death become lower then? In the last setting, history of recurrences is considered in the estimation of the parameters using the joint model (1) like in the two first settings, but it is not considered in the prediction calculation. That is, it corresponds to an average risk of death at a population level and not to an individual prediction considering the history of the patient. An example is a patient was just diagnosed, and we want to know his probability of death during the first 5 or 10 years. And if he is still alive at 5 years, what would be his probability of death during the next 5 or 10 years? This last probability can also be compared with the first and second ones to assess the contribution of the recurrent history to the prediction.

For each of these three settings, we propose a marginal prediction. That is, we integrated the prediction conditional on the random effect over the distribution of the random effects. This makes prediction for new patients possible. Indeed, frailty of a patient not included in the population used to build the prediction model is unknown. Thus, it is not possible to make conditional predictions for this patient. Moreover, the marginal approach only needs the estimation of the frailties distribution, via the estimation of the variance θ, without estimating frailty of each patient.

The following notations are shared by the three settings. Let t and w be the time of prediction and the window of prediction, respectively. We are interested in the probability of death between t and $t + w$. Let D_i denote the time of death for subject i and X_{ij} the j^{th} observed recurrent time for subject i (time since the origin of the study). Consider a new subject i alive at time t (i.e., $D_i > t$), for whom we observe J recurrences before t (i.e., we observe $X_{i1} < X_{i2} < \ldots < X_{iJ} < t$) and for whom the vectors of covariates Z_{ij}^R relative to the risk of recurrences and Z_i^D relative to the risk of death are available at time t of prediction. $S_{ij}^R(t) = P(X_{ij} \geq t)$ and $S_i^D(t) = P(D_i \geq t)$ are two survival functions. We denote the vector of all parameters by $\xi = \left(\lambda_0^R(.), \lambda_0^D(.), \boldsymbol{\beta}, \alpha, \theta\right)$.

To define the probability of death given the recurrence history of the patient, we need to define the recurrence history of the patient. We will define two types of history, one complete and one partial, which will be used for the first and second probability setting, respectively. The complete recurrence history of the patient i until time t is defined by $\mathcal{H}_i^{J,1}(t) = \left\{N_i^R(t) = J, X_{i1} < \ldots < X_{iJ} \leqslant t\right\}$, with $X_{i0} = 0$ and $X_{i(J+1)} > t$. In this case, we consider that we observed the patient's complete history and that no more than the J considered recurrences occurred before t. The partial recurrence history of the patient until time t is defined by $\mathcal{H}_i^{J,2}(t) = \left\{N_i^R(t) \geq J, X_{i1} < \ldots < X_{iJ} \leqslant t\right\}$, with $X_{i0} = 0$. In this case, we consider that we observed J recurrences but that others may have occurred before t.

3.1. Probability of death between t and t + w considering exactly J recurrences: $P^1(t, t + w; \xi)$

We are first interested in the probability of death between t and $t + w$ given the patient had exactly J recurrences before t. We use $\mathcal{H}_i^{J,1}(t)$, and we have $X_{i(J+1)} > t$. The posterior probability of death

between t and $t+w$ (i.e., that we observe $t \leqslant D_i \leqslant t+w$) for the parameter values ξ can be computed by

$$\begin{aligned} P^1(t,t+w;\xi) &= P\left(D_i \leqslant t+w | D_i > t, \mathcal{H}_i^{J,1}(t), Z_{ij}^R, Z_i^D, \xi\right) \\ &= \int_0^\infty P\left(D_i \leqslant t+w | D_i > t, \mathcal{H}_i^{J,1}(t), Z_{ij}^R, Z_i^D, u_i, \xi\right) \\ &\quad \times g\left(u_i | D_i > t, \mathcal{H}_i^{J,1}(t), Z_{ij}^R, Z_i^D, \xi\right) \mathrm{d}u_i \end{aligned}$$

where $g\left(u_i | D_i > t, \mathcal{H}_i^{J,1}(t), Z_{ij}^R, Z_i^D, \xi\right)$, the conditional density of the frailty u_i given the patient i is alive at time t, given his history and covariates and given the parameters ξ, is defined in the appendix.

Considering, on one hand, the independence of patient recurrent event times and, on the other hand, the independence of the recurrent event times and the death time given the random effect, we obtain the following probability (details can be found in the Appendix):

$$P^1(t,t+w;\xi) = \frac{\int_0^\infty \left[S_i^D\left(t|Z_i^D,u_i,\xi\right) - S_i^D\left(t+w|Z_i^D,u_i,\xi\right)\right](u_i)^J S_{i(J+1)}^R\left(t|Z_{ij}^R,u_i,\xi\right) g(u_i)\mathrm{d}u_i}{\int_0^\infty S_i^D\left(t|Z_i^D,u_i,\xi\right)(u_i)^J S_{i(J+1)}^R\left(t|Z_{ij}^R,u_i,\xi\right) g(u_i)\mathrm{d}u_i} \tag{5}$$

where $g(u_i)$ is the density of the gamma distribution defined in Equation (2).

The estimated posterior probabilities, $\hat{P}^1\left(t,t+w;\hat{\xi}\right)$, can be obtained by substituting ξ by the maximum penalized likelihood estimates $\hat{\xi} = \left(\hat{\lambda}_0^R(.), \hat{\lambda}_0^D(.), \hat{\boldsymbol{\beta}}, \hat{\alpha}, \hat{\theta}\right)$ and the individual information for the covariates Z_i^D and Z_{ij}^R into this equation. The J recurrence times $(X_{i1},\ldots,X_{iJ})$ have no influence on the value of $P^1(t,t+w;\xi)$. These results of mathematical reductions are possible because of the use of multiplicative model.

3.2. *Probability of death between t and $t+w$ considering at least J recurrences: $P^2(t,t+w;\xi)$*

We are now interested in the probability of death between t and $t+w$ given the patient had at least J recurrences before t. We use $\mathcal{H}_i^{J,2}(t)$, and compared with the previous probability, we do not have the condition $X_{i(J+1)} > t$. The probability of death becomes

$$\begin{aligned} &P^2(t,t+w;\xi) \\ &= P\left(D_i \leqslant t+w | D_i > t, \mathcal{H}_i^{J,2}(t), Z_{ij}^R, Z_i^D, \xi\right) \\ &= \frac{\int_0^\infty \left[S_i^D\left(t|Z_i^D,u_i,\xi\right) - S_i^D\left(t+w|Z_i^D,u_i,\xi\right)\right](u_i)^J S_{iJ}^R\left(X_{iJ}|Z_{ij}^R,\xi,u_i\right) g(u_i)\mathrm{d}u_i}{\int_0^\infty S_i^D\left(t|Z_i^D,\xi,u_i\right)(u_i)^J S_{iJ}^R\left(X_{iJ}|Z_{ij}^R,\xi,u_i\right) g(u_i)\mathrm{d}u_i} \end{aligned} \tag{6}$$

In this setting, and because of the same reduction as in Eq. (5), only the last recurrent event time is taken into account.

3.3. *Probability of death between t and $t+w$ considering the recurrence history only in the parameters estimation: $P^3(t,t+w;\xi)$*

This last probability corresponds to the average probability of death between t and $t+w$ for a patient with characteristics Z_i^D. This setting does not consider the history of past recurrent events in the prediction probability. It is simply equal to

$$\begin{aligned} P^3(t,t+w;\xi) &= P\left(D_i \leqslant t+w | D_i > t, Z_i^D, \xi\right) \\ &= \frac{\int_0^\infty \left[S_i^D\left(t|Z_i^D,u_i,\xi\right) - S_i^D\left(t+w|Z_i^D,u_i,\xi\right)\right] g(u_i)\mathrm{d}u_i}{\int_0^\infty S_i^D\left(t|Z_i^D,\xi,u_i\right) g(u_i)\mathrm{d}u_i} \end{aligned} \tag{7}$$

In this setting, neither the recurrent event times nor the number of recurrent events is directly taken into account.

3.4. Variability of the probability estimators

For each of the three settings, a percentile confidence interval was estimated using the Monte Carlo method. Calculation of the probabilities is based on the estimated values of the parameters $\hat{\xi} = \left(\hat{\lambda}_0^R(.), \hat{\lambda}_0^D(.), \hat{\boldsymbol{\beta}}, \hat{\alpha}, \hat{\theta}\right)$. We drawn V = 1000 vectors ξ^v from the normal approximation of the distribution of ξ estimated by the model: $\mathcal{MN}\left(\hat{\xi}, \hat{\Sigma}_{\hat{\xi}}\right)$. Let $\hat{P}\left(t, t+w; \hat{\xi}\right)$ be a generic term for $\hat{P}^1\left(t, t+w; \hat{\xi}\right)$, $\hat{P}^2\left(t, t+w; \hat{\xi}\right)$, or $\hat{P}^3\left(t, t+w; \hat{\xi}\right)$. For each vector ξ^v, the corresponding probabilities $P^v(t, t+w; \xi^v)$ were computed. A confidence interval for $\hat{P}\left(t, t+w; \hat{\xi}\right)$ was obtained using the 2.5^{th} and 97.5^{th} percentiles of $P^v(t, t+w; \xi^v)$.

4. Prediction error and validation

In order to validate the dynamic prediction tools, we calculate prediction error curves. This error is based on a weighted time-dependent Brier score, which corresponds to a difference between what is observed (the patient survival status at time $t+w$) and what was predicted by the model (the survival probability at time $t+w$: $1-\hat{P}(t, t+w; \hat{\xi})$). We used an inverse probability of censoring weighted error to take the right censoring into account [18]. Let N_t be the number of patients alive and uncensored at prediction time t, that is, patients for whom the prediction can be made. Using T_i^D the right censored survival time, δ_i^D the corresponding event indicator ($\delta_i^D = 1$ if the observed time is a death time, 0 otherwise) and $\widehat{G}_N(t)$ the Kaplan–Meier estimate of the population censoring distribution, and using the generic term $\hat{P}\left(t, t+w; \hat{\xi}\right)$, which can be each of the three prediction probability setting previously proposed, the error of prediction is defined by

$$Err_{t+w} = \frac{1}{N_t} \sum_{i=1}^{N_t} \left[I\left(T_i^D > t+w\right) - \left(1 - \hat{P}\left(t, t+w; \hat{\xi}\right)\right)\right]^2 \hat{w}_i\left(t+w, \hat{G}_n\right)$$

with $\hat{w}_i(t+w, \hat{G_n})$ being the weight that accounts for right censoring:

$$w_i\left(t+w, \hat{G}_n\right) = \frac{I\left(T_i^D \leqslant t+w\right)\delta_i^D}{\hat{G}_n\left(T_i^D\right)/\hat{G}_n(t)} + \frac{I\left(T_i^D > t+w\right)}{\hat{G}_n(t+w)/\hat{G}_n(t)}$$

In order to see if the prediction tool is able to predict accurately for new patients not used to build the model, a 10-fold cross-validation is carried out [19]. The whole population is randomly split in 10 near-equal size partitions. In each of the 10 steps, one partition is left out. The joint model is built on the remaining nine partitions, and estimated parameters are used to predict probability of death for patients from the left-out partition. At the end, each patient of the population had a predicted probability of death. The prediction error curve was then based on these probabilities.

5. Application: risk of death after recurrences in patients with operable breast cancer

5.1. Breast cancer population and prognostic joint model

The development of these prediction tools was based on a cohort of patients from Institut Bergonié, a French comprehensive cancer center. Between 1989 and 1993, 1161 patients with primary operable invasive ductal carcinoma or invasive lobular carcinoma were included and followed-up until 2010. All were operated as first treatment, and the surgery date was chosen as origin date for the definition of survival times. The median follow-up according to reverse Kaplan–Meier (estimating the censoring dis-

Table I. Joint model estimations on breast cancer population (n=1067 patients, 427 recurrent events).

		For recurrent events		For death	
Variable	% of patients	HR	(95%CI)	HR	(95%CI)
Age (years)	range: 28–84				
[40 − 55] vs >55	36.6	1.18	(0.92–1.51)	0.36	(0.19–0.66)
⩽ 40 vs >55	7.7	2.54	(1.82–3.56)	1.76	(0.82–3.81)
Peritumoral vascular invasion	26.7	1.47	(1.15–1.88)	3.35	(1.80–6.25)
Tumor size (> 20 mm vs ⩽ 20 mm)	22.7	1.86	(1.47–2.37)	4.68	(2.70–8.12)
HER2 positive	11.2	1.43	(1.03–1.99)	1.31	(0.62–2.77)
Hormonal receptors (positive vs negative)	83.0	0.81	(0.57–1.16)	0.23	(0.10–0.54)
Nodal involvement (yes vs no)	42.3	1.82	(1.42–2.32)	4.52	(2.43–8.41)
Grade					
II vs I	45.7	2.14	(1.55–2.95)	7.99	(3.39–18.85)
III vs I	24.6	2.21	(1.48–3.31)	10.80	(4.13–33.76)
θ		1.04	(SE = 0.06)		
α		4.61	(SE = 0.28)		
LCV		2.04			
κ_1		1000000			
κ_2		13000			

HR, hazard ratio; CI, confidence interval; LCV, likelihood cross-validation criterion; HER2, human epidermal growth factor receptor 2; SE, standard error.

tribution, i.e., where the stop of follow-up is considered as an event that may be censored by the death time) was 13.8 years. During the follow-up, patients underwent disease relapses, which could be loco-regional relapses or distant metastases or both at the same time. Patients underwent a maximum of three successive events per patient. Joint models account for the dependence between these recurrent events in one patient and also the dependence between these recurrent events and death. However, the influence of the recurrences on the prediction of the risk of death has not been studied.

Among the 1067 patients without missing data, 362 underwent at least one disease relapse. Among them, 301 had only one relapse (114 were alive at the end of the follow-up and 187 died), 57 had two relapses (20 alive and 37 died), and four had three relapses (three alive and one died). Among the 705 patients without relapse, 600 were alive at the end of follow-up and 105 died.

Studied covariates, measured at time of surgery, were age (younger than 40 years or between 40 and 55 years versus older than 55 years), menopausal status (menopause and post-menopause versus other), genomic tumoral classification (Luminal A, Luminal B, or triple negative versus human epidermal growth factor receptor 2 [HER2]-enriched), pathological tumor size (greater than 20 mm versus 20 mm and less), peritumoral vascular invasion (yes versus no), HER2 expression (positive versus negative), hormonal (estrogen or progesterone) receptor (HR, positive versus negative), proliferation index Mib-1 (positive versus negative), pathological node involvement (yes=at least one, versus no), and histological grade of the tumor (grade II or grade III versus grade I). We fitted a joint model in the calendar timescale (time elapsed since the origin of the study). For parameter estimation, we used all the information available, that is all patients and all recurrences, irrespective of the time of occurrence. Covariate selection was made through backward stepwise selection. Results of the final model are shown in Table I. Seven covariates are associated with the risk of relapse or the risk of death. The variance of the frailty effect $\theta = 1.04$ (standard error, SE = 0.06) is significantly higher than 0 and the parameter $\alpha = 4.61$ (SE = 0.28). This indicates that the risk of a recurrent event and the risk of death are significantly and positively associated and that frailty has a greater effect on the risk of death. That means that the non-observed covariates have a greater effect on risk of death than on risk of recurrent event, which is coherent with the results for observed covariates.

5.2. *Predicted risk of death knowing the history of recurrent events (relapses)*

After estimating the parameters, the probability of death was calculated using the information known at the time of prediction t and ignoring what happened after this time. The maximum horizon for prediction was 15 years, considering that information was insufficient to make accurate prediction after this time.

Table II. Probabilities of death between 5 and 10 or 15 years knowing the generated history at 5 years of an average* patient, with P^1, P^2, and P^3, the three predicted probabilities of death from the joint model (Equations (5), (6), and (7)), with their standard errors.

	Risk of death between 5 and 10 years (%)			Risk of death between 5 and 15 years (%)		
Recurrence history	$P^1(5, 10; \hat{\xi})$	$P^2(5, 10; \hat{\xi})$	$P^3(5, 10; \hat{\xi})$	$P^1(5, 15; \hat{\xi})$	$P^2(5, 15; \hat{\xi})$	$P^3(5, 15; \hat{\xi})$
No recurrence	10.8 (4.2)	10.8 (4.2)	12.7 (4.5)	22.7 (4.8)	22.7 (4.8)	25.6 (4.7)
One recurrence						
$X_{i1} = 1$	30.3 (8.9)	33.3 (8.9)	12.7 (4.5)	53.0 (6.9)	56.2 (6.3)	25.6 (4.7)
$X_{i1} = 2.5$	30.3 (8.9)	32.3 (8.9)	12.7 (4.5)	53.0 (6.9)	55.2 (6.5)	25.6 (4.7)
$X_{i1} = 4.9$	30.3 (8.9)	30.4 (8.9)	12.7 (4.5)	53.0 (6.9)	53.1 (6.9)	25.6 (4.7)
Two recurrences						
$X_{i1} = 1, X_{i2} = 2$	50.6 (11.4)	53.2 (11.1)	12.7 (4.5)	75.6 (6.0)	77.5 (5.4)	25.6 (4.7)
$X_{i1} = 2, X_{i2} = 4$	50.6 (11.4)	51.5 (11.3)	12.7 (4.5)	75.6 (6.0)	76.3 (5.8)	25.6 (4.7)
$X_{i1} = 4, X_{i2} = 4.9$	50.6 (11.4)	50.7 (11.4)	12.7 (4.5)	75.6 (6.0)	75.6 (6.0)	25.6 (4.7)
Three recurrences						
$X_{i1} = 1, X_{i2} = 2, X_{i3} = 3$	67.4 (11.9)	68.9 (11.4)	12.7 (4.5)	88.4 (4.1)	89.2 (3.7)	25.6 (4.7)
$X_{i1} = 1, X_{i2} = 2.5, X_{i3} = 4.9$	67.4 (11.9)	67.5 (11.9)	12.7 (4.5)	88.4 (4.1)	88.4 (4.1)	25.6 (4.7)
$X_{i1} = 3, X_{i2} = 4, X_{i3} = 4.9$	67.4 (11.9)	67.5 (11.9)	12.7 (4.5)	88.4 (4.1)	88.4 (4.1)	25.6 (4.7)

*Average patient: patient with age $>$ 55 years, no peritumoral vascular invasion, tumor size $\leq$ 20 mm, human epidermal growth factor receptor 2 negative, hormonal receptors positive, no lymph node involvement and tumor grade II.

Table II presents some probabilities for an average patient: age $>$ 55 years, no peritumoral vascular invasion, tumor size $\leqslant$ 20 mm, HER2 negative, HR positive, no lymph node involvement, and tumor grade II. The prediction time is $t = 5$ years. We compared predicted risks for different scenarios: different windows of prediction ($w = 5$ and $w = 10$ years), different number of relapses (0 to 3), and different generated relapse times (early or late). The number of relapses before 5 years increases the risk of death

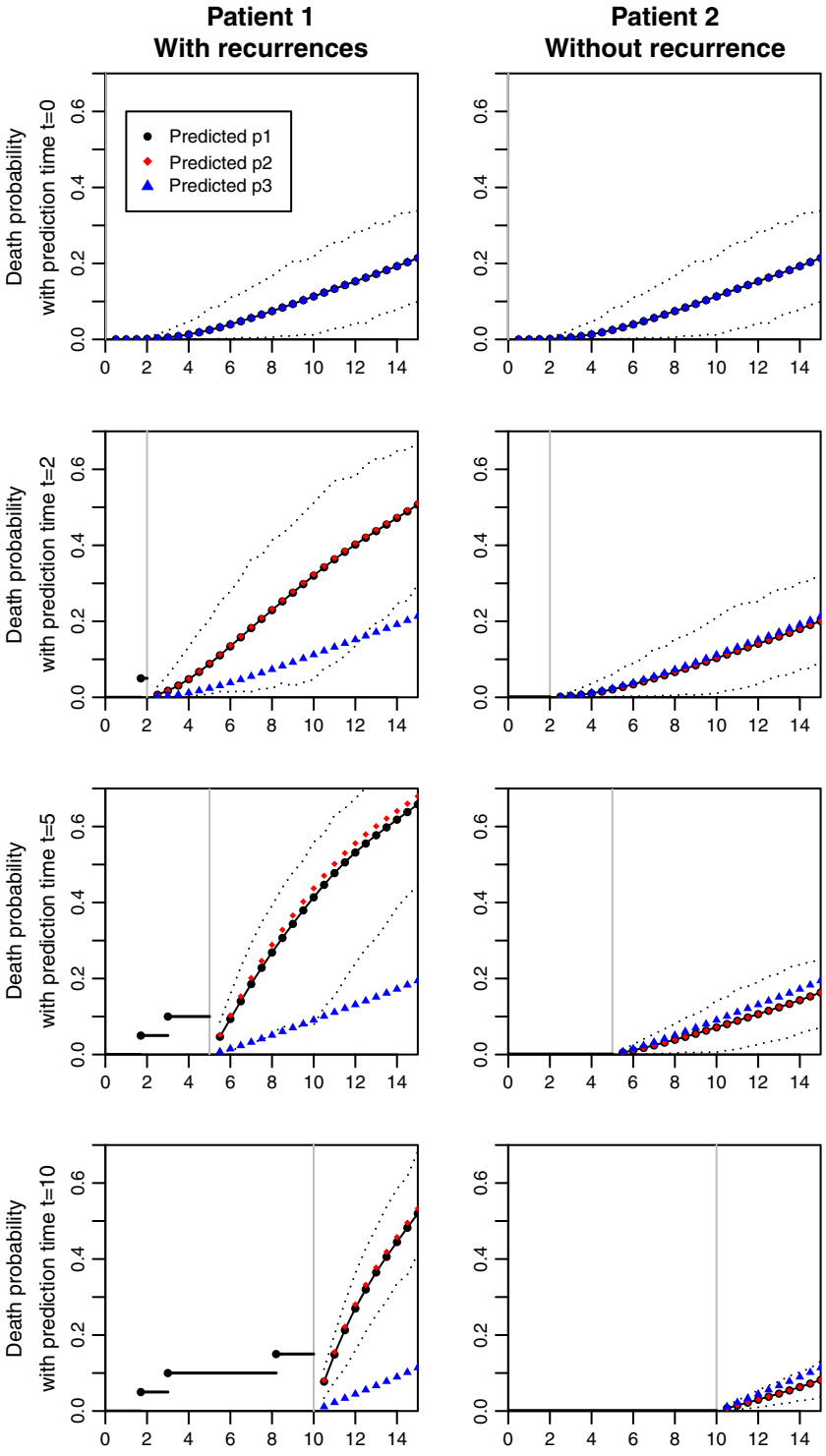

Figure 2. Evolution of the three prediction settings for two patients sharing same characteristics except the occurrence of recurrences. The vertical gray line represents the time of prediction t. Horizontal points and lines represent the recurrent event process. Dashed lines represent the confidence interval bands for the prediction p1.

between 5 and 10 years, from $P^1 = 10.8\%$ for no relapse to $P^1 = 67.4\%$ for three relapses. Similar results are observed for risks of death between 5 and 15 years, but probabilities of death were higher: from $P^1 = 22.7\%$ for no relapse to $P^1 = 88.4\%$ for three relapses. Corresponding probabilities of death predicted by a Cox model would be 9.7% between 5 and 10 years and 19.1% between 5 and 15 years. As expected, only the two first probabilities are influenced by the occurrence of relapses, whereas the third is not. P^1 and P^2 have close values, especially when last relapse time is close to the prediction time t. Because of the use of a multiplicative model, P^1 only depends on the number of relapses and not on the time of these relapses, whereas P^2 also depends on the time of the last relapse. The value of P^2 decreases when the last relapse occurs later. Finally, two thirds of the population did not have any relapse. This explains that the value of the average population risk, estimated by P^3, is close to the value of the probabilities P^1 and P^2 for the 'no recurrence' scenario.

Figure 2 compares the predicted risk of death for two patients from the cohort having the same characteristics: patient aged between 40 and 55 years, no peritumoral vascular invasion, tumor size $\leqslant 20$ mm, HER2 negative, HR positive, no lymph node involvement, and tumor grade II. The first patient underwent three relapses, at 1.7, 3.0, and 8.2 years. This patient died 10.8 years after the surgery. The second patient is a patient who was followed for 18.9 years, without any relapse, and who was alive at the end of follow-up. The figure represents predicted risks at prediction times $t = 0$, $t = 2$, $t = 5$, and $t = 10$ years, with horizon up to 15 years. At $t = 0$, the three settings gave the same prediction, for both patients. The predicted risks P^1 and P^2 are always very close and increase with the occurrence of relapses in the first patient. They slightly decrease for the second patient in the absence of relapse. P^3 stays equal for both patients as it only depends on the time of prediction t but not on the occurrence of relapses.

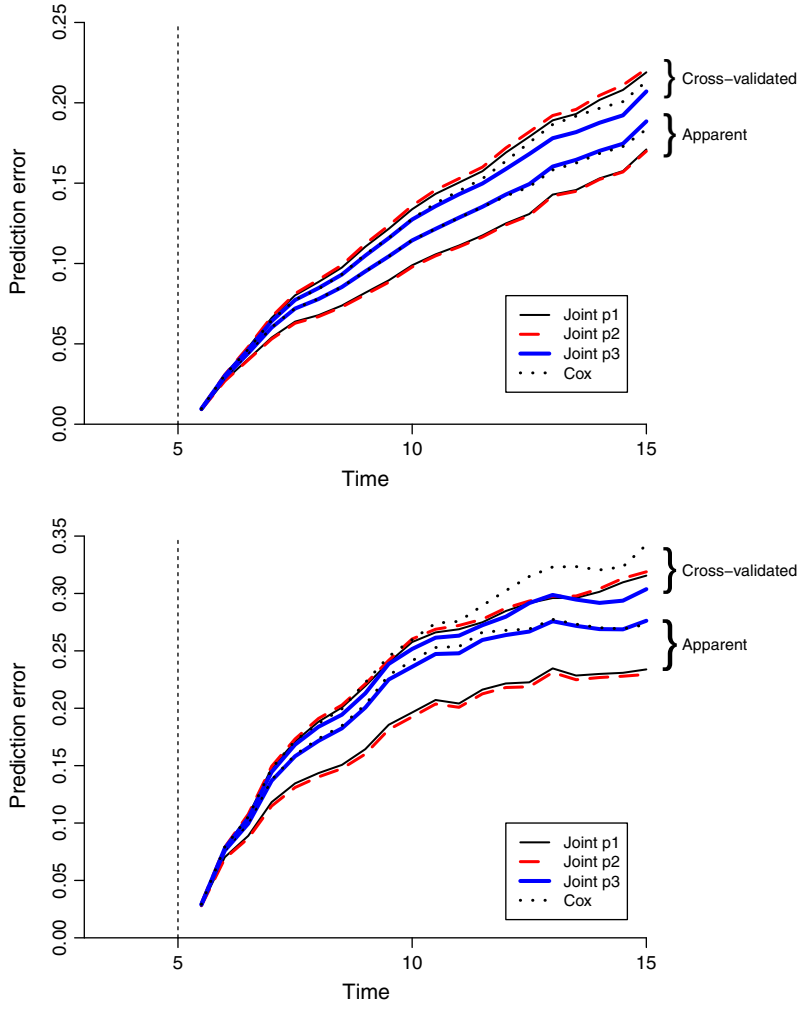

Figure 3. A—Apparent and 10-fold cross-validated error of prediction at $t = 5$ years (n=949 patients alive). B—Apparent and 10-fold cross-validated error of prediction at $t = 5$ years for patients with recurrent events (n=267 patients alive).

5.3. Validation of the prediction

The dynamic prediction tool was validated by measuring an error of prediction. The three predictions were made for all patients of the cohort at a prediction time $t = 5$ years, and the predicted survivals $(1 - \hat{P}(t, t + w; \hat{\xi}))$ were compared with the actual status of the patient at different horizons, up to 15 years. An apparent error of prediction was computed and is illustrated in Figure 3A. Error of prediction on the whole population was very close for the two probabilities P^1 and P^2 and was higher for the probabilities ignoring the relapses P^3 and the probability of the Cox model. This shows that considering the relapses in the prediction of death could lead to more accurate prediction. When the prediction was made on independent patients, using the 10-fold cross-validation method, the cross-validated error of prediction was higher than the apparent error for all settings, as expected. This error of prediction was very close for all settings and no longer lower for P^1 and P^2.

The main advantage of the proposed prediction tools is to take recurrent event information into account. Thus, it appears useful to measure the error of prediction in a subgroup of patients having at least one relapse during follow-up (Figure 3B). We can see that for these patients, the error of prediction is lower for the predictions from joint modeling than from a Cox model. This result is also seen in cross-validation, mainly after 10 years.

6. Discussion

We developed three prediction tools in order to take into account the occurrence of recurrent events in the prediction of risk of death in breast cancer patients. Similar predictions were previously developed in the framework of joint models for longitudinal data and survival time [2, 3]. However, they were not proposed for joint models for recurrent events and a terminal event. Instead of considering evolution of a time-dependent marker, we considered occurrence of a repeated event. We focused on three settings: considering the exact information that we have before the prediction time t, considering partial information, and considering a population risk ignoring the occurrence of recurrent events.

Similar works were also performed on the framework of multi-state models in various applications (see for examples [20–23]). In particular, Putter *et al.* [22] studied the prognosis of patients with breast cancer using multistate model, possible states being surgery, local recurrence only, distant metastasis only, local recurrence and distant metastasis combined, and death. In this work, they were able to evaluate the impact of the occurrence of intermediate events on the risk of death. However, this type of model does not allow quantifying the correlation between the risk of relapses and the risk of death. Moreover, to mimic the joint frailty modeling approach, which can account for several occurrences of a given event, a multistate model would require additional transitions, increasing the potential number of parameters to be estimated. Further work is required to investigate the pros and cons of both approaches and offer the best predictive tool in breast cancer for clinical use.

For simplicity purposes, in our development, we did not consider the possibility of time-dependent covariates, for example, modification in patient's treatment. However, in the same way that these covariates are allowed in joint modeling, they are allowed in prediction calculation. Care must be taken in the value to be used at the time of prediction and evolution of the time-dependent covariate beyond the prediction time t must be ignored in the prediction calculation.

Recurrent events were defined as loco-regional relapse or distant metastasis. However, it is possible that these two types of event do not similarly impact on the risk of death. An extension of the proposed prediction tools could be to consider separately these two types of event. This could be performed either with adequate covariate adjustments, or with two separate models, or through a multivariate joint model.

When studying recurrent events, different timescales can be used [24]. The timescale that is most often used is the gap time: after an event, the subject starts again at time 0, and the time to the next event corresponds to the number of days elapsed between the two successive events. An alternative timescale is the calendar time, also called the counting process approach, which keeps track of time since randomization. The duration of the time at risk for an event corresponds to the duration of the time at risk in the gap time representation. However, the starting time of the period at risk is not reset to 0. A subject is, therefore, not considered to be at risk for the j^{th} event until after the $(j-1)^{th}$ event. A subject can experience different periods at risk during the total observation time. The choice of the timescale has hardly any impact on the parameter estimation and has to be made with the advice of clinicians.

Only marginal predictions were proposed, although conditional prediction could be considered. Expression of the conditional prediction is

$$P^{cond}(t,t+w;\xi) = P\left(D_i \leqslant t+w | D_i > t, Z_i^D, u_i, \xi\right)$$
$$= \frac{S_i^D\left(t|Z_i^D,u_i,\xi\right) - S_i^D\left(t+w|Z_i^D,u_i,\xi\right)}{S_i^D\left(t|Z_i^D,u_i,\xi\right)} = 1 - \left(\frac{S_0^D(t+w)}{S_0^D(t)}\right)^{u_i^\alpha \exp(\beta' Z_i^D)}$$

where u_i is the frailty for the patient i considered and $S_0^D(.)$ is the baseline survival function. The $\hat{u}_i s$ are obtained from the posterior distribution of the $u_i s$ conditional on the observed data, knowing the estimated values of the regression parameters [25, 26]. Thus, conditional prediction is only possible for patients used in the model estimation, that is, patients for whom we already know what happened (at least mostly). Consequently, the interest of such prediction is limited in the framework of recurrent events. Moreover, Komarek *et al.* stated, when they were interested in prediction using a multivariate linear mixed model in a bayesian framework, that the conditional prediction ignores variability in the estimation of the individual random effect [27]. Same criticisms apply to the present framework.

In the model estimation, two components may influence the prediction probability: the baseline hazard functions and the distribution of the frailty term. We chose to approximate the baseline hazard functions by splines to allow flexible functions, as we did not have any a priori on the shape of these risk functions. However, parametric approximations could be used, for example, Weibull or piecewise constant functions. Indeed, flexibility is possible at the price of a higher number of parameters to estimate. This high number of parameters favors over-fitting and could be a concern in the prediction framework. Models with a large number of parameters could lead to a large variance and so be unstable and have poor predictive quality [28]. This may explain in part the error of prediction observed with the 10-fold cross-validation. Moreover, a misspecification of the frailty term distribution may lead to an inaccurate prediction of the individual frailty terms and so an inaccurate prediction. In order to assess the influence of a misspecification of the frailty distribution, we performed a simulation study. We simulated 100 datasets with gamma distributed frailty and estimated the prediction probability (first setting) corresponding to a joint frailty model assuming (i) a gamma distribution and (ii) a log-normal distribution of the frailty terms. The error of prediction obtained with these two models was very close whatever the time of prediction. This suggests that the impact of the choice of the frailty distribution is limited. It was previously demonstrated that the impact of the frailty distribution on the parameter estimation in the framework of joint model is limited [29].

We used here an error of prediction based on the inverse probability of censoring weighted, with a Kaplan–Meier estimate. In the weights definition, the censoring distribution was also estimated by a Cox model adjusted on the covariates taken into account in the model, instead of Kaplan–Meier estimation. Results were very similar to the ones presented here. The values were slightly lower than the ones presented for all models, and so, the conclusion was not modified.

Joint gamma frailty models are multiplicative models. This makes some reductions possible in probability calculation. In the end, no recurrence time was taken into account for the first setting and only the last recurrent event time had an influence in the second setting. The use of an additive model, such as the one proposed by Zeng *et al.* [9], may let all recurrence times have an influence in prediction. It could be a matter of further development to compare such results to the current ones.

Finally, we only explored the probability of death. However, it is also possible to use joint modeling to predict the risk of recurrent event along with the risk of death. The probability of both events (recurrence and death) could also be made using the evolution of a biomarker, as joint modeling of longitudinal data and recurrent events processes with a terminal event have been developed [14, 30].

Appendix

We calculate the probability of death between t and $t+w$ considering exactly J recurrences at known times $P^1(t,t+w;\xi)$. We use the complete recurrence history of the patient $\mathcal{H}_i^{J,1}(t) = \left\{N_i^R(t) = J, X_{i1} < \ldots < X_{iJ} \leqslant t\right\}$, with $X_{i0} = 0$ and $X_{i(J+1)} > t$.

We have

$$P^1(t,t+w;\xi) = P\left(D_i \leqslant t+w | D_i > t, \mathcal{H}_i^{J,1}(t), Z_{ij}^R, Z_i^D, \xi\right)$$
$$= \int_0^\infty P\left(D_i \leqslant t+w | D_i > t, \mathcal{H}_i^{J,1}(t), Z_{ij}^R, Z_i^D, u_i, \xi\right) \times g\left(u_i | D_i > t, \mathcal{H}_i^{J,1}(t), Z_{ij}^R, Z_i^D, \xi\right) \mathrm{d}u_i$$

We have first

$$P\left(D_i \leqslant t+w | D_i > t, \mathcal{H}_i^{J,1}(t), Z_{ij}^R, Z_i^D, u_i, \xi\right) = \frac{P\left(D_i \leqslant t+w, D_i > t | \mathcal{H}_i^{J,1}(t), Z_{ij}^R, Z_i^D, u_i, \xi\right)}{P\left(D_i > t | \mathcal{H}_i^{J,1}(t), Z_{ij}^R, Z_i^D, u_i, \xi\right)}$$

Using the conditional independence of the recurrent times, and of the death times with the recurrent times, conditionally on u_i, we obtain

$$\begin{aligned} P\left(D_i \leqslant t+w | D_i > t, \mathcal{H}_i^{J,1}(t), Z_{ij}^R, Z_i^D, u_i, \xi\right) &= \frac{P\left(D_i \leqslant t+w, D_i > t | Z_i^D, u_i, \xi\right)}{P\left(D_i > t | Z_i^D, u_i, \xi\right)} \\ &= \frac{S_i^D\left(t | Z_i^D, u_i, \xi\right) - S_i^D\left(t+w | Z_i^D, u_i, \xi\right)}{S_i^D\left(t | Z_i^D, u_i, \xi\right)} \end{aligned}$$

Secondly, we have

$$\begin{aligned} g\left(u_i | D_i > t, \mathcal{H}_i^{J,1}(t), Z_{ij}^R, Z_i^D, \xi\right) &= \frac{g\left(u_i, D_i > t, \mathcal{H}_i^{J,1}(t) | Z_{ij}^R, Z_i^D, \xi\right)}{P\left(D_i > t, \mathcal{H}_i^{J,1}(t) | Z_{ij}^R, Z_i^D, \xi\right)} \\ &= \frac{P\left(D_i > t, \mathcal{H}_i^{J,1}(t) | Z_{ij}^R, Z_i^D, \xi, u_i\right) g\left(u_i | Z_{ij}^R, Z_i^D, \xi\right)}{\int_0^\infty P\left(D_i > t, \mathcal{H}_i^{J,1}(t) | Z_{ij}^R, Z_i^D, \xi, u_i\right) g(u_i) \mathrm{d}u_i} \end{aligned}$$

In joint models, frailty terms are assumed to be independent of the covariates. Thus, $g(u_i | Z_{ij}^R, Z_i^D, \xi) = g(u_i)$. Using the independence of the recurrent times, and of the death times with the recurrent times, conditionally on u_i, we obtain

$$\begin{aligned} & g\left(u_i | D_i > t, \mathcal{H}_i^{J,1}(t), Z_{ij}^R, Z_i^D, \xi\right) \\ &= \frac{P\left(D_i > t | Z_i^D, \xi, u_i\right) P\left(\mathcal{H}_i^{J,1}(t) | Z_{ij}^R, \xi, u_i\right) g(u_i)}{\int_0^\infty P\left(D_i > t | Z_i^D, \xi, u_i\right) P\left(\mathcal{H}_i^{J,1}(t) | Z_{ij}^R, \xi, u_i\right) g(u_i) \mathrm{d}u_i} \\ &= \frac{S_i^D\left(t | Z_i^D, \xi, u_i\right) \prod_{k=1}^J \frac{\lambda_{ik}^R\left(X_{ik} | Z_{ij}^R, \xi, u_i\right) S_{ik}^R\left(X_{ik} | Z_{ij}^R, \xi, u_i\right)}{S_{ik}^R\left(X_{i(k-1)} | Z_{ij}^R, \xi, u_i\right)} \frac{S_{i(J+1)}^R\left(t | Z_{ij}^R, \xi, u_i\right)}{S_{i(J+1)}^R\left(X_{iJ} | Z_{ij}^R, \xi, u_i\right)} g(u_i)}{\int_0^\infty S_i^D\left(t | Z_i^D, \xi, u_i\right) \prod_{k=1}^J \frac{\lambda_{ik}^R\left(X_{ik} | Z_{ij}^R, \xi, u_i\right) S_{ik}^R\left(X_{ik} | Z_{ij}^R, \xi, u_i\right)}{S_{ik}^R\left(X_{i(k-1)} | Z_{ij}^R, \xi, u_i\right)} \frac{S_{i(J+1)}^R\left(t | Z_{ij}^R, \xi, u_i\right)}{S_{i(J+1)}^R\left(X_{iJ} | Z_{ij}^R, \xi, u_i\right)} g(u_i) \mathrm{d}u_i} \\ &= \frac{S_i^D\left(t | Z_i^D, \xi, u_i\right) \prod_{k=1}^J \lambda_{ik}^R\left(X_{ik} | Z_{ij}^R, \xi, u_i\right) S_{i(J+1)}^R\left(t | Z_{ij}^R, \xi, u_i\right) g(u_i)}{\int_0^\infty S_i^D\left(t | Z_i^D, \xi, u_i\right) \prod_{k=1}^J \lambda_{ik}^R\left(X_{ik} | Z_{ij}^R, \xi, u_i\right) S_{i(J+1)}^R\left(t | Z_{ij}^R, \xi, u_i\right) g(u_i) \mathrm{d}u_i} \\ &= \frac{S_i^D\left(t | Z_i^D, \xi, u_i\right) \prod_{k=1}^J \left[\lambda_0^R(X_{ik}) u_i \exp\left(\beta' Z_{ij}^R\right)\right] S_{i(J+1)}^R\left(t | Z_{ij}^R, \xi, u_i\right) g(u_i)}{\int_0^\infty S_i^D\left(t | Z_i^D, \xi, u_i\right) \prod_{k=1}^J \left[\lambda_0^R(X_{ik}) u_i \exp\left(\beta' Z_{ij}^R\right)\right] S_{i(J+1)}^R\left(t | Z_{ij}^R, \xi, u_i\right) g(u_i) \mathrm{d}u_i} \\ &= \frac{S_i^D\left(t | Z_i^D, \xi, u_i\right) (u_i)^J S_{i(J+1)}^R\left(t | Z_{ij}^R, \xi, u_i\right) g(u_i)}{\int_0^\infty S_i^D\left(t | Z_i^D, \xi, u_i\right) (u_i)^J S_{i(J+1)}^R\left(t | Z_{ij}^R, \xi, u_i\right) g(u_i) \mathrm{d}u_i} \end{aligned}$$

where $X_{i0} = 0$ and $S_{i0}^R(X_{i0} | Z_{ij}^R, \xi, u_i) = 1$.

We deduce from the two previous results

$$P^1(t,t+w;\xi) = \int_0^\infty \frac{S_i^D\left(t|Z_i^D,u_i,\xi\right) - S_i^D\left(t+w|Z_i^D,u_i,\xi\right)}{S_i^D\left(t|Z_i^D,u_i,\xi\right)} \times \frac{S_i^D\left(t|Z_i^D,\xi,u_i\right)(u_i)^J S_{i(J+1)}^R\left(t|Z_{ij}^R,\xi,u_i\right) g(u_i)}{\int_0^\infty S_i^D\left(t|Z_i^D,\xi,u_i\right)(u_i)^J S_{i(J+1)}^R\left(t|Z_{ij}^R,\xi,u_i\right) g(u_i)\mathrm{d}u_i}\mathrm{d}u_i$$

$$= \frac{\int_0^\infty \left[S_i^D\left(t|Z_i^D,u_i,\xi\right) - S_i^D\left(t+w|Z_i^D,u_i,\xi\right)\right](u_i)^J S_{i(J+1)}^R\left(t|Z_{ij}^R,\xi,u_i\right) g(u_i)\mathrm{d}u_i}{\int_0^\infty S_i^D\left(t|Z_i^D,\xi,u_i\right)(u_i)^J S_{i(J+1)}^R\left(t|Z_{ij}^R,\xi,u_i\right) g(u_i)\mathrm{d}u_i}$$

The results for probabilities $P^2(t,t+w;\xi)$ and $P^3(t,t+w;\xi)$ are derived similarly.

Acknowledgements

The authors thank Pippa McKelvie-Sebileau for medical editorial assistance and the Institut National du Cancer for funding this research.

References

1. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. J. Wiley: New York, 2002.
2. Proust-Lima C, Taylor JM. 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.
3. Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* 2011; **67**(3):819–829. DOI: 10.1111/j.1541-0420.2010.01546.x.
4. Cook R, Lawless J. *The Statistical Analysis of Recurrent Events*. Springer: New York, 2007.
5. Liu L, Wolfe RA, Huang X. Shared frailty models for recurrent events and a terminal event. *Biometrics* 2004; **60**(3):747–756. DOI: 10.1111/j.0006-341X.2004.00225.x.
6. Huang X, Liu L. A joint frailty model for survival and gap times between recurrent events. *Biometrics* 2007; **63**(2):389–397. DOI: 10.1111/j.1541-0420.2006.00719.x.
7. Rondeau V, Mathoulin-Pelissier S, Jacqmin-Gadda H, Brouste V, Soubeyran P. Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics* 2007; **8**(4):708–721. DOI: 10.1093/biostatistics/kxl043.
8. Rondeau V. Statistical models for recurrent events and death: application to cancer events. *Mathematical and Computer Modelling* 2010; **52**(7-8):949–955.
9. Zeng D, Cai J. A semiparametric additive rate model for recurrent events with an informative terminal event. *Biometrika* 2010; **97**(3):699–712.
10. Huang CY, Qin J, Wang MC. Semiparametric analysis for recurrent event data with time-dependent covariates and informative censoring. *Biometrics* 2009; **66**(1):39–49.
11. Zhangsheng Y, Liu L. A joint model of recurrent events and a terminal event with a nonparametric covariate function. *Statistics in Medicine* 2011; **30**(22):2683–2695. DOI: 10.1002/sim.4297.
12. Zhu L, Sun J, Srivastava DK. *et al.* Semiparametric transformation models for joint analysis of multivariate recurrent and terminal events. *Statistics in Medicine* 2011; **30**(25):3010–3023. DOI: 10.1002/sim.4306.
13. Zhao X, Liu L, Liu Y, Xu W. Analysis of multivariate recurrent event data with time-dependent covariates and informative censoring. *Biometrical Journal* 2012; **54**(5):585–599. DOI: 10.1002/bimj.201100194.
14. Kim S, Altman D, Chambless L, Li Y. Joint models of longitudinal data and recurrent events with informative terminal event. *Statistics in Biosciences* 2012:1–20. DOI: 10.1007/s12561-012-9061-x.
15. Li Y, Wileyto E, Heitjan D. Prediction of individual long-term outcomes in smoking cessation trials using frailty models. *Biometrics* 2011; **67**(4):1321–1329. DOI: 10.1111/j.1541-0420.2011.01578.x.
16. Rondeau V, Pignon JP, Michiels S. A joint model for the dependence between clustered times to tumour progression and deaths: a meta-analysis of chemotherapy in head and neck cancer. *Statistical Methods in Medical Research* 2011. DOI: 10.1177/0962280211425578. [In press].
17. Marquardt D. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics* 1963; **11**(2):431–441.
18. Gerds TA, Schumacher M. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal* 2006; **48**(6):1029–1040. DOI: 10.1002/bimj.200610301.
19. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* 2005; **21**(15):3301–3307. DOI: 10.1093/bioinformatics/bti499.
20. Klein JP, Keiding N, Copelan EA. Plotting summary predictions in multistate survival models: probabilities of relapse and death in remission for bone marrow transplantation patients. *Statistics in Medicine* 1993; **12**(24):2315–2332.

21. Dabrowska DM, Sun Gw, Horowitz MM. Cox regression in a markov renewal model: an application to the analysis of bone marrow transplant data. *Journal of the American Statistical Association* 1994; **89**(427):867–877.
22. Putter H, van der Hage J, de Bock GH, Elgalta R, van de Velde CJH. Estimation and prediction in a multi-state model for breast cancer. *Biometrical Journal* 2006; **48**(3):366–380. DOI: 10.1002/bimj.200510218.
23. Liquet B, Timsit JF, Rondeau V. Investigating hospital heterogeneity with a multi-state frailty model: application to nosocomial pneumonia disease in intensive care units. *BMC Medical Research Methodology* 2012; **12**(1):79.
24. Duchateau L, Janssen P. *The Frailty Model*. Springer: New York, 2008.
25. Louis T. Using empirical bayes methods in biopharmaceutical research. *Statistics in Medicine* 1991; **10**(6):811–829.
26. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. Springer: New York, 2009.
27. Komarek A, Hansen B, Kuiper E, van Buuren H, Lesaffre E. Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Statistics in Medicine* 2010; **29**(30):3267–3283.
28. Burnham KP, Anderson DR. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. Springer: New York, 2002.
29. Mazroui Y, Mathoulin-Pelissier S, Soubeyran P, Rondeau V. General joint frailty model for recurrent event data with a dependent terminal event: application to follicular lymphoma data. *Statistics in Medicine* 2012; **31**(11-12):1162–1176. DOI: 10.1002/sim.4479.
30. Liu L, Huang X. Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2008; **58**(1):65–81.

4.2.2 Additional remarks

On the misspecification of the frailty distribution

As suggested by an anonymous reviewer, two components are of importance when doing individual prediction with joint frailty models: the baseline hazard and the frailty distribution. The flexible form of the baseline hazard that we used prevents from misspecification. However, the gamma distribution that we chose for the frailty terms can have some impact on the accuracy of the calculated probabilities if this distribution is misspecified. In order to check this, we have performed a small simulation study. One hundred datasets of 500 subjects, having up to five recurrent events were generated. Two covariates were considered, and the frailty terms were simulated using a gamma distribution with mean one and variance $\theta = 1$. Baseline hazards were simulated using a Weibull distribution. Survival times were censored at time 0.8. Predictions were made at prediction time $t = 0.3$, chosen to let sufficient time to observe some recurrent events. Prediction horizon goes from time $t = 0.4$ to $t = 0.8$, every 0.1.

For each dataset, we estimated two models: one assuming a gamma distribution of the frailty terms, and one assuming a log-normal distribution of the frailty terms, and we calculated the corresponding predictions. We studied only the impact on the first setting $P^1(t, t + w; \xi)$, which is the most dependent on the frailty terms. Two graphs are given: figure 4.2 shows prediction error curves of the first 25 datasets (we show only 25 for readability purpose), and figure 4.3 shows how the difference between the error with the log-normal and the error with the gamma distribution range over the 96 datasets (100 generated minus 4 for which one of the two models did not converge). The prediction errors were very close with both estimations used, changing only at 10^{-3} scale (figure 4.3). This result suggests that the distribution of the frailty terms has only a minor impact on the estimated prediction over the population. This is in accordance with a previous paper, showing that in the framework of a joint frailty model for recurrent events and a terminal event, the considered distribution of the frailty terms has little impact on the parameter estimations (Mazroui et al., 2012).

On the consideration of different types of event and the multistate model

The main competing method to study the prognosis of breast cancer patients while accounting for intermediate events are multistate models as in Putter et al. (2006). Multistate models are models that allow patients to move between different states. In our application, the corresponding multistate model is as shown in figure 4.4 (Putter et al., 2006). At each prediction time, the probability of transiting to one state can be com-

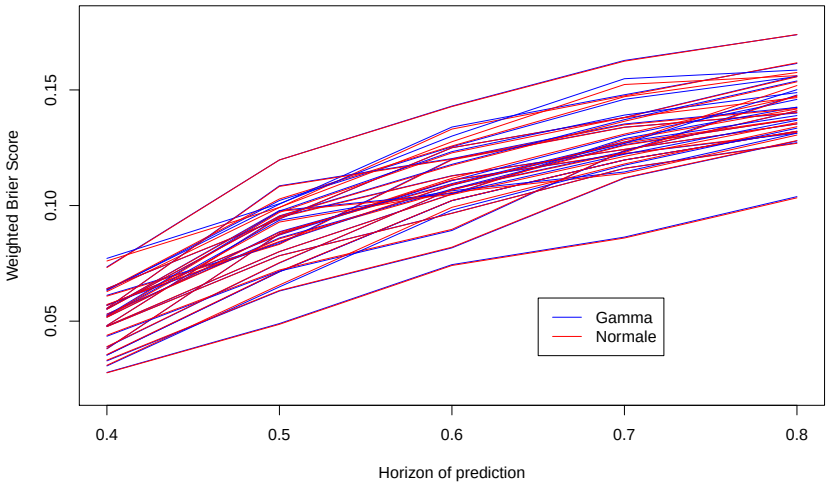

Figure 4.2: Simulation study results: prediction error for 25 randomly selected simulated datasets, analysed assuming either a gamma or a log-normal distribution for the frailty terms.

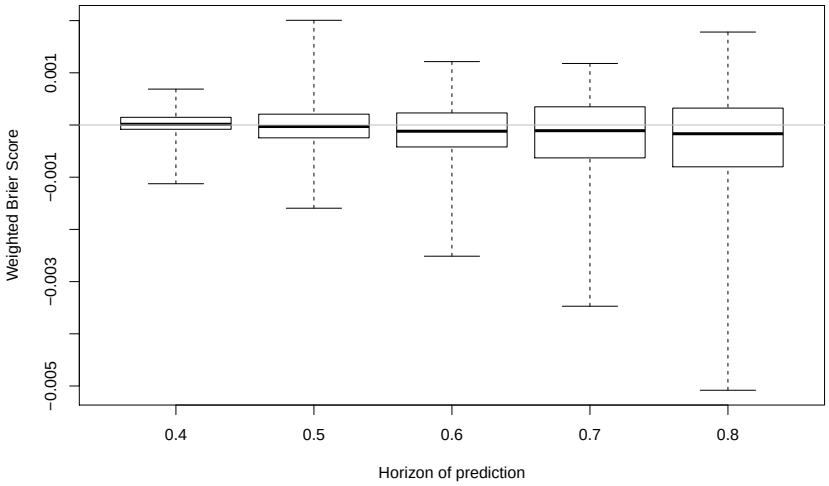

Figure 4.3: Simulation study results: average difference in prediction error between estimation assuming a log-normal and a gamma distribution of the frailty terms, among 96 simulated datasets.

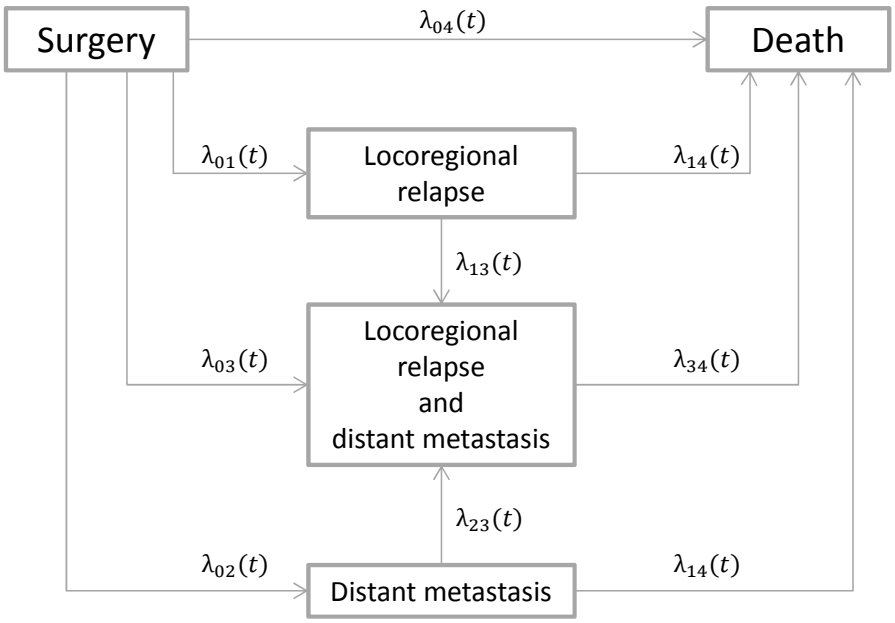

Figure 4.4: Illustration of the multistate model to study breast cancer survival (adapted from Putter et al. (2006)). The $\lambda_{lm}(t)$ represent the transition probability from state l to state m.

puted, especially the probability of transiting to death. This model requires nine transitions, i.e. nine survival models, to be estimated.

Multistate models are an appropriate and rather simple framework to do predictions of risk of death accounting for intermediate events in cancer patients. The main differences between multistate and joint models are as follows. While multistate model easily consider different types of intermediate events, they make it difficult to consider a recurrent event, even with four or five occurrences. Indeed, the number of states would be increased, increasing the number of parameters to be estimated. Moreover, the number of events can vary from one patient to another. It is thus possible to have very few patients going from one state to another, making difficult to estimate this transition. Most of the time, only the first occurrence of each event type is considered. The main advantage is that these different events are separately accounted for, both for the risk of event, and for the risk of death after a disease event. On the opposite, the joint model

is an appropriate framework to study the effect of an event that can occur many times. The main, strong, assumption is that the baseline hazard of event is the same, whatever the rank of the event. We did not consider yet different types of event in our prediction. The multivariate frailty model has recently been developed (Mazroui et al., 2013). The development of the corresponding predictions is one of the possible extension of the current work.

A careful comparison of the prediction performances of the two types of model in the breast cancer framework would be of great interest to be able to propose the best prediction tool to clinicians.

On the interval censoring and the time scale

The date of occurrence of the studied recurrent events, loco-regional relapses and distant metastases, are not exactly known. What we know is that the patient was free of relapse at the date of the previous visit, and has relapsed before the date of the last visit. Thus, what we actually study in our model is not the prognostic value of the time of the relapse itself, but of the time of the diagnosis of a relapse. However, the time elapsed between two hospital visits is usually not too long, and the patients having some symptoms may consult a clinician fast. The impact of not taking this interval-censoring is thus, to our thoughts, limited. Specific methods for interval-censored data in shared frailty models with dynamic prediction tools have been proposed in the context of clustered data (Rondeau et al., 2014, Submitted). To date, no joint model with interval-censored data has been developed to our knowledge.

On the relative interest of the three proposed prediction settings

In our developments, we proposed three prediction settings, which account for the patient recurrence history in three different ways. We initially proposed the second setting to be able to make some prediction from a time point in the future, without knowing exactly what will happen until then (partial information). However, its interest is limited as compared to the first setting. Also, it appears that these two settings give some very closed results, both in term of predicted probabilities and in term of prediction error. This is not surprising as the predictions differ only regarding a small part of the follow-up. Their values are equal if the prediction is done at the time of the last observed relapse. As a consequence, we chose to externally validate only the first prediction setting, assuming that the results could be extrapolated to the second one.

The third setting differs from the two others as it is an average prediction, which ignores the specific recurrence history of a given patient. In this sense, comparing the first and the third settings makes it possible to assess the usefulness of the recurrence history in predicting the risk of death. This is what is presented in the next section on external validation.

4.3 External validation

For the validation purpose, we used data from population similar in terms of disease but coming from different countries and from registries, i.e. from general population without selection. The description in terms of number of events and survival can be found in Table 4.1 in addition to description in the following article. We can see on Figure 4.5 that the survival is higher in the French population. This is due to the patient selection in this series: only patients followed and treated at one hospital (Bergonié comprehensive cancer centre) were included. The two other populations are general population based, meaning they include all the cases of cancer in a geographical area, irrespective of their gravity or treatment.

The model used was presented in the first article (see section 4.2.1). However, due to unavailability of some factors in the two registry populations: HER2 status and the hormonal receptors status (available in less than 5% of the patients), the model without these factors was validated (see article of the current section). This explains the slight differences between the model in the development and the validation articles. Moreover, in the Dutch data, information about peritumoural vascular invasion was not available. The model used for this population is presented in Appendix E.

4.3.1 Submitted publication

Table 4.1: Description of the number of events (disease relapses and deaths) in the three populations.

N events	0	1	2	3	4	All
French						
Alive	600	114	20	3	-	737
Died	105	187	37	1	-	330
All	705	301	57	4	-	1067
Dutch						
Alive	23072	840	1	-	-	23913
Died	4159	2994	9	-	-	7162
All	27231	3834	10	-	-	31075
English						
Alive	536	38	6	3	0	583
Died	359	202	43	7	2	613
All	895	240	49	10	2	1196

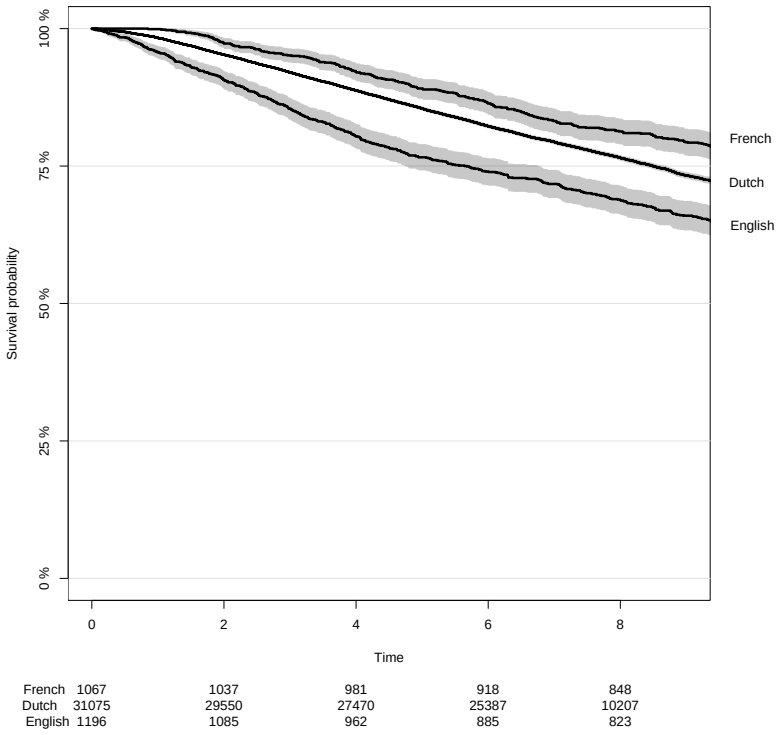

Figure 4.5: Overall survival in the three populations used to develop (French) and validate (Dutch and English) the prediction. Note that the confidence interval for the Dutch population is displayed, but tiny.

RESEARCH

Death prediction after breast cancer relapses using joint models: validation on English and Dutch population-based data and comparison with landmark approach.

Audrey Mauguen[1,2*], Bernard Rachet[2], Simone Mathoulin-Pélissier[3,4], Gill M Lawrence[5], Sabine Siesling[6], Gaëtan MacGrogan[3], Alexandre Laurent[1] and Virginie Rondeau[1]

*Correspondence: audrey.mauguen@isped.u-bordeaux2.fr
[1]Biostatistic unit, INSERM U897, ISPED, Université de Bordeaux, 146 rue Léo Saignat, 33076 Bordeaux Cedex, France
Full list of author information is available at the end of the article

Abstract

Background: Cancer relapses may be of great interest to predict the risk of death. To consider such information, the Landmark approach is popular. We propose as an alternative the joint frailty model for a recurrent event and a terminal event to derive dynamic predictions of the risk of death.

Methods: The proposed prediction settings can account or not for the relapse history. In this work, predictions developed on a French hospital series of patients with breast cancer are externally validated on UK and Netherlands registry data. The performances in term of prediction error and calibration are compared to those from a Landmark Cox model.

Results: The error of prediction was decreased when relapses information was considered. The prediction was well-calibrated, although it was developed and validated on really different populations. Joint modelling and Landmark approaches gave similar performance.

Conclusions: When predicting the risk of death, accounting for relapses led to better prediction performance. Joint modelling appeared to be suitable to do such prediction. It gave similar performances than a landmark Cox model, while directly quantifying the correlation between relapses and death.

Keywords: Breast cancer; Joint frailty model; Landmark; Prediction; Relapse history; Survival

Background

Individual predictions are more and more sought after to help guide treatment decisions and patient care. Accurate predictions are especially important in the framework of personalised medicine, where the ultimate goal is to give personalised targeted treatment to every patient. To do so, it is important to evaluate the prognosis of patients, accounting for their individual characteristics. In recent years, prognosis research in cancer has mainly focused on the presence of biomarkers that can be targeted by treatments. Less focus however has been given to the impact of relapses that patients may experience, such as loco-regional relapses or distant metastases in cancer patients which may explain a large part of the risk of death, despite adequate methods of analysis now available. The impact of these relapses

on the risk of death may vary according to the type of cancer. Relapses can be considered as a surrogate for the frailty of the patient or for the disease aggressiveness. It is of interest to therefore investigate how these events can be used to predict the survival of the patient.

Relapses are recurrent events, evolving during the follow-up of the patients. In cancer, because the relapses can only be measured if the patient is alive, they cannot be included in a survival model as a standard time-dependent covariate to study the risk of death. The Landmark approach [1, 2] gets around the problem by updating the population of interest. At each chosen prediction time, the model is performed again on the alive patients. Thus, the relapses history, observed before the prediction time, can be resumed as a baseline covariate, such as the number of previous relapses. This method has the advantage that the model used can be simple and robust. However, to do some dynamic predictions using several prediction times, several models have to be run on sub-populations of alive patients. Moreover, summarising the relapse process in a single variable may result in a loss of information.

Alternatively, joint modelling can be used to study recurrent event [3, 4]. The hazard of the recurrent event and the hazard of the terminal event (death) are jointly modelled. Such models allow to fully consider the correlation between the two processes using a shared random effect (frailty). Dynamic predictions can then be derived, accounting for all previous events. Once the parameters are estimated, predictions can be updated without running the model again. A recent work investigated the impact of relapses on the risk of death in breast cancer using joint frailty model [5]. The proposed method was shown to be an adequate framework for prediction, and the model seemed to give satisfying performance on the population used to develop the model (development population).

However, in the prognosis research, it is generally agreed that several steps are required to be able to propose an accurate and reliable prediction rule [6–8]. These steps are as follows: 1) identification of potentially relevant predictors, 2) choice of an adequate modelling method, 3) selection of predictors and estimation of their effects (model development) and 4) application of the model on a new population and assessment of the prediction ability (model validation) [9, 10].

Here, we validate a prediction of the risk of death accounting for the cancer relapses, such as loco-regional relapses or distant metastases. We present the results of the external validation of the previously developed tool [5]. We also compare this new tool's performance with that of the Landmark Cox model. As in the development step, we apply the proposed prediction on breast cancer cases, here from two population-based registries, in West Midlands (England) and the Netherlands.

Section 2 of this paper explains the prediction probabilities within the framework of a Landmark Cox model and a frailty joint model, and the tools to validate them. The validation of the prediction on the West Midlands and Netherlands registry datasets is developed in section 3. Finally, section 4 contains discussion and conclusion elements.

Methods

We are interested in the prediction of the risk of death between a prediction time s and a prediction horizon $s + w$ considering all the information available at time

s. The information includes some baseline covariates, but also history of recurrent event (loco-regional relapse or distant metastasis) until time s. In this context, the predicted risk of death can be updated after each new recurrence.

Prediction of death using a Landmark Cox model

The Landmark approach entails fixing a prediction time s and fitting the model on the sub-population of patients still at risk of death at this time, that is patients alive and not lost to follow-up [1]. Thus, the number of relapses having occurred before time s can be treated as a baseline covariate, and the recurrences occurring after s are ignored. This covariate can be updated when another Landmark time s is considered, and a new model is fitted. With this approach, a robust model can be used, requiring few parameters, and the time-dependent effects are easily dealt with. However, only a sub-population of alive patients is included to fit the model, which may result in a loss of information in the parameter estimation.

Let $\lambda^D_{s,i}(.|Z^D_{s,i})$ denote the hazard function of death conditional on being alive at the Landmark time s, $\lambda^D_{s,0}(.)$ be the baseline hazard function, $Z^D_{s,i}$ be the covariate vector at time s and β_s be their effect estimated at time s. The Landmark Cox model is then written as follows:

$$\lambda^D_{s,i}(t|Z^D_{s,i}) = \lambda^D_{s,0}(t)\exp(\beta'_s Z^D_{s,i}), \text{ for } t \geq s$$

This model is estimated with the information available at the Landmark time s. The prognostic factors of interest $Z^D_{s,i}$ may include information about previous recurrent events, for example, their frequency and timing. The baseline hazard of death $\lambda^D_{s,0}(.)$ is estimated using splines and the parameters of the model $\xi^{LM}_s = (\lambda^D_{s,0}(.), \beta_s)$ are estimated using penalized maximum likelihood estimators as in [4]. The corresponding prediction of the risk of death is written as:

$$\begin{aligned} & P^{LM}(s, s+w; \xi^{LM}_s) \\ & = P(D_i \leq s+w | D_i > s, Z^D_{s,i}, \xi^{LM}_s) \\ & = \frac{S^D_i(s|Z^D_{s,i}, \xi^{LM}_s) - S^D_i(s+w|Z^D_{s,i}, \xi^{LM}_s)}{S^D_i(s|Z^D_{s,i}, \xi^{LM}_s)} \end{aligned} \quad (1)$$

where $S^D_i(.|Z^D_{s,i}, \xi^{LM}_s)$ is the survival function associated to the hazard of death.

Prediction of death in joint modelling framework

The frailty joint model for a recurrent event and a terminal event is defined as follows: for subject i ($i = 1, ..., N$), let X_{ij} be the j^{th} recurrent time ($j = 1, ..., n_i$), C_i be the censoring time (not by death) and D_i be the death time. $T^R_{ij} = min(X_{ij}, C_i, D_i)$ corresponds to each follow-up time and δ^R_{ij} is a binary indicator for recurrent events which is 0 if the observation is censored or if the subject died and 1 if X_{ij} is observed ($\delta^R_{ij} = I[T^R_{ij} = X_{ij}]$ where $I[.]$ denotes indicator function). Similarly, we note T^D_i the last follow-up time for subject i, which is either a time of censoring or a time of death ($T^D_i = min(C_i, D_i)$) and $\delta^D_i = I[T^D_i = D_i]$. We actually observe the sequence $(T^R_{ij}, \delta^R_{ij}, T^D_i, \delta^D_i)$. Finally, we denote by Z^R_{ij} and Z^D_i the vectors of covariates associated with the hazard of recurrent events and death, respectively. Both death and recurrent times are in the calendar timescale, that is,

measured by the time elapsed since the origin of the study. However, a patient is considered at risk of a j^{th} recurrence only after the $(j-1)^{st}$ recurrence. The joint model is then written as:

$$\begin{cases} \lambda_{ij}^R(t|u_i) = u_i \lambda_0^R(t) \exp(\beta_1' Z_{ij}^R) = u_i \lambda_{ij}^R(t) \\ \lambda_i^D(t|u_i) = u_i^\alpha \lambda_0^D(t) \exp(\beta_2' Z_i^D) = u_i^\alpha \lambda_i^D(t) \end{cases} \tag{2}$$

where $\lambda_0^R(.)$ is the baseline hazard of a recurrent event, irrespective of event rank, and $\lambda_0^D(.)$ the baseline hazard of death. The effects of explanatory variables β_1 and β_2 are assumed to be different for the hazard of recurrent events and the risk of death. The two processes are linked by the patient-specific frailty effect u_i. The frailty terms are independent and identically distributed following a gamma distribution with variance θ and, without loss of generality, a mean equal to 1. That is:

$$u_i \sim \text{Gamma}(\frac{1}{\theta}; \frac{1}{\theta}) \quad \text{and} \quad g(u_i) = \frac{u^{1/\theta - 1} \exp(-u/\theta)}{\theta^{1/\theta} \Gamma(1/\theta)} \tag{3}$$

The baseline hazard functions ($\lambda_0^R(.)$ for recurrent events and $\lambda_0^D(.)$ for death) are approximated using splines. The parameters of the model $\xi^J = (\lambda_0^R(.), \lambda_0^D(.), \beta_1, \beta_2, \alpha, \theta)$ are estimated using penalized maximum likelihood estimators. For more details on the inference method, please see [4]. As opposed to the Landmark Cox model, this model is estimated using the covariates observed at time 0.

Using the joint modelling framework, we are interested in two prediction settings previously defined [5]. The first prediction P^{Rec} is calculated based on all relapses information. In this setting, the J relapses occurring before the prediction time s are considered. We consider the patient history $\mathcal{H}_i^J(s) = \{N_i^R(s) = J, X_{i1} < \ldots < X_{iJ} \leq s\}$, with $X_{i0} = 0$ and $X_{i(J+1)} > s$, to define the conditional probability of death P^{Rec} as follows:

$$\begin{aligned} & P^{Rec}(s, s+w; \xi^J) \\ &= P(D_i \leq s+w | D_i > s, \mathcal{H}_i^J(s), Z_{s,ij}^R, Z_{s,i}^D, \xi^J) \\ &= \frac{\int_0^\infty [S_i^D(s|Z_{s,i}^D, u_i, \xi^J) - S_i^D(s+w|Z_{s,i}^D, u_i, \xi^J)](u_i)^J S_{i(J+1)}^R(s|Z_{s,ij}^R, u_i, \xi^J) g(u_i) \mathrm{d}u_i}{\int_0^\infty S_i^D(s|Z_{s,i}^D, u_i, \xi^J)(u_i)^J S_{i(J+1)}^R(s|Z_{s,ij}^R, u_i, \xi^J) g(u_i) \mathrm{d}u_i} \end{aligned} \tag{4}$$

where $Z_{s,ij}^R$ and $Z_{s,i}^D$ are the values of the covariates at time s, and $g(u_i)$ is the density of the gamma distribution defined in equation (3).

The second setting P^{Ign} also uses the joint modelling framework. However, the information about previous recurrences is not considered in the prediction, and it can be missing. It is defined by:

$$
\begin{aligned}
&P^{Ign}(s, s+w; \xi^J) \\
&= P(D_i \leq s+w | D_i > s, Z^D_{s,i}, \xi^J) \\
&= \frac{\int_0^\infty [S^D_i(s|Z^D_{s,i}, u_i, \xi^J) - S^D_i(s+w|Z^D_{s,i}, u_i, \xi^J)] g(u_i) \mathrm{d}u_i}{\int_0^\infty S^D_i(s|Z^D_{s,i}, u_i, \xi^J) g(u_i) \mathrm{d}u_i}
\end{aligned} \tag{5}
$$

Both settings are dynamic in the sense that the prediction can be updated by changing the prediction time s, thus the quantity of available information, and/or the prediction window w. The first setting considers the individual relapses history whereas the second ignores it.

External validation of the prediction

In order to do predictions using the three proposed settings P^{LM}, P^{Rec} and P^{Ign} on new patients, the model parameters are estimated on the development population. Based on these estimators, predictions for new patients are obtained by replacing the patient level information, $Z^D_{s,i}$ in equation (1) or J, $Z^R_{s,ij}$ and $Z^D_{s,i}$ in equations (4) and (5), with actual information on the new patients.

The quality of fit of the two models can be compared on the development data using an approximate likelihood cross-validation criterion as in [12].

Prediction error

To estimate if the predictions are accurate, error of prediction curves are used, based on the Brier score. The Brier score aims at measuring how far the prediction is from the actual outcome of the patients. We used a weighted estimator of the Brier score to account for the right censoring using Inverse Probability of Censoring Weights (IPCW) [13].

Let N_s be the number of patients alive and uncensored at prediction time s, that is, patients for whom the prediction can be made. Given T^D_i the right censored survival time, δ^D_i the corresponding event indicator ($\delta^D_i = 1$ if the observed time is a death time, 0 otherwise) and $\widehat{G}_N(.)$ the Kaplan-Meier estimate of the censoring distribution in the population, and using the generic term $\hat{P}(s, s+w; \hat{\xi})$ which can be each of the three prediction probability settings previously described, the error of prediction is defined by:

$$
Err_{s,w} = \frac{1}{N_s} \sum_{i=1}^{N_s} [I(T^D_i > s+w) - (1 - \hat{P}(s, s+w; \hat{\xi}))]^2 \hat{w}_i(s+w, \hat{G}_N(.))
$$

with $\hat{w}_i(s+w, \hat{G}_N(.))$ being a weight that accounts for right censoring:

$$
w_i(s+w, \hat{G}_N(.)) = \frac{I(T^D_i \leq s+w)\delta^D_i}{\hat{G}_N(T^D_i)/\hat{G}_N(s)} + \frac{I(T^D_i > s+w)}{\hat{G}_N(s+w)/\hat{G}_N(s)}
$$

The performance of the models is compared using a measure of explained residual variation defined as follows [14]:

$$R^2 = 1 - \frac{Err_{s,w}}{Err_{s,w;KM}}$$

where $Err_{s,w}$ is the error of one of the predictions (P^{LM}, P^{Rec} or P^{Ign}) described as above and $Err_{s,w;KM}$ is the error of prediction using the Kaplan-Meier estimate at $s+w$ in the entire set of patients. It can be interpreted as how much the prediction error is decreased using the model-based prediction as compared to an average prediction estimated by Kaplan-Meier.

The proposed error of prediction is calculated in two different ways: either s is fixed and w varies, or s varies while w is fixed.

Calibration plot

Another indicator of the accuracy of the prediction tool proposed is the calibration. A well-calibrated prediction means that, among 100 patients with a predicted event risk of p%, p of them will actually experience the event. This can be done only for a binary endpoint, meaning we must choose a time of prediction. We held it at $s + w = 10$ years.

The calibration is illustrated using a calibration plot. The predicted risks of death are grouped according to the deciles of their distribution. For each decile, the observed proportion of event is plotted against the mean predicted value, along with the 95% confidence interval for the observed proportion. For a well-calibrated prediction, all points should fall very close to the first bisector.

On the calibration plot the histogram of the predicted values is also represented, showing how they are distributed between 0 and 1.

Results

Population comparison

The first validation population consisted of all breast cancer cases diagnosed in West Midlands, England, in 1996 and followed up to 2012. The second validation population consisted of cases from the Netherlands Cancer Registry, South Netherlands region excluded, diagnosed between 2003 and 2006 and followed up to the end of 2012. In these two populations, we assume that there is no lost to follow-up. The development cohort consisted of 1,067 patients operated in a comprehensive cancer centre between 1989 and 1993, and with a median follow-up of 14 years. Thus the two validation populations differ from the development population in terms of country and selection of population (general population in West Midlands and Netherlands; hospital-based patients in France) and inclusion period (1996 and 2003-2006 vs 1989-1993).

In the two validation populations, a high rate of missing data was observed: out of the 3168 cases recorded in the year 1996 in West Midlands, 1,196 (38%) had non-missing values for all of the five studied prognostic factors. In the Dutch population, information about peritumoural vascular invasion was not recorded. Of the 41,676 recorded patients, 31,075 (75%) had non-missing values for the four remaining factors. In our validation population, we included only patients with complete

information in the two datasets. This decision is discussed in the last part of the paper.

Table 1 compares the repartition of the prognosis factors in the three populations, as well as the number of relapses per patient, and the overall survival. The patients from both validation populations had more severe disease, i.e., more often a peritumoural vascular invasion (38.5% in West Midlands versus 26.7%), a tumour size greater than 20 mm (46.8% in West Midlands and 39.8% in Netherlands versus 22.7%) and grade III disease (37.1% in West Midlands and 33.8% in Netherlands versus 24.6%), despite a similar age. As a result, overall survival in both West Midlands and Netherlands was lower than in the development cohort.

The number of relapses per patient also varied. There were up to four relapses recorded in the West Midlands population compared with a maximum of three in the Dutch population and the development cohort. In the West Midlands registry, relapses were not collected but retrieved from the treatment information with an algorithm that uses the treatment type and time interval between successive treatments [15]. In the Dutch population, relapse data were obtained directly from patient files; both clinically and pathologically confirmed relapses were recorded. The recording was limited to relapses occurring during the first five years after diagnosis and, in some regions, to the first relapse of each type (local relapse, regional relapse or distant metastasis). In the French cohort, the relapses (loco-regional recurrence or distant metastasis) were recorded following a clinical examination. That resulted in 75% of the patients without registered relapse in the West Midlands, 88% in the Dutch population, and 66% in the French cohort.

Validation of the prediction

Models

The results of the joint frailty models, and the Landmark Cox model (thereafter called Landmark model), estimated on the French data are shown in Table 2. The prognostic factors kept for prediction were those associated with the risk of recurrent event or with the risk of death in the joint model. The Landmark model provided lower estimated effects than with the joint model. The Landmark model also showed an important effect from the number of previous relapses. The likelihood cross-validation criterion was lower for the Landmark model, suggesting that this model fitted the data better than the joint model. However, a better fit does not necessarily result in a better prediction.

Prediction error

Overall, all of the studied prediction settings gave better results than with the Kaplan-Meier, with a higher R^2 for both predictions accounting for the relapses (Figure 1). In West Midlands, when the time of prediction s is at five years (Figure 1-B), R^2 was as high as 80% for early predictions and regularly decreased with increasing prediction horizon (30% at 10 years). The gain in the prediction error diminished with the prediction horizon, being around 50% at seven years and ending around 17% at 15 years, showing that the information from the model had a higher impact on short-term prediction. R^2 was still around 20% at horizon 15 years, but very similar for the three settings. R^2 was low, under 20%, for shorter prediction

time ($s = 2$ years; see Figure 1-A). This illustrates the fact that the information gathered up to two years was not enough to obtain good prediction, especially considering relapses. In the Dutch population (Figures 1-C and D), the limited follow-up prevented us from studying prediction times longer than three years. The results were very similar to those in West Midlands population at $s = 2$ years, but the difference between the three settings was larger. Results were better for the prediction from the Landmark model, and worse for the prediction ignoring the relapse information.

When holding the prediction window at $w = 2$ or $w = 5$ years, results were similar (Figures 2-A and B, respectively). The setting ignoring relapses always gave lower performance, while the performances of the two other settings were very similar. The more information that was collected and used, the more accurate the prediction was, as showed by the curves increasing with time of prediction t, for both window times. As expected, the entire curve was higher (i.e. lower error of prediction) when the prediction was made for a shorter window (two years as compared to five years).

Calibration

All three settings gave good calibration, with points around the first bisector (Figure 3). Interestingly, both prediction approaches accounting for relapses identified a group of patients with high risk of death in both populations. For these high-risk patients, the mean predicted risk was somewhat lower than the observed risk (40% versus 50% in West Midlands and in the Netherlands using P-Recurrence). The histograms show that predicted values were lower overall for the Landmark approach (rarely exceeding 20%) whereas both predictions from the joint model gave some higher risks. This may explain why the observed probability of death seemed a little underestimated with the Landmark approach in West Midlands.

Additional validation by subgroups

The populations used to validate the prediction differed in many aspects from the development population. Thus, even if good results are observed for the proposed prediction both in terms of prediction error and of calibration, it is crucial to check the accuracy of the prediction on a more similar population. Here we selected a subpopulation of operated patients, as it was the main selection criterion in the development population.

Similar results were observed with the 602 operated patients from the West Midlands (Figure 4). Large confidence intervals were observed due to the reduced number of subjects included in this analysis (n=417 patients alive at five years). Calibration was not better than the one observed on the entire West Midlands data.

A second subgroup analysis was performed comparing the performance of the proposed predictions between subjects who relapsed at least once before the prediction time of five years and those who did not (Figure 5). As expected, in the subgroup of patients without relapse, the results were very similar to those in the entire population. However, no high-risk subject groups were identified, as observed in the main analysis. In the population with relapses, the prediction ignoring the relapses underestimated the observed probabilities of events (all the points are above the line) and had a really low R^2, even negative after 7.5 years.

Conclusion and discussion

The present work shows how recurrent events occurring in breast cancer patients may be used to obtain accurate prediction of death. The resulting calibration and error of prediction show that the proposed model is useful to predict the risk of death, in particular when enough variability in the number of recurrences is observed. Good calibration was obtained, especially considering the validation populations differed from the development population with respect to inclusion criteria for the patients and period, country and therefore, health care system. Using different incidence years is of great interest, since the care (especially treatment and screening) of breast cancer patients evolved during the 1990s, influencing the patients' survival. It finally seems that despite these differences, the effect of covariates and relapses remained similar and was still of interest.

Initially, the proposed prediction incorporated the information about the human epidermal growth factor receptor 2 (HER-2) status and hormonal receptors status. However, considering the non-availability of this information at the general population level at the time of this data collection, we have re-estimated the model and prediction without this information. On the initial development population, we compared the prediction error of the two joint models, with and without this biological information. The prediction error was very similar for both models (data not shown).

Prognostic research literature is scant when it comes to the consequences of missing data on the validation process, i.e. after the development phase. No prediction can be done if one of the covariates is missing. However, the impact of such exclusion on the validation process remains unclear. Multiple imputation has proved to be useful approach for model estimation (e.g. [16]), and could also be used for the validation stage. However, the benefit of such imputation to estimate the model performance is uncertain. To reproduce the conditions of the clinical practice, we keep in our validation only the patients with complete information. The subjects with missing data were more likely to be 55 years and older, and to have more nodal involvement, for similar stage and tumour size (data not shown). As the predictions that take into account the relapses information showed to be more appropriate in predicting high risk of death, it is possible that the performances of the prediction accounting for the relapses were underestimated. Finally, the survival results of the analysed patients were in accordance with the results of the EUROCARE-4 study [17]. In this study, the age-adjusted 5-year survival was 81% for French patients diagnosed between 1990 and 1994, and 78% and 83% for the patients diagnosed between 1995 and 1999 in England and Netherlands, respectively.

To account for the recurrent events and their association with the risk of death, we used frailty joint model for recurrent events and a terminal event. This framework appears suitable to derive such predictions. The obtained prediction gave similar error of prediction to a Cox model in a Landmark approach. This result illustrates the fact that a more complex model can be used to derive a useful prediction, avoiding the overfitting problem normally associated with a high number of parameters to estimate. As compared to the Landmark approach, the joint modelling has the advantage that all patients are used to derive predictions at all times. In the Landmark approach, only alive patients are taken into account. This may lead to small

populations in late prediction times, and thus less accurate prediction. However, it normally requires a simpler, thus more stable, model. The Landmark model makes it possible to fully model the recurrent event process. In the Landmark approach, a modelling choice has to be made about the previous events: using the number of previous events (continuous or not) and/or using the time of the previous events (linear or not). The prediction error seems similar between both approaches. So the choice between the joint and the Landmark approach should be mainly guided by the willingness (i) to fully describe both processes (recurrent event and death) and their correlation (joint model); or (ii) to focus only on the death (Landmark approach). A Landmark approach using a non-parametric prediction was also recently proposed to predict a long-term outcome accounting for a short-term event [18].

In the context of the joint modelling framework, the prediction of recurrent events can also be derived. Moreover, it appeared in this work that each prognostic factor considered separately adds very little prediction information once the baseline hazard and recurrent event processes are adequately modelled (data not shown). The covariates may be of higher interest when predicting the risk of a recurrent event. Considering the type of relapse differently (loco-regional relapse and metastasis) can also be of interest as they reflect various severity levels of the disease [19]. Finally, these predictions could be extended in the framework of competing risks or excess of mortality, where it would be possible to focus on predicting only the risk of death from the cancer.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

AM, BR and VR designed the study, interpret the data and drafted the manuscript. SMP, GL, SS and GMG collected and shared the data. AM performed the statistical analyses. AL contributed to the software development. All authors reviewed the manuscript and approved the final version.

Ethical approval and availability of supporting data

For the french hospital series, ethical approval from the national ethics committee (Commission Nationale de l'Informatique et des Libertés) was obtained for this study, which allowed the use of data recorded in this clinical and pathological database. In this comprehensive cancer center, each patient was informed that medical data can be use in observational research. The procedure follows the French law for medical research.

All data used are confidential. Researchers may access the data by sending a formal request to the appropriate institution (Institut Bergonié for the French series, West Midlands Cancer Intelligence Unit for the UK registry data and Comprehensive Cancer Centre The Netherlands for the Dutch registry data).

Acknowledgements

This research was funded by the Institut National du Cancer and BR is funded by a Cancer Research UK Program grant (C1336/A11700). We thank Catherine Lagord, Jackie Walton and Christopher Lawrence from the West Midlands Cancer Intelligence Unit, and Mirian Brink from the Comprehensive Cancer Centre The Netherlands (IKNL), for their help with the data preparation and sharing. We thank Tsion Solomon for editorial review.

Author details

[1]Biostatistic unit, INSERM U897, ISPED, Université de Bordeaux, 146 rue Léo Saignat, 33076 Bordeaux Cedex, France. [2]Cancer Research UK Cancer Survival Group, Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, WC1E 7HT London , UK. [3]Clinical epidemiology and research, Institut Bergonié, 229 Cours de l'Argonne, 33000 Bordeaux, France. [4]INSERM CIC-EC7, ISPED, Université de Bordeaux, 146 rue Léo Saignat, 33076 Bordeaux Cedex, France. [5]West Midlands Cancer Intelligence Unit, 5, St Philip's Place, B3 2PW Birmingham, UK. [6]Comprehensive Cancer Centre The Netherlands (IKNL), Godebaldkwartier 419 ingang Janssoenborch, 3511 Utrecht, The Netherlands.

References

1. Van Houwelingen, H.: Dynamic prediction by landmarking in event history analysis. Scandinavian Journal of Statistics **34**(1), 70–85 (2007). doi:10.1111/j.1467-9469.2006.00529.x
2. Van Houwelingen, H., Putter, H.: Dynamic Prediction in Clinical Survival Analysis, 1st edn. CRC Press Inc, Boca raton (2011)
3. Liu, L., Wolfe, R., Huang, X.: Shared frailty models for recurrent events and a terminal event. Biometrics **60**(3), 747–756 (2004). doi:10.1111/j.0006-341X.2004.00225.x
4. Rondeau, V., Mathoulin-Pelissier, S., Jacqmin-Gadda, H., Brouste, V., Soubeyran, P.: Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. Biostatistics **8**(4), 708–721 (2007). doi:10.1093/biostatistics/kxl043
5. Mauguen, A., Rachet, B., Mathoulin-Pelissier, S., MacGrogan, G., Laurent, A., Rondeau, V.: Dynamic prediction of risk of death using history of cancer recurrences in joint frailty models. Statistics in Medicine **32**(30), 5366–5380 (2013). doi:10.1002/sim.5980
6. Moons, K., Royston, P., Vergouwe, Y., Grobbee, D., Altman, D.: Prognosis and prognostic research: what, why, and how? BMJ **338**(1), 375–375 (2009). doi:10.1136/bmj.b375
7. Moons, K.G.M., Kengne, A.P., Woodward, M., Royston, P., Vergouwe, Y., Altman, D.G., Grobbee, D.E.: Risk prediction models: I. development, internal validation, and assessing the incremental value of a new (bio)marker. Heart **98**(9), 683–690 (2012). doi:10.1136/heartjnl-2011-301246
8. Riley, R., Hayden, J., Steyerberg, E., Moons, K., Abrams, K., Kyzas, P., et al.: Prognosis research strategy (PROGRESS) 2: Prognostic factor research. PLoS Med **10**(2), 1001380 (2013). doi:10.1371/journal.pmed.1001380
9. Altman, D., Royston, P.: What do we mean by validating a prognostic model? Statistics in Medicine **19**(4), 453–473 (2000)
10. Konig, I., Malley, J., Weimar, C., Diener, H., Ziegler, A.: Practical experiences on the necessity of external validation. Statistics in Medicine **26**(30), 5499–5511 (2007). doi:10.1002/sim.3069
11. Moons, K., Kengne, A., Grobbee, D., Royston, P., Vergouwe, Y., Altman, D., Woodward, M.: Risk prediction models: II. external validation, model updating, and impact assessment. Heart **98**(9), 691–698 (2012). doi:10.1136/heartjnl-2011-301247
12. Rondeau, V., Pignon, J.-P., Michiels, S.: A joint model for the dependence between clustered times to tumour progression and deaths: A meta-analysis of chemotherapy in head and neck cancer. Stat Methods Med Res **897**, 1–19 (2011). doi:10.1177/0962280211425578
13. Gerds, T., Schumacher, M.: Consistent estimation of the expected brier score in general survival models with right-censored event times. Biometrical Journal **48**(6), 1029–1040 (2006). doi:10.1002/bimj.200610301
14. Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M.: Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine **18**(17-18), 2529–2545 (1999). doi:10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5
15. Walton, J., Lagord, C., Lawrence, C., Lawrence, G.: The development of an algorithm to identify breast cancer recurrences., Brighton. NCIN Cancer Outcomes Conference 2013. Brighton, June 12th-14th, 2013. [Poster 121, url: http://www.ncin.org.uk/news_and_events/conferences/2013_posters, consulted on December 2013] (2013)
16. Vergouwe, Y., Royston, P., Moons, K., Altman, D.: Development and validation of a prediction model with missing predictor data: a practical approach. Journal of Clinical Epidemiology **63**(2), 205–214 (2010). doi:10.1016/j.jclinepi.2009.03.017
17. Berrino, F., De Angelis, R., Sant, M., Rosso, S., Lasota, M., Coebergh, J., Santaquilani, M.: Survival for eight major cancers and all cancers combined for european adults diagnosed in 1995-99: results of the EUROCARE-4 study. The Lancet Oncology **8**(9), 773–783 (2007). doi:10.1016/S1470-2045(07)70245-0. Accessed 2014-02-14
18. Parast, L., Cheng, S., Cai, T.: Incorporating short-term outcome information to predict long-term survival with discrete markers. Biometrical Journal **53**(2), 294–307 (2011)
19. Mazroui, Y., Mathoulin-Pelissier, S., MacGrogan, G., Brouste, V., Rondeau, V.: Multivariate frailty models for two types of recurrent events with a dependent terminal event: Application to breast cancer data. Biometrical Journal **55**(6), 866–884 (2013). doi:10.1002/bimj.201200196

Table 1 Description of the three populations used to develop (n=1067) and validate the model (n=3168 and n=31,075)

	French cohort (1989-1993)		West Midlands (1996)		Netherlands (2003-2006)	
Variable	N=1067	%	N=1196	%	N=31075	%
Age						
Age $\leq$40	82	7.7	73	6.1	2126	6.8
Age]40-55]	391	36.6	456	38.1	10681	34.4
Age >55	594	55.7	667	55.8	18268	58.8
Peritumoural vascular invasion	285	26.7	460	38.5	-	-
Tumour size $>$ 20 mm	242	22.7	560	46.8	12365	39.8
Nodal involvement	451	42.3	496	41.5	12588	40.5
Grade						
Grade I	316	29.6	226	18.9	6565	21.1
Grade II	488	45.7	526	44.0	13993	45.0
Grade III	263	24.6	444	37.1	10517	33.8
Number of recurrent event						
None	705		895		2248	
1	301		240		27042	
2	57		49		1018	
3	4		10		458	
4	0		2		209	
$\geq$ 5	0		0		100	
5-year survival	89.1	(87.3-91.0)	76.6	(74.2-79.0)	85.5	(85.1-85.9)
10-year survival	77.1	(74.6-79.7)	63.1	(60.5-65.9)	-	
15-year survival	65.4	(62.2-68.2)	51.6	(48.8-54.5)	-	

Table 2 Joint and Landmark Cox models estimations on the French cohort (n=1067 patients, 427 recurrent events)

	For recurrent events		**For death**		**For death - Cox Landmark**	
Variable	HR	(95% CI)	HR	(95% CI)	HR	(95% CI)
Age						
$]40-55]$ *vs* >55	1.17	(0.91-1.51)	0.31	(0.16-0.60)	0.56	(0.41-0.76)
≤ 40 *vs* >55	2.41	(1.73-3.37)	1.57	(0.73-3.38)	0.54	(0.31-0.92)
Peritumoural vascular invasion	1.61	(1.26-2.06)	4.74	(2.54-8.85)	1.04	(0.76-1.43)
Tumour size (>20 mm vs ≤ 20 mm)	1.95	(1.52-2.50)	6.21	(2.99-12.86)	1.20	(0.88-1.65)
Nodal involvement	1.84	(1.44-2.36)	4.89	(2.47-9.67)	1.95	(1.45-2.60)
Grade						
II vs I	2.18	(1.57-3.01)	7.48	(2.71-20.66)	1.07	(0.75-1.52)
III vs I	3.09	(2.16-4.41)	44.33	(15.61-125.93)	1.25	(0.83-1.88)
Recurrences before $t=5$ years						
One previous recurrence					7.18	(5.25-9.83)
Two previous recurrences					6.94	(3.05-15.83)
$\theta = var(u_i)$	1.07	(se=0.06)				
α	4.45	(se=0.33)				
LCV	1.19				0.93	

HR: Hazard ratio; CI: Confidence interval; LCV: Likelihood cross-validation criterion
Cox Landmark at time $t=5$ years

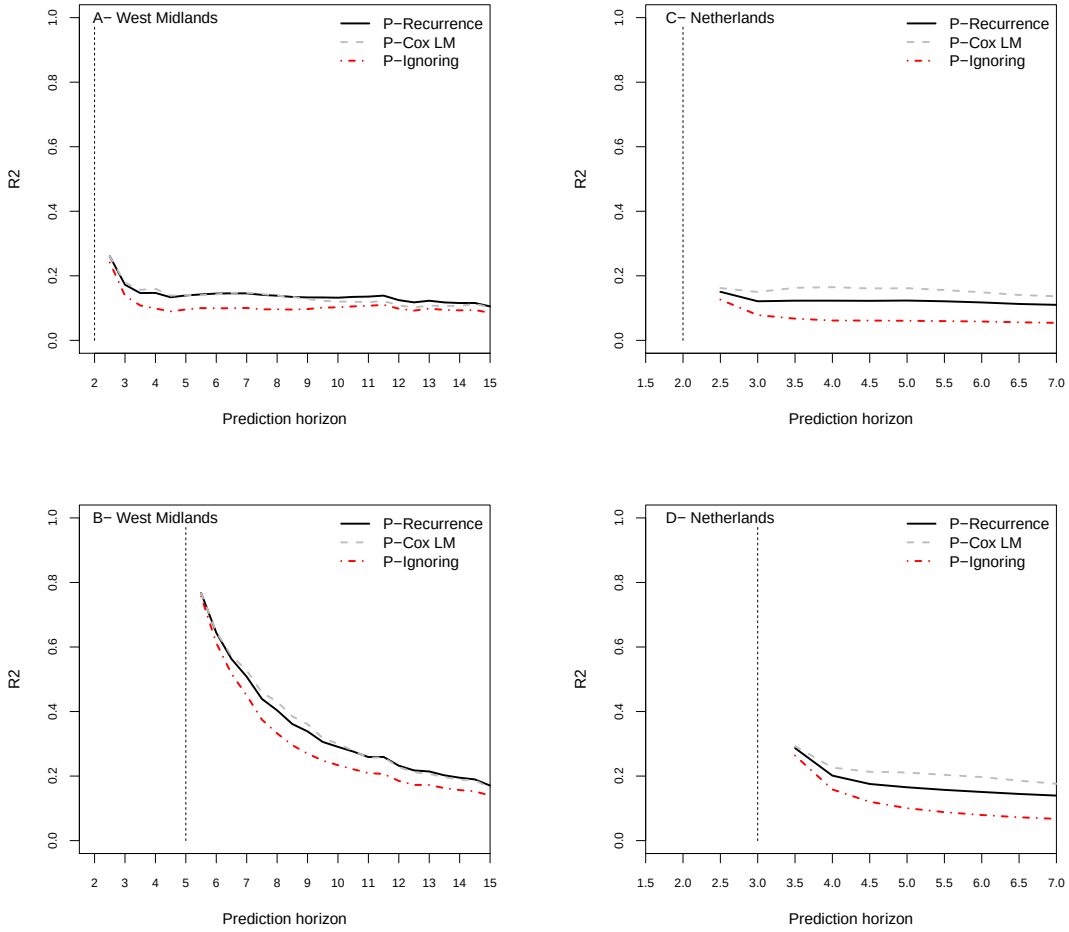

Figure 1. Error of prediction on validation populations for the three predictions. A- On the West Midlands population, at the prediction time $t = 2$ years and a prediction horizon going from 2.5 to 15 years. B- On the West Midlands population, at the prediction time $t = 5$ years and a prediction horizon going from 5.5 to 15 years. C- On the Dutch population, at the prediction time $t = 2$ years and a prediction horizon going from 2.5 to 7 years. D- On the Dutch population, at the prediction time $t = 3$ years and a prediction horizon going from 2.5 to 7 years.

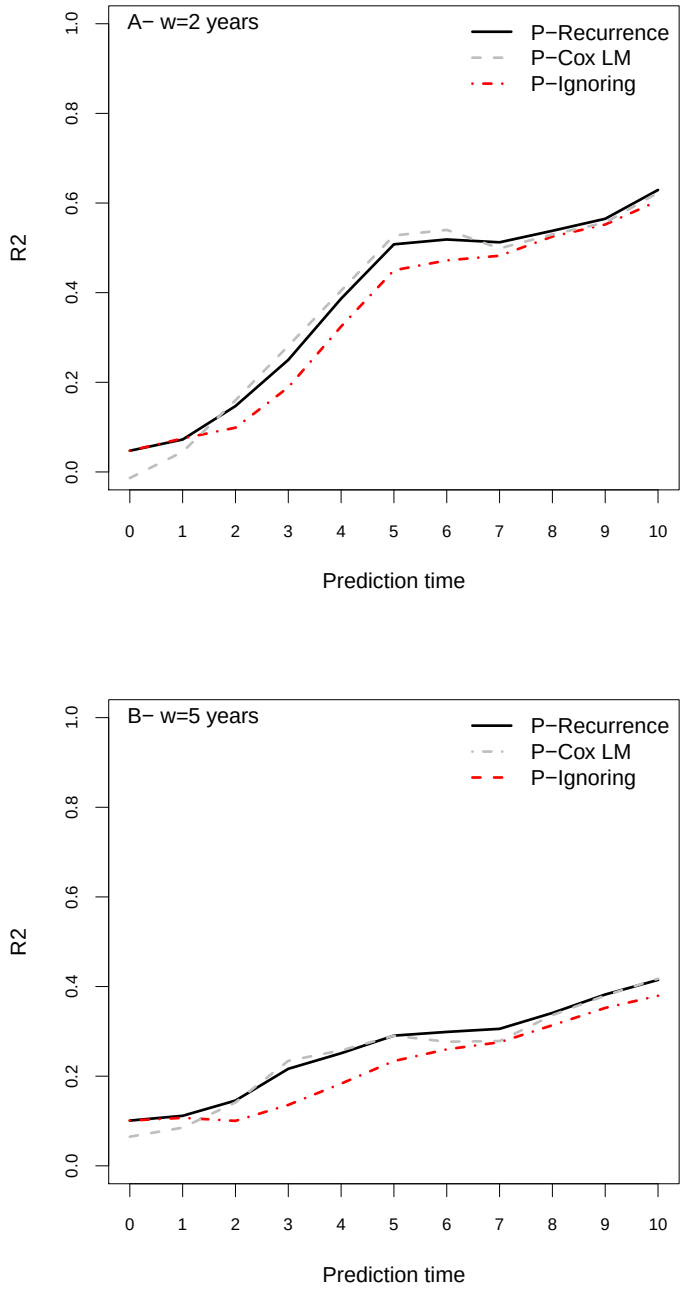

Figure 2. Error of prediction on the West Midlands population when the prediction time t is increasing from 0 to 10 and the window of prediction is held at A- 2 years and B- 5 years.

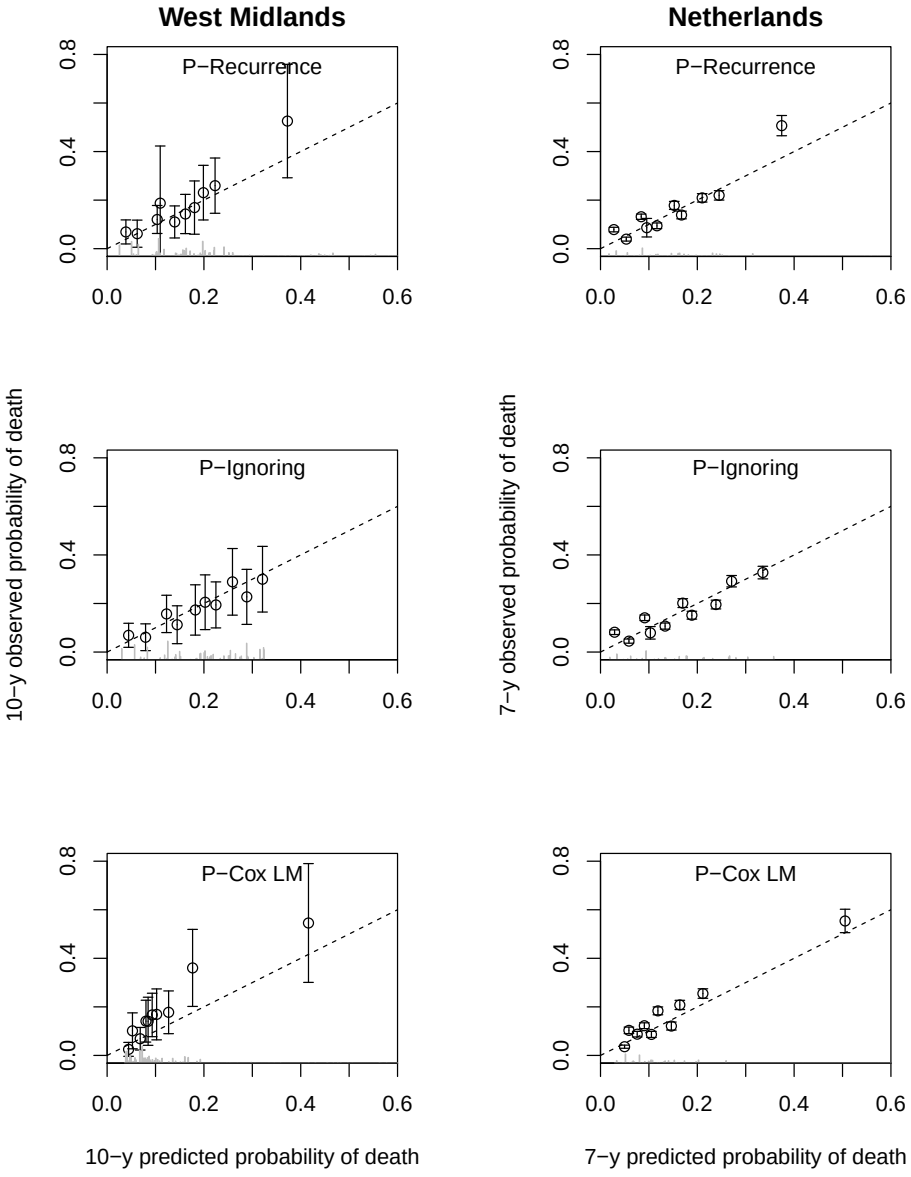

Figure 3. Calibration plot for the three predictions of dying between 5 and 10 years in the West Midlands population (left panel) and between 2 and 7 years in the Dutch population (right panel).

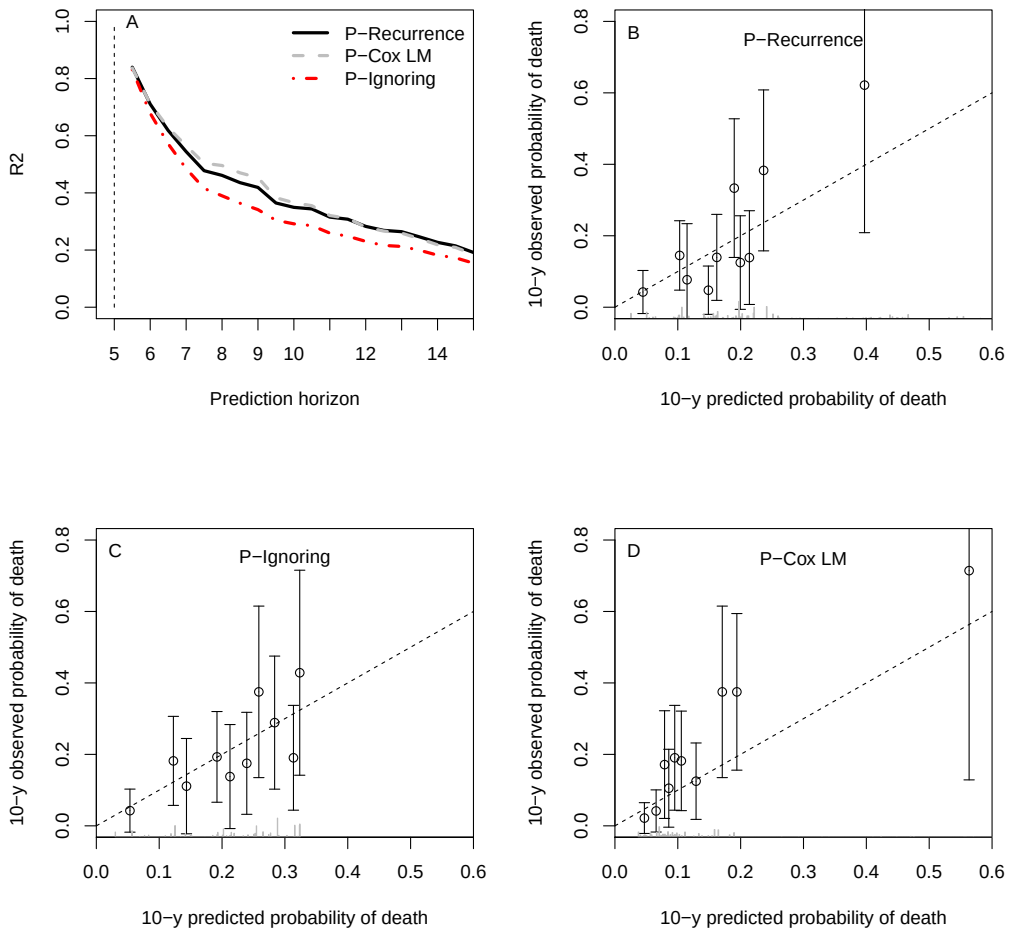

Figure 4. Result of the prediction on the operated patients from West Midlands. A- Relative prediction error at the prediction time $t = 5$ years and a prediction horizon going from 5.5 to 15 years. B,C,D - Calibration plot for the three predictions.

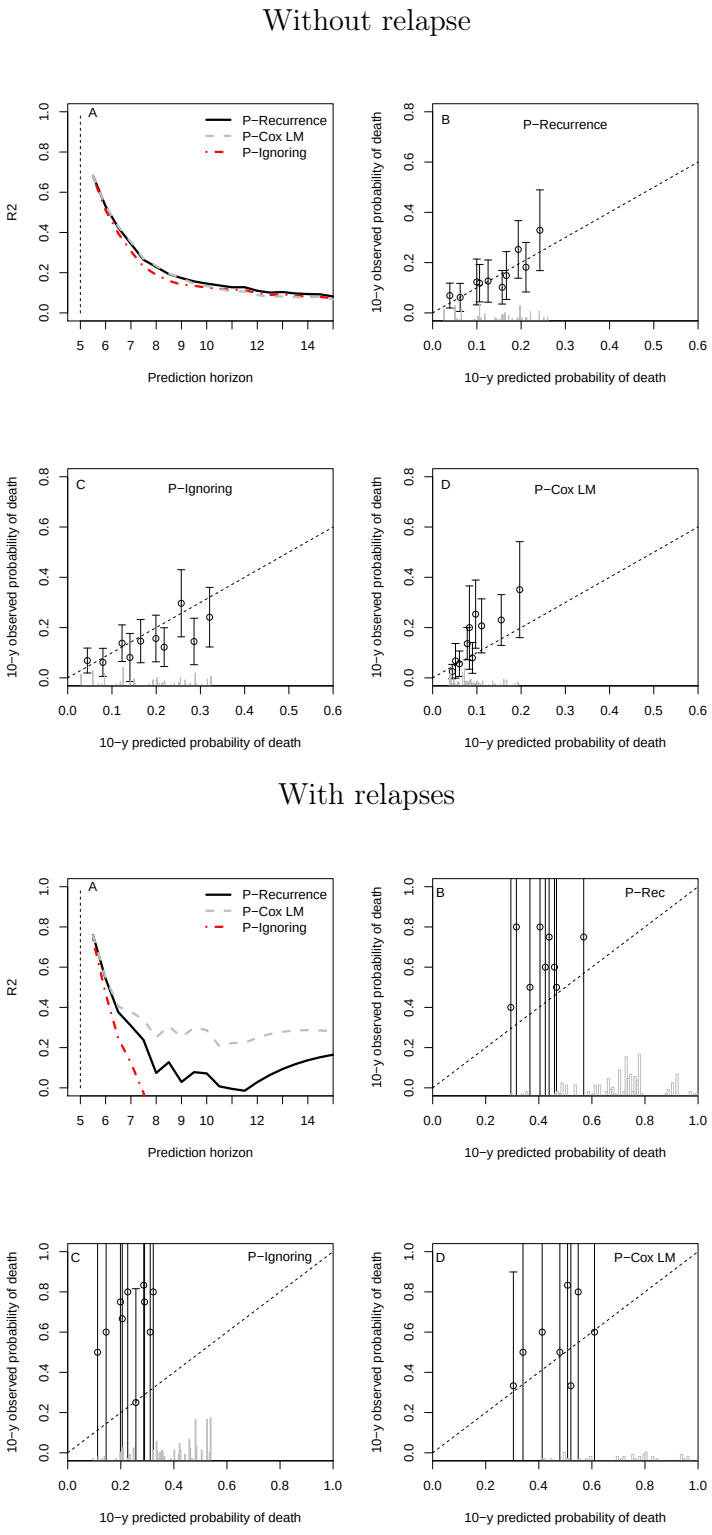

Figure 5. Result of the prediction on the patients from West Midlands. A- Relative prediction error at the prediction time $t = 5$ years and a prediction horizon going from 5.5 to 15 years. B,C,D - Calibration plot for the three predictions. The upper part is for patients without relapse before 5 years; the lower part is for patients with at least one relapse observed before 5 years.

4.3.2 Additional remarks

On the difference between internal and external validation

In this external validation, the prediction from a joint model seems to perform better than in the internal validation. This may be explained by the internal validation method that we chose. Molinaro et al. (2005) compared the performance of different resampling methods to validate predictions. They found that the prediction errors estimated with the 10-fold cross-validation approximate those obtained with leave-one-out-cross-validation in almost all settings, and stated that this method may be prefer for some computationally intense analyses. We thus used a 10-fold cross-validation for our internal validation. At each step, we used 90% of our dataset to develop the model and estimate the parameters, and the remaining 10% were used to estimate the prediction performances. However, it is possible that the 90% of our dataset was not enough to get an accurate estimation of our parameters. On the opposite, in the external validation, we used 100% of the French patients to develop the model. It gave more accurate estimations, with lower prediction error on the new populations. It seems however difficult to perform a leave-one out or a bootstrap internal validation as it would require to fit the model N times (N being the number of patients), which is too intensive when using joint models.

On the R^2

The R^2 measures were displayed in the article. The Figure 4.6 briefly explains the shape of the observed curves for the probability of death accounting for relapses in joint model. The absolute difference between the two Brier score curves is stable, but the curves are increasing with time. As a consequence, a large part of the prediction error can be explained by the model at early prediction horizons, leading to a high R^2. At late prediction times, the difference between the two curves is smaller relatively to the Brier score value, explaining the low R^2.

On the time for calibration plot

The calibration plot compares the observed versus the predicted probabilities of event, at a given time. We chose to look at the calibration at the 10-year calibration in the West Midlands population, given that the prediction is made at 5 years. As seen in the article, the calibration at 10 years was good. We also had a look at the calibration at 15 years (see Figure 4.7 on page 99). At this time, less patients were still under observations, and thus the confident interval are larger. The calibration was worst than at 10 years. This is

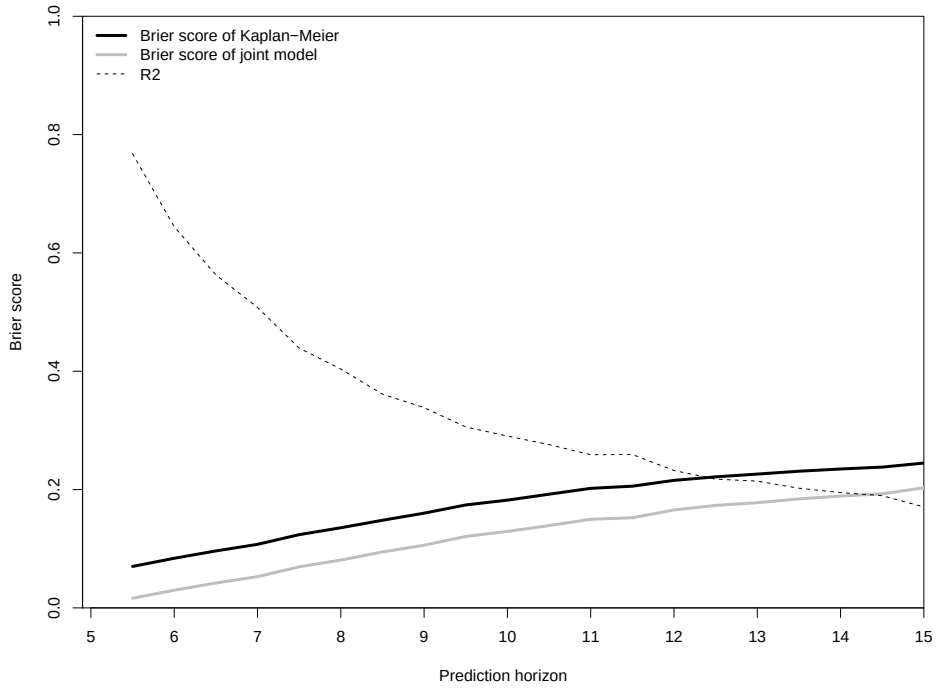

Figure 4.6: Illustration of the relation between the weighted estimator of the Brier score for the Kaplan-Meier model, for the joint model (accounting for relapses) and the R^2. The prediction time is $t = 5$ years and the window goes from 0.5 to 10 years. Illustration on West Midlands data.

coherent with the prediction error curves showing a prediction error that increases with horizon time.

On the missing factors

HER2 status and hormonal receptors status are some well-known prognostic factors in cancer patients. They were thus included in our prediction at the development step. However, they were not available at the general-population level for the validation step. They are different possibilities to handle such missing data in prediction calculation. The first one is to impute them. This requires careful imputation, and few missing data in the population, which was not our case (more than 95% of the HER2 and hormonal receptor status were missing in the English dataset). A second possibility is not to allowed for missing data, that is to calculate predictions only if all information is available. This is the solution adopted by the online calculator Prostate Cancer Calculator which

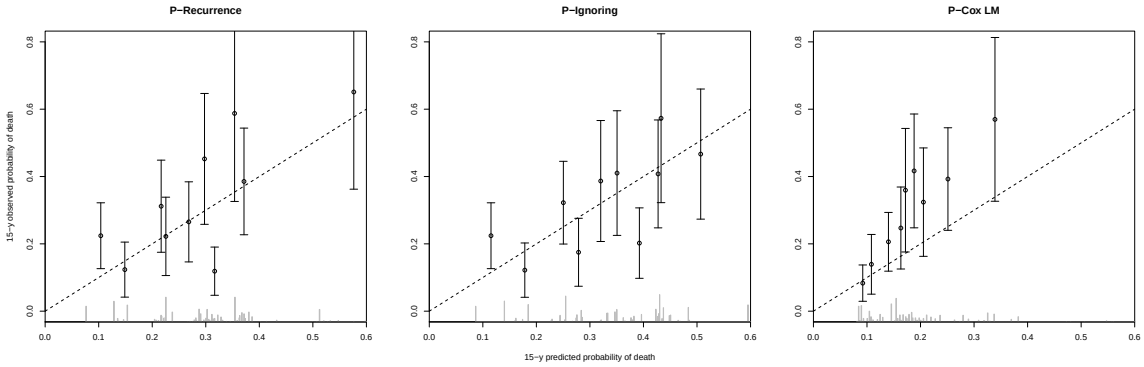

Figure 4.7: Calibration of the three proposed prediction in the West Midlands population, at 15 years (prediction time=5 years).

predicts the risk of relapses in prostate cancer patients according to their prostate-specific antigen values and other characteristics (Taylor et al., 2013). Missing values for Gleason score or T-stage are not allowed, and no prediction is calculated if they are missing. A third possibility is to give a prediction that is average on the missing value. This solution is done in the online calculator Adjuvant!, which estimate the risk of death or relapse in breast cancer patients, according to their characteristics and the treatment received (Ravdin et al., 2001). The risk estimation are calculated on population data from the SEER (Surveillance, Epidemiology, and End-Results) registry. Values allowed for oestrogen receptor status and grade include an *undefined* value. In that case, the risk calculated on all cases irrespective of the oestrogen receptor or grade status is given. Only the other factors are accounted for (such as age, tumour size or positive nodes). However, missing values for tumour size and positive nodes are not allowed.

In the validation context, we did not found literature in the impact of the missing data, when an already proposed prediction has to be validated but that frequent missing data are observed. We chose the third possibility to account for the missing data, that is proposed an average prediction whatever their value. For that, we re-estimated the model excluding the two missing factors, HER2 status and hormonal receptors status. On the development dataset, excluding these two factors had almost no impact on the prediction error (Figure 4.8). The lack of availability of some covariates in our validation datasets raises the question of the applicability of the model. Indeed, a useful prognostic model should be easily available. However, the aim of the proposed prediction is to help clinical decision. That is, be used in the patient's individual care, most of the time in hospital, where it is nowadays possible to obtain information on HER2 and hormonal

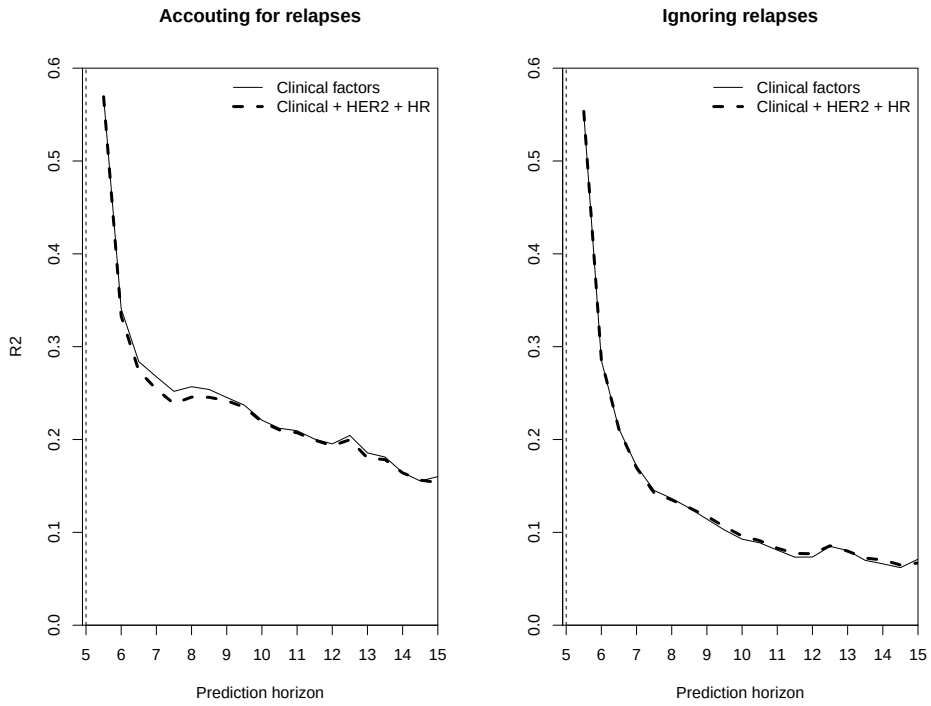

Figure 4.8: Comparison of the prediction error with and without the factors HER2 and hormonal receptor status in addition to the other clinical factors (age, peritumoural vascular invasion, tumour size, nodal involvement and grade). Prediction are from a joint model.

receptors status.

In this context, the next step is to assess one of the practical uses of the validated predictions: their use to help reducing clinical trial duration, as explained in the next chapter.

Chapter 5

Use of the individual predictions to reduce clinical trial duration

5.1 Question and data

Clinical trials on cancer research can take several years, as the main endpoint is often the overall survival. One way to reduce this time is the use of surrogate endpoint, which are information observable at earlier time point, that can be used to conclude earlier on the treatment effect. Especially, it was proposed to use the surrogate endpoint to impute the missing information on death (Faucett et al., 2002; Conlon et al., 2011; Parast et al., 2014). In this section, we use the dynamic prediction tool previously developed to impute the missing death time accounting for the observed relapses information, in the framework of joint modelling. We want to compare three different ways to impute the survival times based on the prediction. The first one is to use a mean survival time. The second one is to sample a death time from the predicted distribution of survival. In the third method, predictions are only used to define nearest neighbours. Then, the survival is non-parametrically estimated on these nearest neighbours and the survival time is sample from this distribution.

The motivating dataset was two randomised clinical trials which studied the effect of an adjuvant chemotherapy in breast cancer. Trials last for about 16 years. We kept in our analysis only the 935 patients that were included in the Institut Gustave Roussy cancer center, which represent 83% of all included patients. As seen on the Figure 5.1, the survival is not significantly different between the two treatment arms. The question of an earlier conclusion with sufficient confidence to conclude on the treatment effect (or non effect) would have been of interest in this example.

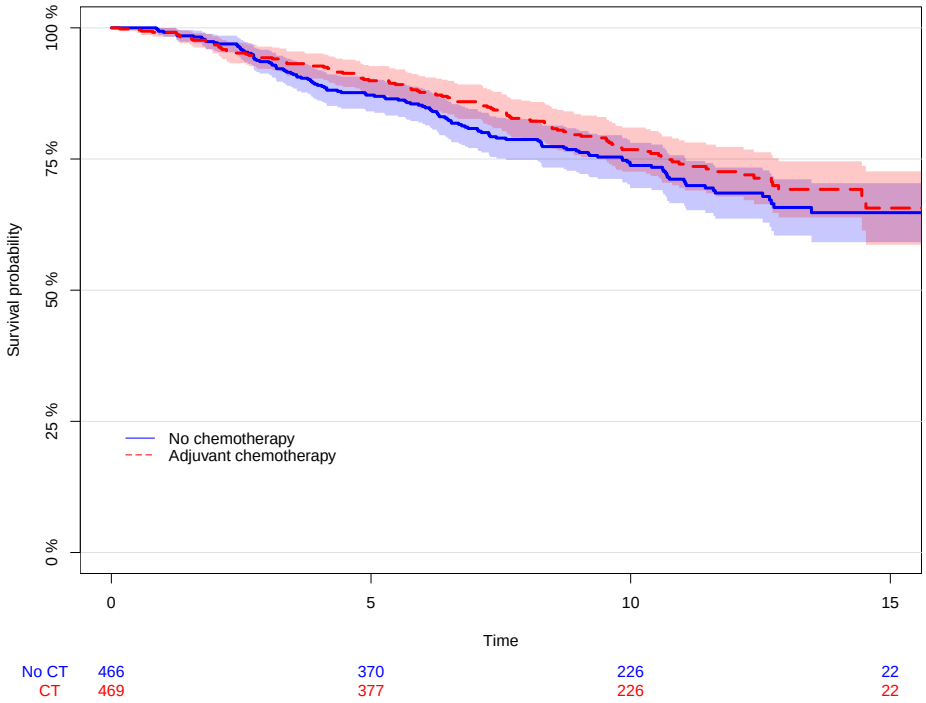

Figure 5.1: Illustration of the overall survival in the two adjuvant chemotherapy trials according to the randomised treatment.

5.2 Publication in preparation

Predict treatment effect on overall survival using cancer relapses information: imputation with joint modelling

Audrey Mauguen, Stefan Michiels, Virginie Rondeau

Abstract

Clinical trial duration may be a concern in clinical research, especially in cancer trials where the endpoint is often the death. We propose the use of a surrogate endpoint as an auxiliary variable to analyse the treatment effect earlier. At an early time point, the high number of censored deaths can be offset by the imputation of the non observed deaths times. We propose to use predictions of the risk of death from a joint model for a recurrent event and a terminal event, that account for disease relapses information. Three imputation methods are compared: the use of a restricted residual mean survival, a sampling in the parametric estimation of the survival time and a sampling in its non-parametric estimation. The treatment effect and its standard-error are estimated via multiple imputations. The performances of the three methods are compared in terms of bias in the estimates, standard-errors, type-I error, power of the analysis and coverage probability. The sampling in the parametric estimation of the survival, which appears to be the best method, is retrospectively applied on two randomised clinical trials studying the effect of an adjuvant chemotherapy in breast cancer patients.

keywords: cancer; imputation; joint model; prediction; randomized clinical trial; surrogate endpoint; survival.

1 Introduction

Clinical trial duration has been a concern for years in clinical research. In cancer treatment randomized clinical trials (RCT), the main endpoint is mainly the overall survival that takes years to be observed. It has motivated research on surrogate endpoints and auxiliary variables, that would enable to reach earlier a conclusion on treatment effect. Different approaches can be found in Prentice (1989); Pepe et al. (1994); De Gruttola et al. (1997) and Buyse and Molenberghs (1998) for example. Most of the developments have focused on the validation of these surrogate endpoints, which is the mandatory first step. Less has been done on how to effectively use them in practice. Burzykowski and Buyse (2006) have proposed the concept of Surrogate Threshold Effect, calculated during the meta-analytic surrogate validation. It represents the minimum treatment

effect that has to be observed on the surrogate endpoint to be able to conclude that the treatment will be effective on overall survival. The idea is thus to estimate the treatment effect on the surrogate endpoint, then to extrapolate the conclusion on the overall survival. This methods requires data from several similar clinical trials, which are often available when the meta-analytic approach is used for the validation. Faucett et al. (2002) and Conlon et al. (2011) proposed a different approach based on auxiliary variable. They used prediction from the joint modelling framework to impute the non observed deaths, and then estimate the treatment effect on overall survival based on the imputed data. The proposition was done using a longitudinal surrogate, the CD4 count, to estimate the effect of zidovudine on the time-to-AIDS. Recently, Parast et al. (2014) proposed a similar approach to compare survival in two treatment groups based on intermediate events. The predictions are done using landmark approach. They proposed a non-parametric estimation of the survival functions to limit the impact of model misspecification.

The progression and event-free survival have been studied as surrogate endpoint in several cancers, including breast cancer for which there is no consensus, especially in advanced setting (Miksad et al., 2008; Saad et al., 2010; Beauchemin et al., 2014). In the adjuvant setting, disease-free survival is implicitly accepted as a surrogate endpoint for overall survival (Cameron, 2007). It is thus of interest to study the disease relapses as an help to study the treatment effect on overall survival and conclude earlier. We propose an approach using joint model for a recurrent event and a terminal event. In our approach, we focus on the estimation of an hazards ratio, which is the main measure used in cancer clinical trials. We impute censored survival times according to a new approach based on the mean. We also extended the approach proposed by Faucett et al. (2002), as well as the non-parametric approach proposed by Taylor et al. (2002), to the framework of recurrent event. We compared the performance of the three methods in terms of parameter estimation and precision, coverage probability and in terms of type-I error and power for the test of treatment effect.

The motivating application of our work was data from two breast cancer randomized clinical trials studying effect of an adjuvant chemotherapy on overall survival (OS). Patients were included during a 6-year period, and the final analysis and conclusion were obtained after a follow up of 16 years. Typically, in these trials, information about first included patients could have been used to predict outcome of last included patients and conclude earlier. Our aim was to assess whether similar conclusion could have been reached earlier using relapses information. Results of the so-called early analysis, using multiple imputation, are compared to the results of the late analysis, especially in terms of standard error of the hazards ratio and rejection of the null hypothesis.

In section 2, we present the three methods to impute the death times not yet observed using relapses information using a joint model. Simulation plan and results are presented in section 3, while application of the method on breast cancer trials is presented in section 4. Finally, section

5 contains concluding remarks.

2 Methods

To predict the risk of death accounting for disease relapses information, the joint model for a recurrent event and a terminal event as proposed by Liu et al. (2004) and Rondeau et al. (2007) appears as an ideal framework. Indeed, the disease relapses can be correctly estimated, accounting for the death as a non-independent censoring. At the same time, the correlation between relapses in a same patient, as well as the association between the recurrent event and the death processes can be estimated, and then taken into account in the prediction.

2.1 Prediction using joint model

We use the following model: for subject i ($i = 1, ..., N$), let X_{ij} be the j^{th} recurrent time ($j = 1, ..., n_i$), C_i be the censored time (not by death) and D_i be the death time. $T_{ij}^R = min(X_{ij}, C_i, D_i)$ corresponds to each follow-up time and δ_{ij}^R is a binary indicator for recurrent events which is 0 if the observation is censored or if the subject died and 1 if X_{ij} is observed ($\delta_{ij}^R = I[T_{ij}^R = X_{ij}]$ where $I[.]$ denotes indicator function). Similarly, we note T_i^D the last follow-up time for subject i, which is either a time of censoring or a time of death ($T_i^D = min(C_i, D_i)$) and $\delta_i^D = I[T_i^D = D_i]$. We actually observe the sequence $(T_{ij}^R, \delta_{ij}^R, T_i^D, \delta_i^D)$. Finally, we denote by Z_{ij}^R and Z_i^D the vectors of covariates associated with the risk of recurrent events and death, respectively. Recurrent times are in the calendar timescale, that is, measured by the time elapsed since the origin of the study. However, a patient is considered at risk of a j^{th} recurrence only after the $(j-1)^{st}$ recurrence. The joint model is then written as:

$$\begin{cases} \lambda_{ij}^R(t|u_i, Z_{ij}^R) = u_i \lambda_0^R(t) \exp(\beta_1' Z_{ij}^R) = u_i \lambda_{ij}^R(t|Z_{ij}^R) \\ \lambda_i^D(t|u_i, Z_i^D) = u_i^\alpha \lambda_0^D(t) \exp(\beta_2' Z_i^D) = u_i^\alpha \lambda_i^D(t|Z_i^D) \end{cases} \tag{1}$$

with $\lambda_0^R(.)$ the baseline risk of event (irrespective of event rank), $\lambda_0^D(.)$ the baseline risk of death, β_1 and β_2 the effects of explanatory variables (assumed to be different), u_i the shared frailty effect, iid:

$$u_i \sim \text{Gamma}(\frac{1}{\theta}; \frac{1}{\theta}) \quad \text{and} \quad g(u_i) = \frac{u_i^{1/\theta - 1} \exp(-u_i/\theta)}{\theta^{1/\theta} \Gamma(1/\theta)} \tag{2}$$

The baseline hazard functions $\lambda_0^R(.)$ and $\lambda_0^D(.)$ are approximated using splines and $\xi = (\lambda_0^R(.), \lambda_0^D(.), \beta_1, \beta_2, \alpha, \theta)$ the parameter vector of the model is estimated using penalized max-

imum likelihood. The corresponding variance-covariance matrix Σ is estimated by using the inversion of the negative Hessian matrix of the penalized likelihood. All estimation details can be found in Rondeau et al. (2007).

We define the history of relapses as the number of relapses occurring before the prediction time s and their time of occurrence: $\mathcal{H}_i^J(s) = \{N_i^R(s) = J, X_{i1} < \ldots < X_{iJ} \leq s\}$, with $X_{i0} = 0$ and $X_{i(J+1)} > s$. We then define a dynamic individual prediction, $P^D(s, s+w; \xi)$ being the probability of death between the prediction time s and the horizon $s+w$, given that the patient is alive at time s, given his relapse history and his covariates. The prediction is defined as follows (Mauguen et al., 2013):

$$
\begin{aligned}
&P^D(s, s+w; \xi) \qquad (3)\\
&= P(D_i \leq s+w | D_i > s, \mathcal{H}_i^J(s), Z_{s,ij}^R, Z_{s,i}^D, \xi)\\
&= \frac{\int_0^\infty [S_i^D(s|Z_{s,i}^D, u_i, \xi) - S_i^D(s+w|Z_{s,i}^D, u_i, \xi)](u_i)^J S_{i(J+1)}^R(s|Z_{s,ij}^R, u_i, \xi) g(u_i) \mathrm{d}u_i}{\int_0^\infty S_i^D(s|Z_{s,i}^D, u_i, \xi)(u_i)^J S_{i(J+1)}^R(s|Z_{s,ij}^R, u_i, \xi) g(u_i) \mathrm{d}u_i}
\end{aligned}
$$

We thus have the corresponding conditional survival:

$$
\begin{aligned}
S^D(s+w|s, \xi) &= S^D(t = s+w | t > s, \mathcal{H}_i^J(s), Z_{s,ij}^R, Z_{s,i}^D, \xi)\\
&= P(D_i > s+w | D_i > s, \mathcal{H}_i^J(s), Z_{s,ij}^R, Z_{s,i}^D, \xi)\\
&= 1 - P^D(s, s+w; \xi)
\end{aligned}
$$

2.2 Imputation of the censored data

The predictions from the joint model were used to impute missing deaths in three different ways. The model is estimated on all included patients, but mainly the first included patients, who have the longer follow-up, and patients not censored bring the information needed to do prediction for other censored patients. This is illustrated in Figure 1. For each of these imputation methods, M different datasets were created for the purpose of multiple imputation.

2.2.1 Imputation based on the mean survival time

We propose to use the classical way to turn a survival risk (obtained from the prediction) into a survival time, that is the mean survival time. The mean survival time is defined by $E[D_i] = \int_0^\infty S(u)\mathrm{d}u$. However, it is dependent on the behaviour of the right-hand tail of the distribution, and thus can be impacted by the presence of censoring (Aalen et al., 2008). In our context of intermediate analysis in a clinical trial, this may be a serious concern, because the patients' follow-

up time is limited. It is thus advise to use the *restricted* mean survival time $E[D_i \wedge \tau] = \int_0^\tau S(u)\mathrm{d}u$, with τ equal the maximum follow-up time (Royston and Parmar, 2011). This quantity is also called the τ-year life expectancy. Another restriction in our context is that the predictions are made given that the patient is still alive at his censoring time C_i. To take this condition into account, a *residual* survival time is used: $E[D_i|D_i > C_i] = \frac{\int_{C_i}^{\infty} S(u)\mathrm{d}u}{S(C_i)}$. Combining these concepts, the missing survival observations were imputed by the restricted residual mean survival time. It corresponds to a τ-year residual life expectancy. The imputation steps were as follows.

For each of the $m = 1, \ldots, M$ dataset to be imputed:

1. sample $\xi_m \sim \mathcal{MN}(\hat{\xi}, \hat{\Sigma})$
2. estimate $\hat{P}^D_{i,m}(C_i, t; \xi_m), \forall t \in [C_i, \tau]$ and thus $\hat{S}^D_{i,m}(t|C_i, \xi_m)$
3. for each censored patient i, compute the new observation time $T^{D*}_{i,m}$ using the restricted residual mean survival time conditional on his history and covariates

$$\begin{aligned} T^{D*}_{i,m} &= C_i + E\left[(D_i - C_i) \wedge (\tau - C_i)|D_i > C_i, \xi_m\right] \\ &= \frac{\int_{C_i}^{\tau} \hat{S}^D_{i,m}(t|C_i, \xi_m)\mathrm{d}t}{\hat{S}^D_{i,m}(C_i|\xi_m)} \end{aligned}$$

   where $\hat{S}^D_{i,m}(C_i|\xi_m) = S^D(t = C_i|\mathcal{H}^J_i(s), Z^R_{s,ij}, Z^D_{s,i}, \xi_m)$.
4. replace C_i by $T^{D*}_{i,m}$, and $\delta^D_i = 0$ by 1 if $T^{D*}_{i,m} < \tau$, 0 otherwise ($T^{D*}_{i,m} = \tau$).

2.2.2 Imputation based on the predicted probability of event

The second method used was initially proposed for longitudinal data by Faucett et al. (2002) and for recurrent event in the context of a cure model by Conlon et al. (2011). We extend it to joint model for recurrent event and a terminal event. It consists in sampling in the distribution of the probability of event and imputing the corresponding time of event. The imputation steps were as follows.

For each of the $m = 1, \ldots, M$ dataset to be imputed:

1. sample $\xi_m \sim \mathcal{MN}(\hat{\xi}, \hat{\Sigma})$
2. for each censored patient i
   (a) estimate $\hat{P}^D_{i,m}(C_i, t; \xi_m), \forall t \in [C_i, \tau]$
   (b) draw a random variable $a_{i,m} \sim U[0; 1]$

(c) find the value of $T_{i,m}^{D*}$ which verifies $\hat{P}_{i,m}^{D}(C_i, T_{i,m}^{D*}; \xi_m) = a_{i,m}$ for $T_{i,m}^{D*} \in [C_i, \tau]$
Note that no closed-form solution exists for the integration on u_i in equation (3), unless $\alpha = 1$. Thus, we used an iterative optimisation method, through the R function *optimize*, to minimize $|\hat{P}_{i,m}^{D}(C_i, T_{i,m}^{D*}; \xi_m) - a_{i,m}|$ with $T_{i,m}^{D*} \in [C_i, \tau]$.

(d) replace C_i by $T_{i,m}^{D*}$, and $\delta_i^D = 0$ by 1 if $T_{i,m}^{D*} < \tau$, 0 otherwise ($T_{i,m}^{D*} = \tau$).

2.2.3 Non parametric imputation

The last method used was initially proposed in Taylor et al. (2002) and Hsu and Taylor (2009) to estimate and compare survival distributions. It uses the Kaplan-Meier estimation of the survival rate of nearest neighbours. Similarly to the previous section, the method consists in sampling in the non-parametric estimation of the distribution of the probability of event and imputing the corresponding time of event. The imputation steps were as follows.

For each of the $m = 1, \ldots, M$ dataset to be imputed, for each censored patient i:

1. define the risk set $R_{i,m}^{+}$ of patients still at risk at time C_i
2. estimate $\hat{P}_i^D(C_i, \tau; \hat{\xi})$ the predicted probability of event between C_i and τ
3. select the NN nearest neighbours in $R_{i,m}^{+}$, defined by the NN patients with the smallest distance from patient i; the distance between patients i and j is defined by $\sqrt{\left(\hat{P}_i^D(C_i, \tau; \hat{\xi}) - \hat{P}_j^D(C_i, \tau; \hat{\xi})\right)^2}$
4. estimate the cumulative distribution function $1 - \hat{S}_{i,m}^{NN}(.)$, where $\hat{S}_{i,m}^{NN}(.)$ is the Kaplan-Meier estimator of the overall survival in the NN selected patients
5. draw a random variable $a_{i,m} \sim U[0; 1]$
6. find empirically the solution $T_{i,m}^{D*}$ of $1 - \hat{S}_{i,m}^{NN}(T_{i,m}^{D*}) = a_{i,m}$
7. replace C_i by $T_{i,m}^{D*}$, and $\delta_i^D = 0$ by 1 if $T_{i,m}^{D*} < \tau$, 0 otherwise ($T_{i,m}^{D*} = \tau$).

Note that imputed times are not included in the risk set $R_{i,m}^{+}$ for subsequent patients. Also, if the last observation time of the NN selected patients t_{NN} is censored, the imputed time $T_{i,m}^{D*}$ will be a censored time with probability $1 - \hat{S}_{i,m}^{NN}(t_{NN})$. Finally, and this is true for the three imputation method, a survival time can only be imputed once, even if the imputation value is a censoring time.

Initially, the distance used to define the NN nearest neighbours accounted for the probability of censoring in addition to the probability of event. However, the simulation results suggested that the censoring probability did not add much to the performance of the method (Hsu and Taylor,

2009). Therefore, we focused on the probability of event. The number NN was advised between 5 and 20.

2.3 Multiple imputation

As discussed in Rubin (1996), a single imputation is not enough to get some accurate results. For each of the imputation methods presented, multiple imputation was performed. For each of the M datasets created with different imputed values, the treatment effect $\hat{\beta}_m$ was estimated. Its variance $\hat{\sigma}_m^2$ was estimated by the inverse of the Hessian matrix. Results of the M datasets were then combined following Rubin's rules (Rubin, 1996). The resulted treatment effect $\hat{\beta}$ is an average of the M coefficients $\hat{\beta}_m$ estimated on each dataset.

$$\hat{\beta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}_m$$

The corresponding variance $\hat{\sigma}^2$ is a sum of two terms: the average within-imputation variance among the M datasets, and a term accounting for the between-imputation variance.

$$\hat{\sigma}^2 = \frac{1}{M} \sum_{m=1}^{M} \hat{\sigma}_m^2 + (1 + 1/M) \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\beta}_m - \hat{\beta} \right)^2$$

2.4 Analysis of the imputed datasets

Once the datasets are imputed, the treatment effect β_m is estimated using a Cox model. A joint model was also used as a sensitivity analysis. Results of the imputation schemes at an early time were compared with 1) the results of the late analysis, on all the observed events and 2) the results of the early analysis without any imputation, on only the events observed at early time. Results compared are the treatment effect estimate $\hat{\beta}$, its estimated standard error $\hat{\sigma}$ and the empirical standard error. The type-I error and power estimated across the 500 simulated trials were also compared, as well as the coverage probability.

3 Simulation study

3.1 Simulation setting

We simulated S=500 trials, including N=1000 patients each. The inclusion time period was defined as the five first years ($t_1 = 5$, see Figure 1). The results of the early analysis at six years were compared to a late analysis at ten years ($t_2 = 6$ and $t_3 = 10$, see Figure 1). The baseline survival followed an exponential distribution for the relapses, and a Weibull distribution

for the death, with a risk increasing with time (shape>1). Relapses and deaths were correlated using a frailty effect following a Gamma distribution with mean one and a variance $\theta = 0.8$. The parameter α in model (1) was equal to 1. Two covariates were considered: $X_{1,i} \sim B(0.5)$ mimicked the treatment, and $X_{2,i} \sim \mathcal{N}(0,1)$ is a continuous covariate. Both were impacting the risk of relapses with effect $HR_{1,relapse} = 0.9$ and $HR_{2,relapse} = 1.1$, respectively. Only $X_{1,i}$ had an impact on the risk of death. The null hypothesis was defined as $H_0 : \beta = 0$ ($HR_{1,death} = 1$) and the alternative hypothesis was defined as $H_1 : \beta = -0.22$ ($HR_{1,death} = 0.8$). Four other scenarios were studied to assess the robustness of the result: a stronger effect of the treatment on the relapses ($HR_{1,relapse} = 0.7$), a stronger effect of the frailty on the death risk ($\alpha = 1.5$), both $HR_{1,relapse} = 0.7$ and $\alpha = 1.5$, and finally a higher heterogeneity ($\theta = 1.2$). The spline basis used to estimate the baseline hazards in joint model for prediction were kept equal in all simulations. The number of imputations was M=10.

We simulated correlated survival times, but we used a Cox model for the analysis, that does not account for the correlations. We thus did not expect the model to retrieve the simulated coefficient β. To calculate the bias of the estimation, we rather used as a reference an asymptotic value β_{as}. This value was obtained by generating 10,000 trials of 10,000 patients using the method described above. The result of the late Cox analysis, that uses all the available information, was calculated for each trial. β_{as} was defined as the mean of this late Cox analysis result among the 10,000 trials.

The bias is defined by the mean difference between the estimated coefficient and β_{as}. Two standard-errors are presented: $\hat{\sigma}$ is the mean of the estimated standard-errors among the 500 simulations; $\hat{\sigma}_{emp}$ is the empirical standard-error of the $\hat{\beta}$ among the 500 simulations.

3.2 Simulation results

The simulation results are presented in Table 1. First, the number of events analysed with the imputation methods is closed to the number of events actually observed in the late analysis, even if a little overestimated. The β coefficients were overall well estimated, both at the late and the early analyses.

The imputations using the parametric or the non-parametric sampling did not introduce any bias in the estimation. However, the bias was higher with the method based on the mean, for all studied scenarios.

The main interest of the imputation was to get smaller standard-errors than using partially observed data at the early time analysis. The estimated standard-errors using the non-parametric sampling were not higher than the early analysis ones, but we can see that the empirical ones were actually higher. This method is thus not efficient.

The only method that seems to give satisfactory results in terms of bias and standard-errors

is the parametric sampling. The bias was low with this method, always lower or equal than the one of the late analysis, except in scenarios 3 and 5 where the bias equals -0.01 for the parametric sampling and was null at the late analysis. The empirical standard-errors were reduced as compared to the early analysis for all scenarios. The empirical standard-errors were slightly lower than the estimated ones. We can thus conclude that this parametric method of imputation is efficient.

This efficiency does not translate in better performance in terms of power (Table 2). Indeed, the power to reject the null hypothesis using imputation based on the parametric sampling is never as high as the power of the late analysis, and even lower that the early analysis without imputation. The type-I error was preserved and lower with the imputation than both the early and the late analyses. A low type-I error prevents from concluding wrongly that the treatment is efficient. The coverage probability was similar or higher with the imputation using the parametric sampling than with the early analysis, showing that the estimation was slightly more accurate.

4 Application

The motivating data was two randomized clinical trials studying the effect of an adjuvant chemotherapy in pre- and post-menopausal breast cancer patients (Arriagada et al., 2005). Only patients included at Gustave Roussy Institute (Villejuif) were kept in the analysis, that is 935 of the 1146 randomized patients (83%) (Conforti et al., 2007). Patients were randomized between surgery alone or surgery plus adjuvant chemotherapy. More details can be found in Arriagada et al. (2005) and Conforti et al. (2007). The patients were included during seven years, from 1989 to 1996, and followed up to 16 years (year 2005). Thus, results at the 16 years after the trial beginning were considered as the results of the late analysis, and compared to results of two early analyses at 8 and 10 years after the trials beginning.

During the follow-up, 236 deaths and a total of 446 relapses were observed. Relapses were observed in 348 patients, with one to four relapses per patient. At the time of the 8-year early analysis, 165 deaths and 357 relapses in 281 patients (one to four per patient) were observed. At the time of the 10-year early analysis, 201 deaths and 396 relapses in 309 patients were observed. Estimation of the treatment effect after 8, 10 and 16 years of follow-up is shown in Table 3. We can see that the same conclusions on the treatment effect could have been drawn 8 or 6 years before the end of the study at the early analysis time using the proposed imputation approach.

5 Discussion

In this article, we showed that it was possible to retrieve information about non observed death using prediction of death in a joint model framework. Indeed, we imputed the missing death times in three different ways, and were able to retrieve the numbers of deaths observed.

The method based on the restricted residual mean gave some biased results. It overestimates the probability of death. Under the null hypothesis, this bias is lower because the overestimation is the same in both arms. However, under the alternative hypothesis, less deaths are observed in the treatment arm. Thus, a larger part of the patients has a probability of death overestimated by the imputation in the treatment arm than in the control arm. This leads to a bias in favour of the control arm.

The method based on the non-parametric sampling gave some unbiased estimation of the treatment effect. However, the standard-error was underestimated. This resulted in a very high type-I error and low coverage probability. One explanation could be the number of nearest neighbours used. We used 10 nearest neighbours, as it seems a reasonable number in the work of Taylor et al. (2002). We compared these results with our imputation procedure using 20 and 50 nearest neighbours, and the results were very similar, included the empirical standard estimations.

The best imputation method was the one based on the parametric sampling, consisting in sampling in the survival time distribution estimated on both observed deaths and deaths imputed using prediction from a joint model. This method gave satisfactory results, both in term of bias and standard-errors. As expected, the standard-errors of the estimations were reduced with the imputation. The gain was small, but consistent with the results of Conlon et al. (2011) and Parast et al. (2014). This imputation method did not improve the power as compared to an early analysis on available data only. However, a small bias was observed for the early analysis, always in favour of the treatment arm. This probably explains that the power of the early analysis was higher than the one obtained using imputation, while the coverage probability was not. As a comparison, Parast et al. (2014) had an actual increase in the power when considering the auxiliary information, but studied only the difference between the two arms at one time point, not the all survival curve. Conlon et al. (2011) did not study the power of the analysis in simulation.

The imputation using a joint model may suffer to be model-based. First, it can be subject to model misspecification. Secondly, the risks are assumed to be proportional and this may not the case when having such a long follow-up.

Finally, another approach to study censored survival times are the jackknife pseudo observations (Andersen and Perme, 2010). Each observation is replaced by its contribution to the estimator of interest (for example, the survival mean). The idea is to have uncensored data to be able to use standard regression methods, such as linear or logistic regressions. They have proposed

to study the restricted mean survival time (Andersen et al., 2004; Royston and Parmar, 2011). Although we chose the imputation procedure instead, as we wanted to use a Cox model as a final analysis to have the hazard ratio estimation, it would be of interest to investigate their use in the context of reducing the duration of clinical trials.

References

Aalen, O., O. Borgan, and H. Gjessing (2008). *Survival and Event History Analysis: A Process Point of View*. Springer, New York.

Andersen, P. K., M. G. Hansen, and J. P. Klein (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime data analysis 10*(4), 335–350.

Andersen, P. K. and M. P. Perme (2010). Pseudo-observations in survival analysis. *Statistical Methods in Medical Research 19*(1), 71–99.

Arriagada, R., M. Spielmann, S. Koscielny, T. Le Chevalier, T. Delozier, M. Reme-Saumon, M. Ducourtieux, T. Tursz, and C. Hill (2005). Results of two randomized trials evaluating adjuvant anthracycline-based chemotherapy in 1 146 patients with early breast cancer. *Acta Oncologica 44*(5), 458–466.

Beauchemin, C., D. Cooper, L. Yelle, J. Lachaine, and M.-E. Lapierre (2014). Progression-free survival as a potential surrogate for overall survival in metastatic breast cancer. *OncoTargets and Therapy 7*, 1101–1110.

Burzykowski, T. and M. Buyse (2006). Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics 5*(3), 173–186.

Buyse, M. and G. Molenberghs (1998). Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics 54*(3), 1014–1029.

Cameron, D. (2007). Clinical outcomes: to be a surrogate or not to be ...? *Breast Cancer Research 9*(Suppl 2), S26.

Conforti, R., T. Boulet, G. Tomasic, E. Taranchon, R. Arriagada, M. Spielmann, M. Ducourtieux, J. Soria, T. Tursz, S. Delaloge, S. Michiels, and F. Andre (2007). Breast cancer molecular subclassification and estrogen receptor expression to predict efficacy of adjuvant anthracyclines-based chemotherapy: a biomarker study from two randomized trials. *Annals of Oncology 18*(9), 1477–1483.

Conlon, A. S. C., J. M. G. Taylor, D. J. Sargent, and G. Yothers (2011). Using cure models and multiple imputation to utilize recurrence as an auxiliary variable for overall survival. *Clinical trials 8*(5), 581–590.

De Gruttola, V., T. Fleming, D. Y. Lin, and R. Coombs (1997). Perspective: Validating surrogate markers—are we being naive? *Journal of Infectious Diseases 175*(2), 237–246.

Faucett, C. L., N. Schenker, and J. M. Taylor (2002). Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. *Biometrics 58*(1), 37–47.

Hsu, C.-H. and J. M. G. Taylor (2009). Nonparametric comparison of two survival functions with dependent censoring via nonparametric multiple imputation. *Statistics in Medicine 28*(3), 462–475.

Liu, L., R. Wolfe, and X. Huang (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics 60*(3), 747–756.

Mauguen, A., B. Rachet, S. Mathoulin-Pélissier, G. MacGrogan, A. Laurent, and V. Rondeau (2013). Dynamic prediction of risk of death using history of cancer recurrences in joint frailty models. *Statistics in Medicine 32*(30), 5366–5380.

Miksad, R. A., V. Zietemann, R. Gothe, R. Schwarzer, A. Conrads-Frank, P. Schnell-Inderst, B. Stollenwerk, and U. Siebert (2008). Progression-free survival as a surrogate endpoint in advanced breast cancer. *International Journal of Technology Assessment in Health Care 24*(04), 371–383.

Parast, L., L. Tian, and T. Cai (2014). Landmark estimation of survival and treatment effect in a randomized clinical trial. *Journal of the American Statistical Association 109*(505), 384–394.

Pepe, M. S., M. Reilly, and T. R. Fleming (1994). Auxiliary outcome data and the mean score method. *Journal of Statistical Planning and Inference 42*(1–2), 137–160.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine 8*(4), 431–440.

Rondeau, V., S. Mathoulin-Pélissier, H. Jacqmin-Gadda, V. Brouste, and P. Soubeyran (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics 8*(4), 708–721.

Royston, P. and M. K. B. Parmar (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine 30*(19), 2409–2421.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association 91*(434), 473–489.

Saad, E. D., A. Katz, P. M. Hoff, and M. Buyse (2010). Progression-free survival as surrogate and as true end point: insights from the breast and colorectal cancer literature. *Annals of Oncology 21*(1), 7–12.

Taylor, J. M. G., S. Murray, and C.-H. Hsu (2002). Survival estimation and testing via multiple imputation. *Statistics & Probability Letters 58*(3), 221–232.

Table 1: Simulation results according to the 5 scenarios: parameter estimation.

Scenario	N_{deaths} fully observed	N_{deaths} observed + imputed	$\hat{\beta}$	Bias*	$\hat{\sigma}$	$\hat{\sigma}_{emp}$
1. H_0**:** $HR_{1,death} = 1$**,** $HR_{1,relapse} = 0.9$**,** $\alpha = 1$**,** $\theta = 0.8$**;** $\beta_{as} = 0.017$						
Late analysis	596.4	596.4	0.02	0.01	0.082	0.080
Early analysis	342.0	342.0	0.03	0.01	0.108	0.105
Imputation: Mean survival	342.0	667.8	0.05	0.03	0.080	0.071
Imputation: Parametric sampling	342.0	663.4	0.02	0.00	0.101	0.096
Imputation: Non-parametric sampling	342.0	545.0	0.04	0.03	0.100	0.178
2. H_1**:** $HR_{1,death} = 0.8$**,** $HR_{1,relapse} = 0.9$**,** $\alpha = 1$**,** $\theta = 0.8$**;** $\beta_{as} = -0.159$						
Late analysis	564.5	564.5	-0.15	0.01	0.084	0.080
Early analysis	317.4	317.4	-0.17	-0.01	0.113	0.110
Imputation: Mean survival	317.4	653.7	-0.05	0.11	0.081	0.074
Imputation: Parametric sampling	317.4	631.5	-0.15	0.01	0.105	0.102
Imputation: Non-parametric sampling	317.4	512.3	-0.15	0.01	0.103	0.196
3. $HR_{1,death} = 0.8$**,** $HR_{1,relapse} = 0.7$**,** $\alpha = 1$**,** $\theta = 0.8$**;** $\beta_{as} = -0.128$						
Late analysis	578.9	578.9	-0.13	0.00	0.083	0.082
Early analysis	322.2	322.2	-0.15	-0.02	0.112	0.108
Imputation: Mean survival	322.2	656.2	-0.04	0.09	0.081	0.076
Imputation: Parametric sampling	322.2	634.9	-0.14	-0.01	0.105	0.098
Imputation: Non-parametric sampling	322.2	518.3	-0.11	0.02	0.102	0.183
4. $HR_{1,death} = 0.8$**,** $HR_{1,relapse} = 0.9$**,** $\alpha = 1.5$**,** $\theta = 0.8$**;** $\beta_{as} = -0.130$						
Late analysis	533.0	533.0	-0.12	0.01	0.087	0.087
Early analysis	320.6	320.6	-0.14	-0.01	0.112	0.113
Imputation: Mean survival	320.6	620.5	-0.04	0.09	0.084	0.071
Imputation: Parametric sampling	320.6	606.9	-0.12	0.01	0.103	0.095
Imputation: Non-parametric sampling	320.6	484.3	-0.12	0.01	0.104	0.186
5. $HR_{1,death} = 0.8$**,** $HR_{1,relapse} = 0.7$**,** $\alpha = 1.5$**,** $\theta = 0.8$**;** $\beta_{as} = -0.099$						
Late analysis	544.8	544.8	-0.09	0.00	0.086	0.083
Early analysis	324.2	324.2	-0.12	-0.02	0.112	0.109
Imputation: Mean survival	324.2	624.8	-0.02	0.07	0.083	0.071
Imputation: Parametric sampling	324.2	611.0	-0.11	-0.01	0.102	0.091
Imputation: Non-parametric sampling	324.2	489.8	-0.10	0.00	0.103	0.176
6. $HR_{1,death} = 0.8$**,** $HR_{1,relapse} = 0.9$**,** $\alpha = 1$**,** $\theta = 1.2$**;** $\beta_{as} = -0.146$						
Late analysis	512.2	512.2	-0.15	0.00	0.089	0.086
Early analysis	292.4	292.4	-0.17	-0.02	0.118	0.120
Imputation: Mean survival	292.4	630.5	-0.04	0.10	0.083	0.073
Imputation: Parametric sampling	292.4	599.1	-0.15	-0.00	0.107	0.104
Imputation: Non-parametric sampling	292.4	462.6	-0.14	0.00	0.108	0.198

β_{as}: true value of the β to be estimated.

* Bias is the difference between the estimator $\hat{\beta}$ and the asymptotic value β_{as}.

Table 2: Simulation results according to the 5 scenarios: type-I error, power and coverage probability.

Method	Type-I error α H_0^*	Power $1-\beta$ H_1^*	Power $1-\beta$ H_1 $HR_{1,relapse}=0.7$	Power $1-\beta$ H_1 $\alpha=1.5$	Power $1-\beta$ H_1 $HR_{1,relapse}=0.7$, $\alpha=1.5$	Power $1-\beta$ H_1 $\theta=1.2$
Late analysis	0.058	0.452	0.352	0.318	0.204	0.416
Early analysis	0.040	0.316	0.278	0.262	0.198	0.310
Expected survival	0.054	0.084	0.082	0.038	0.044	0.068
Parametric sampling	0.032	0.304	0.256	0.216	0.186	0.288
Non-parametric sampling	0.298	0.456	0.394	0.378	0.322	0.416
Coverage probabilities						
Late analysis	0.954	0.954	0.956	0.954	0.970	0.960
Early analysis	0.964	0.954	0.954	0.964	0.940	0.940
Expected survival	0.966	0.758	0.828	0.838	0.898	0.806
Parametric sampling	0.970	0.946	0.968	0.962	0.968	0.960
Non-parametric sampling	0.732	0.682	0.730	0.738	0.766	0.716

*H_0: $HR_{1,death}=1$, $HR_{1,relapse}=0.9$, $\alpha=1$, $\theta=0.8$
H_1: $HR_{1,death}=0.8$, $HR_{1,relapse}=0.9$, $\alpha=1$, $\theta=0.8$

Table 3: Application results: estimation of the treatment effect

Method	N_{deaths} observed	N_{deaths} observed + imputed	$\widehat{HR}$	$\hat{\sigma}$	95% CI
Treatment effect only (N=935 patients)					
Based on 8 year - no imputation	165	165.0	0.80	0.157	[0.59 - 1.08]
Based on 8 year - parametric imputation	165	184.9	0.86	0.159	[0.63 - 1.17]
Based on 10 year - no imputation	201	201.0	0.85	0.141	[0.64 - 1.12]
Based on 10 year - parametric imputation	201	241.4	0.87	0.137	[0.66 - 1.13]
Based on 16 year - no imputation	236	236.0	0.86	0.130	[0.67 - 1.11]
Adjusted on prognostic factors* (N=769 patients)					
Based on 8 year - no imputation	140	140.0	0.87	0.171	[0.62 - 1.22]
Based on 8 year - parametric imputation	140	154.4	0.87	0.176	[0.62 - 1.23]
Based on 10 year - no imputation	171	171.0	0.93	0.154	[0.69 - 1.26]
Based on 10 year - parametric imputation	171	204.4	0.95	0.160	[0.70 - 1.30]
Based on 16 year - no imputation	202	202.0	0.91	0.141	[0.69 - 1.20]

CI: Confidence interval

* Prognostic factors: age, tumour size, grade, positive nodes, hormonal receptor status, and type of surgery.

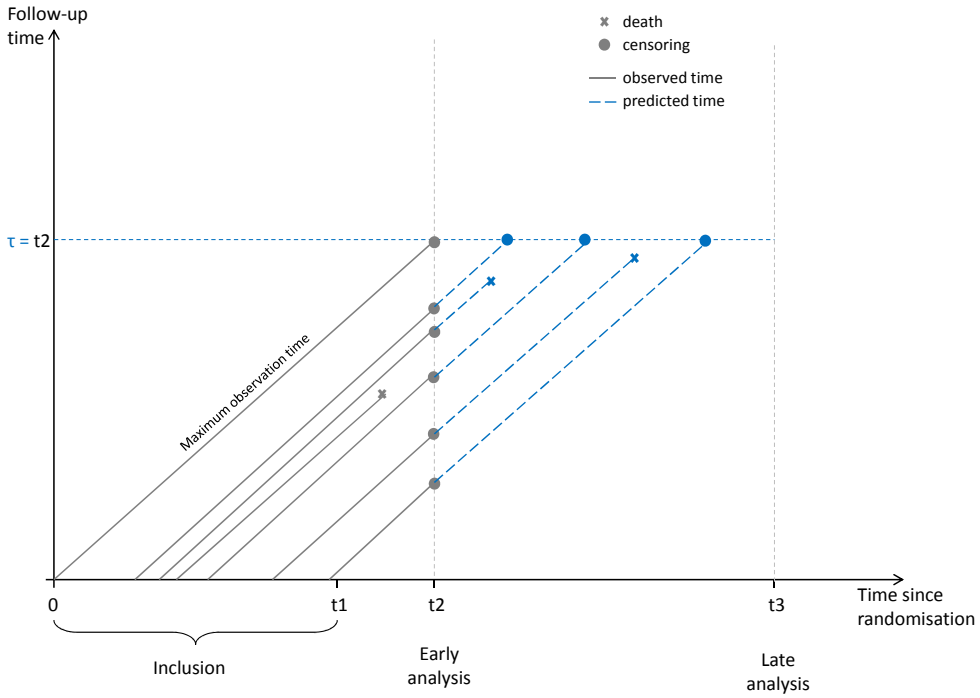

Figure 1: Illustration of the imputation procedure: information is recorded until early analysis time, and used to impute the censored survival time until the maximum observation time τ.

5.3 Additional remarks

Other scenarios are currently being studied, in particular to be closest to the usual power level of 80%.

5.3.1 On the optimization

When computing the imputation based on the parametric prediction of event, we needed to resolve $\hat{P}^D_{i,m}(C_i, T^{D*}_{i,m}; \xi_m) = a_{i,m}$ for $T^{D*}_{i,m} \in [C_i, \tau]$ (see section 2.2.2 of the paper). This equation can not be analytically resolved due to the presence of the integral term which has no close form because of the term α. Thus we used an iterative optimization method to solve this equation. This optimization problem is rather simple: there is only one dimension, and the function is monotonous. Therefore, there is no chance to fall into some local optimum. The optimisation method we used is the one proposed by the *optimize* function in R. This method aims at finding the minimum or the maximum of a function. Thus we actually minimized $f(T^{D*}_{i,m}) = |\hat{P}^D_{i,m}(C_i, T^{D*}_{i,m}; \xi_m) - a_{i,m}|$.

Chapter 6

Frailtypack

frailtypack is an  package (R Core Team, 2013) available on the CRAN since 2005 (Rondeau et al., 2012). It aims at estimating different types of frailty models, including shared or correlated frailty models, joint models for recurrent event and a terminal event, multivariate joint models for two types of recurrent event and a terminal event, and nested models. The package was extended to include part of the work of this thesis as follows.

6.1 Concordance measures

Following the work of this thesis, we added the function *Cmeasures* to the package. It calculates the extension of the Harrell c-index, the Uno c-index and the Gönen and Heller concordance for shared frailty models.

For example, in our application on MACH-NC meta-analysis, the code used for the european dataset was as follows.

```
> shared <- frailtyPenal(Surv(delai,status)~ cluster(groupe) + sexF + age5160
+  + age61plus + stage3 + stage4 +larynx, data=machdev, n.knots=8,
+   kappa=1, cross.validation=TRUE)

Be patient. The program is computing ...
The program took 1.1 seconds

> cindex <-Cmeasures(shared, tau=9.4)

> cindex
```

```
----- Concordance Measures on machdev dataset -----

                   Between Within  Overall
CONDITIONAL
Gonen & Heller's   0.616   0.578   0.613
Uno's              0.638   0.615   0.636

----- Information -----

 Number.patients Number.events Number.groups
            1652          1110            21

         Between Within Overall
Nb pairs 1270643  93083 1363726
```

The available options are as follows:
`Cmeasures(fitc, ties = 1, marginal = 0, cindex = 0, Nboot = 0, tau = 0, data.val)` with `fitc` the model fit (a Cox or a shared model); `ties`: should the ties be included (=1) or excluded (=0) from the concordance calculation; `marginal`: if 1, adds the marginal concordance in the results (by default, only conditional are given); `cindex`: if 1, adds the Harrell c-index values; `Nboot`: Number of bootstrap resamplings for confidence intervals; `tau`: the time limit, if 0, the maximum event time is taken; `data.val`: a data.frame if concordance should be computed on different data than the one used to fit the model –to be used for external validation.

6.2 Prediction in joint models

The function *prediction* was added in the package. It takes as an argument a joint frailty model (*frailtyPenal* object), a time of prediction and a window of prediction. Each of the time and window of prediction can be a vector of values, but not both at the same time. Three probabilities of event corresponding to the three proposed settings are calculated. We have also added a corresponding plot function.

For example, the Table II of the development prediction paper (section 4.2.1) is obtained using the following code:

```
> joint <- frailtyPenal(Surv(tt0,tt1,indR)~cluster(groupe2)
```

```
+age1+age2+emboln+taille+her2n+rhposn+nplusn+grade2+grade3
+ terminal(indDC),
formula.terminalEvent= ~ age1+age2+emboln+taille+her2n+rhposn+nplusn
+grade2+grade3,
data=Btotal, n.knots=4, kappa=c(1000000,13000), recurrentAG=TRUE)

Be patient. The program is computing ...
The program took 37.09 seconds

> # Dataset for prediction - fictional patients (10 patients - 19 rows)
>datapred <- data.frame("tt0"=rep(0,19),    # start time
  "tt1"=c(0,1,2.5,4.9,1,2,2,4,4,4.9,1,2,3,1,2.5,4.9,3,4,4.9),# stop time
  "indR"=c(0,rep(1,18)),                   # recurrence status
  "indDC"=rep(0,19),                       # death status
  "groupe2"=c(1,2,3,4,5,5,6,6,7,7,8,8,8,9,9,9,10,10,10),# patient (cluster)
  "age1"=rep(0,19),"age2"=rep(0,19),       # mean age
  "emboln"=rep(0,19), "taille"=rep(0,19),  # mean PVI and size
  "her2n"=rep(0,19), "rhposn"=rep(1,19),   # mean HER2 and HR status
  "nplusn"=rep(0,19),                      # mean positive nodes status
  "grade2"=rep(1,19),"grade3"=rep(0,19))   # mean grade

> # predicted risk of death between 5 years and 10,
> # and between 5 years and 15 (w=5 and 10)
> predictions <- prediction(joint,datapred,5,c(5,10), MC.sample=500)

Calculating the probabilities ...
Predictions done for 10 subjects and 2 times

> predictions
Call:
prediction(fit = joint, data = datapred, t = 5, window = c(5,
    10), MC.sample = 500)


--------- Prediction 1 (exactly j recurrences) ---------
```

```
       times
ind 1  0.108 0.227
ind 2  0.303 0.530
ind 3  0.303 0.530
ind 4  0.303 0.530
ind 5  0.506 0.756
ind 6  0.506 0.756
ind 7  0.506 0.756
ind 8  0.674 0.884
ind 9  0.674 0.884
ind 10 0.674 0.884

--------- Prediction 2 (at least j recurrences) ---------

       times
ind 1  0.108 0.227
ind 2  0.333 0.562
ind 3  0.323 0.552
ind 4  0.304 0.531
ind 5  0.532 0.775
ind 6  0.515 0.763
ind 7  0.507 0.756
ind 8  0.689 0.892
ind 9  0.675 0.884
ind 10 0.675 0.884

--------- Prediction 3 (only parameters) ---------

       times
ind 1  0.127 0.256
ind 2  0.127 0.256
ind 3  0.127 0.256
ind 4  0.127 0.256
ind 5  0.127 0.256
ind 6  0.127 0.256
ind 7  0.127 0.256
```

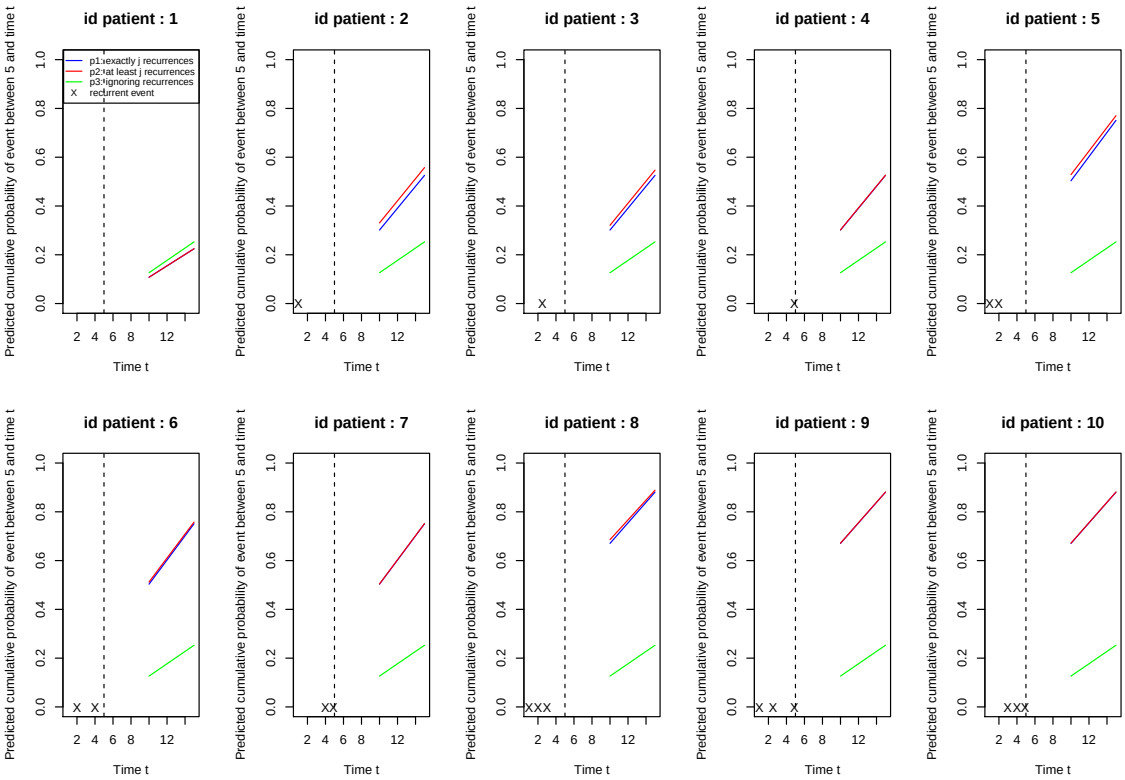

Figure 6.1: Example of plot of predicted probabilities obtained with the package *frailtypack* with a prediction time at 5 years and an horizon equals to 10 and 15 years.

```
ind 8   0.127 0.256
ind 9   0.127 0.256
ind 10 0.127 0.256
```

The available options are as follows:
`prediction(fit, data, t, window, group, MC.sample=0)` with `fit`: the joint model (or cox or shared) used to predict; `data`: the dataset containing the patient for whom prediction has to be made; `t`: the prediction time; `window` the prediction window w; `group`: used to do some conditional predictions in shared frailty model (in joint model, only marginal predictions are proposed); `MC.sample`: number of resampling for confidence interval (=0 if no confidence intervals are wanted).

The command `plot(predictions)` on our example with only two horizon times (5 and 10 years) gives the Figure 6.1. The same prediction with a more continuous horizon from 5.5 to 15 gives the plot in Figure 6.2.

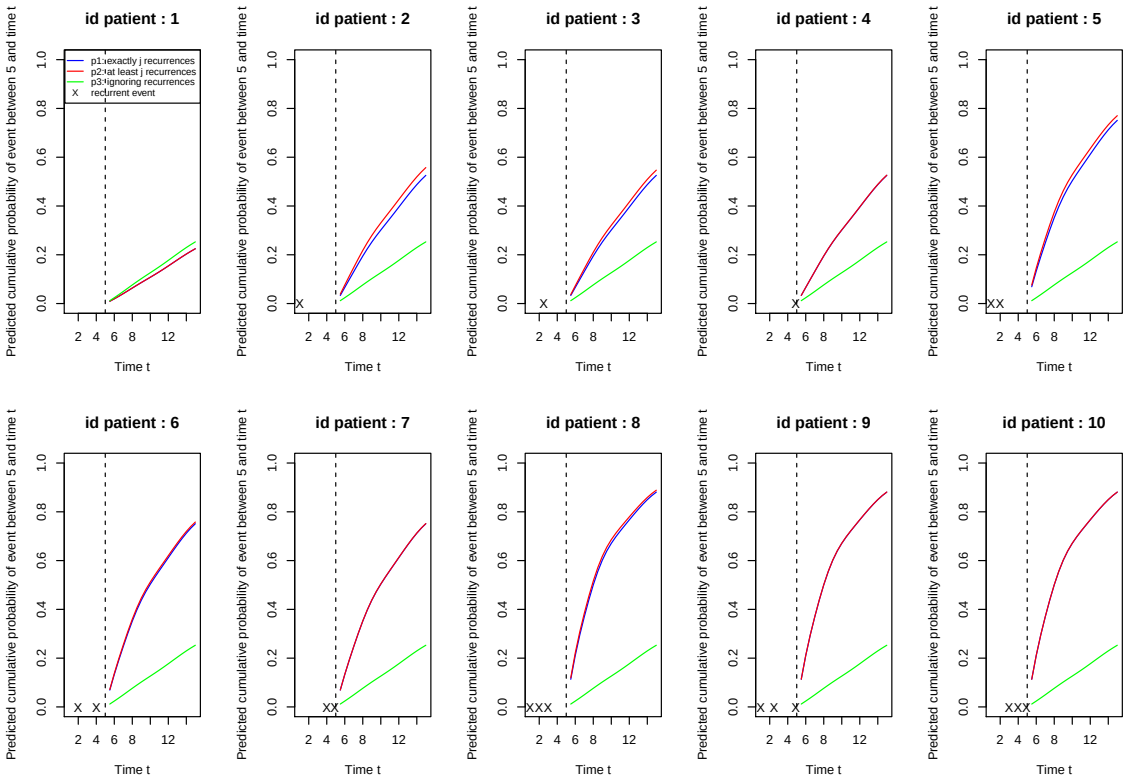

Figure 6.2: Example of plot of predicted probabilities obtained with the package *frailtypack* with a prediction time at 5 years and an horizon going from 5.5 to 15 years (by 0.5 year).

6.3 Some simulations regarding the proposed estimations

During the thesis work, some simulations have been performed to assess the estimation of the parameters. Especially, we assessed the influence of choosing a calendar or a gap time in the joint model. We simulated 100 populations of 500 patients. The baseline hazards were exponential (Weibul with shape=1 and scale=1/0.1). We simulated $\theta = 0.8$ and the $\alpha = 1$. The two coefficients β of the effect of a binary (p=0.5) variable on the recurrent event and the terminal event were both equal to 0.4. The results were as follows (the first line is the gap time and the second line is the calendar time).

```
> # theta (=0.8)
> summary(resJoint$AGtheta); summary(resJoint$theta);
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
 0.5463  0.7427  0.7886  0.7815  0.8173  0.9032
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3143  0.7473  0.7884  0.7822  0.8232  0.8963
> # alpha (=1)
> summary(resJoint$AGalpha);  summary(resJoint$alpha);
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.4532  0.8875  1.0130  1.0420  1.1900  1.6810
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.5214  0.8743  0.9975  1.0320  1.1930  1.6590
> # beta recurrent event (=0.4)
> summary(resJoint$AGcoef.rec); summary(resJoint$coef.rec);
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.03462  0.31870  0.42850  0.42810  0.51090  1.05700
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.03349  0.31340  0.43620  0.42790  0.51850  1.05700
> # beta death(=0.4)
> summary(resJoint$AGcoef.dc);  summary(resJoint$coef.dc);
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.06104 0.31680 0.40900 0.43370 0.54590 0.95390
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.05921 0.31390 0.39670 0.43130 0.55550 0.97030
```

We can see that the choice of the time scale has almost no impact on the parameter estimation. Globally, the coefficients were well-estimated: no bias for the α, the θ was slightly underestimated while the coefficients were slightly overestimated for the death.

Chapter 7

General discussion

7.1 Conclusion on the thesis work

In this thesis, we intended to answer some of the challenges raised by the evolution of cancer research. We specifically wanted to assess the input of joint frailty model in predicting patients survival.

In a first part, we proposed an extension of the concordance measures to clustered data, studied by shared frailty models. This extension made it possible to distinguish between the within-cluster concordance and the between-cluster concordance. These two levels of information answer two questions: 1- how the covariates discriminate between the patients' survival, and 2- in what extent the estimated frailty terms increase this discrimination. One of the limit of this extension is that it is not suitable for the recurrent events data, for which the shared frailty models are also of interest. The specificity of such data is the presence of many clusters but of small size. In addition, the recurrent events are ordered, thus the comparison by pairs without taking the order into account may be a little clumsy.

In a second part, we developed prediction of the risk of death according to the previously observed disease events. We showed that more accurate predictions can be obtained by considering the relapses. We also showed in a validation step that the joint modelling, despite the high number of parameters to be estimated, performed as well as a simpler landmark Cox model. One advantage of the joint model is that it is possible, estimating only one model, to derive dynamic predictions, with a varying prediction time and a varying prediction horizon. It also fully considers the information about relapses.

Finally, in a third part, we illustrated how these predictions can be useful to reduce clinical trial duration. Especially, how predictions can be used to impute the non-

observed death times and conclude earlier and more precisely the treatment effect. Only the sampling in the parametric estimation of the survival distribution appeared suitable. We showed that, as expected, the standard-error of the treatment effect was reduced when using imputations. This did not result in an increased power, but in a preserved type-I error and a better coverage rate for most of the scenarios.

7.2 Critical insight and perspectives

We have shown that the joint modelling approach can be added to the suitable options to study impact of disease recurrences on prediction of death risk. Three main works have previously studied the impact of intermediate disease events on the risk of death in breast cancer patients. Hatteville et al. (2002) proposed a prediction of the risk of death 20 years after the surgery in breast cancer patients, in two steps. The first step consisted in estimated the hazards of events using time-dependent covariates. The second steps was to compute conditional probabilities of death at 20 years, given that an intermediate event occurred before the time of prediction t ($t < 10$ years). They conclude that the survival probability falls from 89% if no event was observed to 9% if both a locoregional relapse and a distant metastasis were observed. An alternative method was a multi-state model as proposed by Putter et al. (2006) to estimate the transition probabilities between the different types of events. This actually allowed to get transition probabilities to death that differ according to the previous event. As compared to the previous method proposed, no internal time-dependent covariate was included in a Cox model. Finally Parast et al. (2011) proposed a non-parametric estimation of the risk of a long-term outcome accounting for the occurrence of a short term outcome. However, in their application on breast cancer patients, Parast and Cai (2013) chose to adapt the landmark approach proposed by Van Houwelingen (2007) with an estimation by Cox model. In this paper, they also derived adequate measures of accuracy. Our work extents the previous ones as it is suitable for recurrent events with many occurrences. It also allows a full comprehension of the two event processes and their inter-dependence. This correlation is then directly taken into account in the prediction. Although all these methods have slightly different approaches and objectives, it would be of great interest to perform a careful comparison of them. We expect the results to depend on the nature of the data and especially the dependence strength between the events, but also the evolution of the covariate effects along time.

Contrary to the multi-state approach, our work makes no distinction between loco-regional relapses and distant metastases so far. However, this two different types of event

should be differently accounted for. We also should consider that the baseline hazard for these two events is not the same. For that purpose, a multivariate joint model was recently proposed (Mazroui et al., 2013). It may be interesting to develop the prediction from this multivariate model to assess whether a more accurate prediction can be obtain by considering separately these two types of events. Another application of joint models may also be to use the loco-regional relapses information to predict the risk of metastasis, as the appearance of metastasis may totally change the disease course. Finally, it may also be of interest to evaluate whether the risk of death is modified in case of multiple successive cancers.

This thesis presents new developments in joint modelling for recurrent events. Most of them have already been proposed for longitudinal data. Indeed, most of the biomarkers measured are continuous variables, such as CD4 count in HIV/AIDS and prostate-specific antigen in prostate cancer. However, many applications are concerned by recurrent events with a terminal event: study of re-hospitalisations, study of epileptic seizures, asthma, and all chronic diseases defined by episodes. The terminal event is often the death. One perspective is to combine longitudinal information, tumour size for example, with the relapses to study the risk of death. This can be done by a multivariate joint model, with three parts.

The scope of this thesis was prognosis research, and to evaluate the input of joint modelling in this area. We thus did not evaluate or discuss much the model itself and its estimation, as it has been studied before (Rondeau et al., 2007). Only few simulations were performed, in which the estimations were accurate.

The joint model may play a role not only in the use of surrogate markers, but also in their validation. Indeed, the meta-analytic approach as proposed by Buyse et al. (2000) is currently done in two steps: one individual level correlation using copulas, and one trial level correlation using linear regression. These two steps may be combined in one step using a joint model with random effect at individual level, and a random treatment effect at trial level. This may lead to more accurate estimations of the correlations between the surrogate and the true endpoints.

Finally, one question should still be resolved: how can we assess the prediction discrimination in the framework of recurrent data? We have seen that the concordance index is not appropriate. We chose to use the prediction error to assess whether the predictions are closed to the reality but the concept is somewhat different to the concept of discrimination. Indeed, the prediction error calculated by a Brier score covers both the discrimination and the calibration concepts. A high prediction error may thus be due to a low discrimination, i.e. the covariates are not predictive enough to separate

patients with different survival times, but may also be due to a low calibration, which can be explained, for example, by a poor estimation of the baseline hazard. As the two concepts do not have the same origin for poor results, and do not have the same solution in case of poor results, it would be interesting to disentangle both.

Bibliography

Aalen, O. (1980). A model for nonparametric regression analysis of counting processes. In D. W. Klonecki, D. A. Kozek, and D. J. Rosinski (Eds.), *Mathematical Statistics and Probability Theory*, Number 2 in Lecture Notes in Statistics, pp. 1–25. Springer, New York.

Aalen, O., O. Borgan, and H. Gjessing (2008). *Survival and Event History Analysis: A Process Point of View*. Springer, New York.

Altman, D. and P. Royston (2000). What do we mean by validating a prognostic model? *Statistics in Medicine 19*(4), 453–473.

Altman, D., Y. Vergouwe, P. Royston, and K. Moons (2009). Prognosis and prognostic research: validating a prognostic model. *BMJ 338*, b605–b605.

Andersen, P., O. Borgan, R. Gill, and N. Keiding (1997). *Statistical Models Based on Counting Processes*. Springer, New York.

Andersen, P. K., M. G. Hansen, and J. P. Klein (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime data analysis 10*(4), 335–350.

Andersen, P. K. and N. Keiding (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine 31*(11-12), 1074–1088.

Andersen, P. K. and M. P. Perme (2010). Pseudo-observations in survival analysis. *Statistical Methods in Medical Research 19*(1), 71–99.

Badve, S., D. J. Dabbs, S. J. Schnitt, et al. (2011). Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists. *Modern Pathology 24*(2), 157–167.

Box-Steffensmeier, J. and S. De Boef (2006). Repeated events survival models: the conditional frailty model. *Statistics in Medicine 25*(20), 3518–3533.

Boyle, P., B. Levin, and International Agency for Research on Cancer (2008). *World cancer report 2008*. IARC Press, Lyon.

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

Burzykowski, T. and M. Buyse (2006). Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics 5*(3), 173–186.

Buyse, M. and G. Molenberghs (1998). Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics 54*(3), 1014–1029.

Buyse, M., G. Molenberghs, T. Burzykowski, D. Renard, and H. Geys (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics 1*(1), 49–67.

Clayton, D. G. (1991, June). A monte carlo method for bayesian inference in frailty models. *Biometrics 47*(2), 467.

Collette, S., L. Collette, T. Budiharto, et al. (2008). Predictors of the risk of fibrosis at 10 years after breast conserving therapy for early breast cancer-a study based on the EORTC trial 22881-10882 'boost versus no boost'. *European Journal of Cancer 44*(17), 2587–2599.

Commenges, D., P. Joly, A. Gégout-Petit, and B. Liquet (2007). Choice between semi-parametric estimators of markov and non-markov multi-state models from coarsened observations. *Scandinavian Journal of Statistics 34*(1), 33–52.

Commenges, D., B. Liquet, and C. Proust-Lima (2012). Choice of prognostic estimators in joint models by estimating differences of expected conditional kullback–leibler risks. *Biometrics 68*(2), 380–387.

Conlon, A. S. C., J. M. G. Taylor, D. J. Sargent, and G. Yothers (2011). Using cure models and multiple imputation to utilize recurrence as an auxiliary variable for overall survival. *Clinical trials 8*(5), 581–590.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), 187–220.

Ducrocq, V. and G. Casella (1996). A bayesian analysis of mixed survival models. *Genetics, Selection, Evolution 28*, 505–529.

Faucett, C. L., N. Schenker, and J. M. Taylor (2002). Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. *Biometrics 58*(1), 37–47.

Ferlay, J., I. Soerjomataram, M. Ervik, et al. (2013). GLOBOCAN 2012 v1.0, cancer incidence and mortality worldwide: IARC CancerBase no. 11 [internet]. lyon, france: International agency for research on cancer. available from http://globocan.iarc.fr [consulted on 26-8-2014].

Freedman, L. S., B. I. Graubard, and A. Schatzkin (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in medicine 11*(2), 167–178.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2013). *Bayesian Data Analysis, Third Edition*. CRC Press, Boca Raton.

Gerds, T. and M. Schumacher (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal 48*(6), 1029–1040.

Gönen, M. and G. Heller (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika 92*(4), 965–970.

Graf, E., C. Schmoor, W. Sauerbrei, and M. Schumacher (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine 18*(17-18), 2529–2545.

Ha, I. and Y. Lee (2005). Comparison of hierarchical likelihood versus orthodox best linear unbiased predictor approaches for frailty models. *Biometrika 92*(3), 717 –723.

Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer, New York.

Harrell, F. E., R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati (1982). Evaluating the yield of medical tests. *JAMA: The Journal of the American Medical Association 247*(18), 2543–2546.

Harrell, F. E., K. L. Lee, and D. B. Mark (1996). Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine 15*(4), 361–387.

Hatteville, L., C. Mahe, and C. Hill (2002). Prediction of the long-term survival in breast cancer patients according to the present oncological status. *Statistics in Medicine 21*(16), 2345–2354.

Hemingway, H., P. Croft, P. Perel, et al. (2013). Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ : British Medical Journal 345*, e5595.

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

Hingorani, A., D. Windt, R. Riley, et al. (2013). Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *BMJ : British Medical Journal 345*, e5793.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime Data Analysis 1*(3), 255–273.

Huang, C. Y., J. Qin, and M. C. Wang (2009). Semiparametric analysis for recurrent event data with time-dependent covariates and informative censoring. *Biometrics 66*(1), 39–49.

Huang, X. and L. Liu (2007). A joint frailty model for survival and gap times between recurrent events. *Biometrics 63*(2), 389–397.

IARC (2002). *IARC Handbooks of Cancer Prevention. Vol.7: Breast Cancer Screening.* (IARCPress, Lyon ed.). Vainio H and Bianchini F.

Kalbfleisch, J. and R. Prentice (2002). *The statistical analysis of failure time data.* John Wiley, Hoboken.

Kim, S., D. Zeng, L. Chambless, and Y. Li (2012). Joint models of longitudinal data and recurrent events with informative terminal event. *Statistics in biosciences 4*(2), 262–281.

Klein, J. (1992). Semiparametric estimation of random effects using the cox model based on the EM algorithm. *Biometrics 48*(3), 795–806.

Konig, I., J. Malley, C. Weimar, H. Diener, and A. Ziegler (2007). Practical experiences on the necessity of external validation. *Statistics in Medicine 26*(30), 5499–5511.

Korn, E. and R. Simon (1990). Measures of explained variation for survival data. *Statistics in Medicine 9*(5), 487–503.

Krommer, A. R. and C. W. Ueberhuber (1998). *Computational Integration*. Society for Industrial and Applied Mathematics, Philadelphia.

Lam, K. and D. Ip (2003). REML and ML estimation for clustered grouped survival data. *Statistics in medicine 22*(12), 2025–2034.

Lawless, J. F. and Y. Yuan (2010). Estimation of prediction error for survival models. *Statistics in Medicine 29*(2), 262–274.

Legrand, C., L. Duchateau, P. Janssen, V. Ducrocq, and R. Sylvester (2009). Validation of prognostic indices using the frailty model. *Lifetime Data Analysis 15*(1), 59–78.

Liu, L., R. Wolfe, and X. Huang (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics 60*(3), 747–756.

Logan, B. R., M.-J. Zhang, and J. P. Klein (2011). Marginal models for clustered time-to-event data with competing risks using pseudovalues. *Biometrics 67*(1), 1–7.

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics 11*(2), 431–441.

Martinussen, T. and T. H. Scheike (2006). *Dynamic Regression Models for Survival Data* (1st ed.). Springer, New York.

Mattei, A., F. Mealli, and D. B. Rubin (2012). Missing data and imputation methods. In *Modern Analysis of Customer Surveys: with Applications using R, First edition.* (John Wiley & Sons, Ltd, Hoboken ed.). Ron S. Kenett and Silvia Salini.

Mauguen, A., S. Collette, J.-P. Pignon, and V. Rondeau (2013). Concordance measures in shared frailty models: application to clustered data in cancer prognosis. *Statistics in Medicine 32*(27), 4803–4820.

Mauguen, A., S. Collette, J.-P. Pignon, and V. Rondeau (2014). Reply to 'interpretation of concordance measures for clustered data'. *Statistics in Medicine 33*(4), 717–718.

Mauguen, A., B. Rachet, S. Mathoulin-Pélissier, G. MacGrogan, A. Laurent, and V. Rondeau (2013). Dynamic prediction of risk of death using history of cancer recurrences in joint frailty models. *Statistics in Medicine 32*(30), 5366–5380.

Mazroui, Y., S. Mathoulin-Pélissier, G. MacGrogan, V. Brouste, and V. Rondeau (2013). Multivariate frailty models for two types of recurrent events with a dependent terminal event: Application to breast cancer data. *Biometrical Journal 55*(6), 866–884.

Mazroui, Y., S. Mathoulin-Pélissier, P. Soubeyran, and V. Rondeau (2012). General joint frailty model for recurrent event data with a dependent terminal event: Application to follicular lymphoma data. *Statistics in Medicine 31*(11-12), 1162–1176.

McLain, A. C., K. J. Lum, and R. Sundaram (2012). A joint mixed effects dispersion model for menstrual cycle length and time-to-pregnancy. *Biometrics 68*(2), 648–656.

Molinaro, A. M., R. Simon, and R. M. Pfeiffer (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics 21*(15), 3301–3307.

Moons, K., D. Altman, Y. Vergouwe, and P. Royston (2009). Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ 338*, b606–b606.

Moons, K., A. Kengne, D. Grobbee, et al. (2012). Risk prediction models: II. external validation, model updating, and impact assessment. *Heart 98*(9), 691–698.

Moons, K., P. Royston, Y. Vergouwe, D. Grobbee, and D. Altman (2009). Prognosis and prognostic research: what, why, and how? *BMJ 338*, b375–b375.

Moons, K. G. M., A. P. Kengne, M. Woodward, et al. (2012). Risk prediction models: I. development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart 98*(9), 683–690.

Munda, M. and C. Legrand (2014). Adjusting for centre heterogeneity in multicentre clinical trials with a time-to-event outcome. *Pharmaceutical Statistics 13*(2), 145–152.

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

Parast, L. and T. Cai (2013). Landmark risk prediction of residual life for breast cancer survival. *Statistics in Medicine 32*(20), 3459–3471.

Parast, L., S. Cheng, and T. Cai (2011). Incorporating short-term outcome information to predict long-term survival with discrete markers. *Biometrical Journal 53*(2), 294–307.

Parast, L., L. Tian, and T. Cai (2014). Landmark estimation of survival and treatment effect in a randomized clinical trial. *Journal of the American Statistical Association 109*(505), 384–394.

Pencina, M., R. D' Agostino Sr, R. D' Agostino Jr, and R. Vasan (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine 27*(2), 157–172.

Pickles, A. and R. Crouchley (1995). A comparison of frailty models for multivariate survival data. *Statistics in Medicine 14*(13), 1447–1461.

Pignon, J.-P., A. Le Maitre, E. Maillard, et al. (2009). Meta-analysis of chemotherapy in head and neck cancer (MACH-NC): an update on 93 randomised trials and 17,346 patients. *Radiotherapy and Oncology 92*(1), 4–14.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine 8*(4), 431–440.

Proust-Lima, C. and J. Taylor (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. *Biostatistics 10*(3), 535–549.

Putter, H., J. van der Hage, G. de Bock, R. Elgalta, and C. van de Velde (2006). Estimation and prediction in a multi-state model for breast cancer. *Biometrical journal 48*(3), 366–380.

R Core Team, . (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science 3*(4), 425–441.

Ravdin, P. M., L. A. Siminoff, G. J. Davis, et al. (2001). Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *Journal of Clinical Oncology 19*(4), 980–991.

Riley, R., J. Hayden, E. Steyerberg, et al. (2013). Prognosis research strategy (PROGRESS) 2: Prognostic factor research. *PLoS Med 10*(2), e1001380.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics 67*(3), 819–829.

Rondeau, V., D. Commenges, and P. Joly (2003). Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime Data Analysis 9*(2), 139–153.

Rondeau, V., A. Laurent, A. Mauguen, C. Berr, and C. Helmer (2014). Dynamic prediction models for clustered and interval-censored outcomes: investigating the intra-couple correlation in the risk of dementia. *Submitted* ., .

Rondeau, V., S. Mathoulin-Pélissier, H. Jacqmin-Gadda, V. Brouste, and P. Soubeyran (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics 8*(4), 708–721.

Rondeau, V., Y. Mazroui, and J. R. Gonzalez (2012). frailtypack: an r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software 47*(4), 1–28.

Royston, P., K. Moons, D. Altman, and Y. Vergouwe (2009). Prognosis and prognostic research: Developing a prognostic model. *BMJ 338*, b604–b604.

Royston, P. and M. K. B. Parmar (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine 30*(19), 2409–2421.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association 91*(434), 473–489.

Rubin, D. B. and N. Schenker (1991). Multiple imputation in health-are databases: An overview and some applications. *Statistics in Medicine 10*(4), 585–598.

Schemper, M. and R. Henderson (2000). Predictive accuracy and explained variation in cox regression. *Biometrics 56*(1), 249–255.

Simon, R. and D. G. Altman (1994). Statistical aspects of prognostic factor studies in oncology. *British Journal of Cancer 69*(6), 979–985.

Sinha, D., T. Maiti, J. G. Ibrahim, and B. Ouyang (2008). Current methods for recurrent events data with dependent termination. *Journal of the American Statistical Association 103*(482), 866–878.

Stare, J., M. Perme, and R. Henderson (2011). A measure of explained variation for event history data. *Biometrics 67*(3), 750–759.

Steyerberg, E. (2010). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating* (2010 ed.). Springer, New York.

Steyerberg, E., K. Moons, D. van der Windt, and others for the PROGRESS Group (2013). Prognosis research strategy (PROGRESS) 3: Prognostic model research. *PLoS Medicine 10*(2), e1001381.

Taylor, J. M. G., Y. Park, D. P. Ankerst, et al. (2013). Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics 69*(1), 206–213.

Therneau, T. and P. Grambsch (2000). *Modeling survival data: extending the Cox model.* Springer, New York.

Uno, H., T. Cai, M. Pencina, R. D'Agostino, and L. Wei (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine 30*(10), 1105–1117.

Van Houwelingen, H. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics 34*(1), 70–85.

Van Houwelingen, H. and H. Putter (2011). *Dynamic Prediction in Clinical Survival Analysis* (1 ed.). CRC Press Inc, Boca Raton.

van Klaveren, D., E. W. Steyerberg, P. Perel, and Y. Vergouwe (2014). Assessing discriminative ability of risk models in clustered data. *BMC Medical Research Methodology 14*(1), 5.

van Klaveren, D., E. W. Steyerberg, and Y. Vergouwe (2014). Interpretation of concordance measures for clustered data. *Statistics in medicine 33*(4), 714–716.

Van Oirbeek, R. and E. Lesaffre (2010). An application of harrell's c-index to PH frailty models. *Statistics in Medicine 29*(30), 3160–3171.

Van Zee, K. J., U. Rudloff, E. Brogi, and S. Patil (2011). Reply to c. mazouni et al. *Journal of Clinical Oncology 29*(2), e45–e46.

Varadhan, R., Q.-L. Xue, and K. Bandeen-Roche (2014). Semicompeting risks in aging research: methods, issues and needs. *Lifetime data analysis 20*, 538–562.

Wishart, G. C., E. M. Azzato, D. C. Greenberg, et al. (2010). PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Research 12*, R1.

Ye, Y., J. D. Kalbfleisch, and D. E. Schaubel (2007). Semiparametric analysis of correlated recurrent and terminal events. *Biometrics 63*(1), 78–87.

Zeng, D. and J. Cai (2010). A semiparametric additive rate model for recurrent events with an informative terminal event. *Biometrika 97*(3), 699–712.

Zeng, D. and D. Y. Lin (2009). Semiparametric transformation models with random effects for joint analysis of recurrent and terminal events. *Biometrics 65*(3), 746–752.

Zhangsheng, Y. and L. Liu (2011). A joint model of recurrent events and a terminal event with a nonparametric covariate function. *Statistics in Medicine 30*(22), 2683–2695.

Zhao, X., L. Liu, Y. Liu, and W. Xu (2012). Analysis of multivariate recurrent event data with time-dependent covariates and informative censoring. *Biometrical journal 54*(5), 585–599.

Appendixes

Appendix A: Description of the dataset used in the concordance application

EORTC dataset

Table 1: Description of clusters in EORTC data.

Institution	Boost arm		No Boost arm	
	N	N fibrosis	N	N fibrosis
1	8	3	6	1
2	233	46	226	25
3	59	8	61	10
4	90	27	80	11
5	30	10	44	6
6	207	48	210	25
7	69	24	67	7
8	61	6	56	1
9	122	76	124	24
10	57	21	45	6
11	37	27	42	29
12	177	48	178	16
13	146	78	137	47
14	129	64	129	15
15	70	26	67	15
16	410	51	406	20
17	302	13	297	6
18	90	30	90	14
19	51	27	55	15
20	8	3	12	2
21	68	40	73	26

MACH-NC dataset

Table 2: Description of clusters in MACH-NC data.

Trials	N	N deaths	Median survival
Europe (development data)			
1	113	96	1.0
2	62	47	1.8
3	78	62	1.2
4	127	103	1.4
5	70	50	1.6
6	194	146	1.3
7	111	65	2.4
8	161	92	2.6
9	56	42	1.1
10	43	36	1.0
11	55	20	3.1
12	100	51	5.0
13	34	19	3.1
14	172	108	2.4
15	61	36	2.7
16	113	77	2.3
17	47	35	2.3
18	16	5	Not reached
19	12	4	Not reached
20	14	9	1.1
21	13	7	1.3
United States (validation data)			
1	50	33	4.6
2	59	47	2.0
3	228	131	2.8
4	184	79	5.2
5	101	89	1.0
Others (validation data)			
1	23	20	0.5
2	227	132	1.0
3	17	11	1.2

Appendix B: Development steps of the concordance probability estimator (Gönen and Heller)

The concordance probability in frailty models is defined by: (global development, then extended to the between-group concordance)

$$K(\beta) = \mathbb{P}(T_2 > T_1 | \mathbf{ln(u_i)} + \beta' x_1 \geq \mathbf{ln(u_j)} + \beta' x_2) = \frac{\mathbb{P}(T_2 > T_1, \mathbf{ln(u_i)} + \beta' x_1 \geq \mathbf{ln(u_j)} + \beta' x_2)}{\mathbb{P}(\mathbf{ln(u_i)} + \beta' x_1 \geq \mathbf{ln(u_j)} + \beta' x_2)}$$

First step: $\mathbb{P}\left\{T(\beta' x_2, u_j) > T(\beta' x_1, u_i)\right\}$

$$\begin{aligned}
&\mathbb{P}\left\{T(\beta' x_2, u_j) > T(\beta' x_1, u_i)\right\} \\
&= \int_0^\infty S(t; x_2, \beta, u_j) f(t; x_1, \beta, u_i) \mathrm{dt} \\
&= -\int_0^\infty S(t; x_2, \beta, u_j) \frac{\mathrm{d}S(t; x_1, \beta, u_i)}{\mathrm{dt}} \mathrm{dt} \\
&= -\int_0^\infty \exp\left\{-u_j e^{\beta' x_2} \Lambda_0(t)\right\} \frac{\mathrm{d}\exp\left\{-u_i e^{\beta' x_1} \Lambda_0(t)\right\}}{\mathrm{dt}} \mathrm{dt} \\
&= -\int_0^\infty \exp\left\{-u_j e^{\beta' x_2} \Lambda_0(t)\right\} \left(-u_i e^{\beta' x_1}\right) \lambda_0(t) \exp\left\{-u_i e^{\beta' x_1} \Lambda_0(t)\right\} \mathrm{dt} \\
&= \frac{-u_i e^{\beta' x_1}}{u_j e^{\beta' x_2} + u_i e^{\beta' x_1}} \int_0^\infty \underbrace{-[u_j e^{\beta' x_2} + u_i e^{\beta' x_1}] \lambda_0(t)}_{v'} \underbrace{exp\{-\Lambda_0(t)[u_j e^{\beta' x_2} + u_i e^{\beta' x_1}]\}}_{e^v} \mathrm{dt} \\
&= \frac{-u_i e^{\beta' x_1}}{u_j e^{\beta' x_2} + u_i e^{\beta' x_1}} \left[\exp\left\{-\Lambda_0(t)(u_j e^{\beta' x_2} + u_i e^{\beta' x_1})\right\}\right]_0^\infty \\
&= \frac{-1}{\frac{u_j e^{\beta' x_2}}{u_i e^{\beta' x_1}} + 1} (0 - 1) \\
&= \frac{1}{1 + \frac{\mathbf{u_j}}{\mathbf{u_i}} e^{\beta'(x_2 - x_1)}} = \frac{1}{1 + e^{ln(u_j) + \beta' x_2 - [ln(u_i) + \beta' x_1]}}
\end{aligned}$$

Second step: $\mathbb{P}\left\{ln(u_i) + \beta' x_1 \geq ln(u_j) + \beta' x_2\right\}$

$$
\begin{aligned}
&\mathbb{P}\left\{ln(u_i)+\beta' x_1 \geq ln(u_j)+\beta' x_2\right\} \\
&= \int_{\mathbb{R}} \mathbb{P}\left\{ln(u_i)+\beta' x_1 \geq ln(u_j)+\beta' x_2, \quad ln(u_j)+\beta' x_2 = b\right\} \mathrm{db} \\
&= \int_{\mathbb{R}} \mathbb{P}\left\{ln(u_i)+\beta' x_1 \geq ln(u_j)+\beta' x_2 | ln(u_j)+\beta' x_2 = b\right\} \mathbb{P}\left\{ln(u_j)+\beta' x_2 = b\right\} \mathrm{db} \\
&= \int_{\mathbb{R}} \mathbb{P}\left\{ln(u_i)+\beta' x_1 \geq b\right\} \mathrm{dF}\left(ln(u_j)+\beta' x_2\right) \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{P}\left\{ln(u_i)+\beta' x_1 \geq b, ln(u_i)+\beta' x_1 = a\right\} \mathrm{dF}\left(ln(u_j)+\beta' x_2\right) \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{P}\left\{ln(u_i)+\beta' x_1 \geq b | ln(u_i)+\beta' x_1 = a\right\} \mathbb{P}\left\{ln(u_i)+\beta' x_1 = a\right\} \mathrm{dadF}\left(ln(u_j)+\beta' x_2\right) \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{P}\left\{a \geq b\right\} \mathbb{P}\left\{ln(u_i)+\beta' x_1 = a\right\} \mathrm{da} \quad \mathrm{dF}\left(ln(u_j)+\beta' x_2\right) \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{P}\left\{a \geq b\right\} \mathrm{dF}\left(ln(u_i)+\beta' x_1\right) \mathrm{dF}\left(ln(u_j)+\beta' x_2\right) \\
&= \int_{-\infty}^{ln(u_i)+\beta' x_1} \int_{\mathbb{R}} \mathrm{dF}\left(ln(u_i)+\beta' x_1\right) \mathrm{dF}\left(ln(u_j)+\beta' x_2\right)
\end{aligned}
$$

Finally we get:

$$
\begin{aligned}
K(\beta) &= \mathbb{P}\left\{T_2 > T_1, ln(u_i)+\beta' x_1 \geq ln(u_j)+\beta' x_2\right\} \\
&= \frac{\int_{-\infty}^{ln(u_i)+\beta' x_1} \int_{\mathbb{R}} \left[1+\frac{u_j}{u_i} e^{\beta'(x_2-x_1)}\right]^{-1} \mathrm{dF}\left(ln(u_i)+\beta' x_1\right) \mathrm{dF}\left(ln(u_j)+\beta' x_2\right)}{\int_{-\infty}^{ln(u_i)+\beta' x_1} \int_{\mathbb{R}} \mathrm{dF}\left(ln(u_i)+\beta' x_1\right) \mathrm{dF}\left(ln(u_j)+\beta' x_2\right)}
\end{aligned}
$$

The estimator of $K(\beta)$ becomes:

$$
K_n(\hat{\beta}, \hat{u}) = \frac{2}{n(n-1)} \sum \sum_{i<j} \left\{ \frac{I(ln(\frac{u_j}{u_i}) + \hat{\beta}' x_{ji} < 0)}{1+\frac{u_j}{u_i} e^{\hat{\beta}' x_{ji}}} + \frac{I(ln(\frac{u_i}{u_j}) + \hat{\beta}' x_{ij} < 0)}{1+\frac{u_i}{u_j} e^{\hat{\beta}' x_{ij}}} \right\}
$$

Appendix C: Concordance measures interpretation: letter to the editor and response

Statistics
in Medicine

Letter to the Editor

(wileyonlinelibrary.com) DOI: 10.1002/sim.5928 Published online in Wiley Online Library

Interpretation of concordance measures for clustered data

David van Klaveren,*† Ewout W. Steyerberg and Yvonne Vergouwe

Mauguen *et al.* [1] extended two censoring-robust estimators of the concordance probability to frailty models. Uno *et al.* [2] and Gönen and Heller [3] proposed the estimators ('Uno' and 'GH' estimators, respectively). The authors followed the suggestion of Van Oirbeek and Lesaffre [4] to derive separate concordance probability estimates for patients within the same cluster and for patients in different clusters and to pool them into an overall estimate. Although the proposed techniques add to the assessment of prognostic model performance in clustered survival data, we would like to discuss three issues related to their interpretation and practical use.

First, the model-based GH estimator does not use observed survival times directly in contrast to Harrell's c-index [5] and the Uno estimator. Instead, the effect of observed survival times is mediated through the regression coefficients. As a consequence, the concordance probability in a new population is estimated under the assumption that the regression coefficients are correct. The GH estimator should therefore be interpreted with care when applied to new populations. The authors applied the GH estimator in clusters of a validation population, using the regression coefficients of the development population. The resulting GH estimates are similar to benchmark estimates suggested before [6] and differ only from the concordance probability estimates in the development population due to differences in patient heterogeneity (case-mix). We undertook a small simulation study with different external validation settings to illustrate the interpretation of the GH estimator (Table I). When both case-mix distribution (X) and coefficient (β) were equal to the development population (validation 1), the concordance probability estimates gave similar results as in the development setting, apart from small differences due to sensitivity for censoring. When we lowered case-mix heterogeneity (validation 2), all concordance measures decreased similarly. When we lowered the coefficient (validation 3), the GH and the Benchmark estimates remained almost the same while the c-index and the Uno estimate decreased further, empirically supporting the aforementioned reasoning.

Second, the authors recommended using cluster-specific (conditional) predictions for validation of a prognostic model. They suggested using the validation data to estimate frailties for new clusters. However, using validation data to derive predictions does not correspond to a direct external validation of a prognostic model's performance in new settings. It might better be labeled a form of internal validation [7]. We recommend to use population (marginal) predictions for external validation and to limit the use of cluster-specific predictions to temporal validation, with frailties estimated on development data and validated on more recent data from the same clusters.

Third, the authors did not give guidance when to use within-cluster, between-cluster, or overall concordance measures. We propose using the within-cluster concordance probability in clinical practice as decisions on interventions are commonly taken within centers (clusters). A valuable prognostic model should be able to separate patients within the same center into those with good outcome and poor outcome. In contrast, we consider the overall concordance measure appropriate when decisions are taken at the population level, where between-center heterogeneity can be used to guide decision making.

Department of Public Health, Erasmus MC, Rotterdam, The Netherlands
**Correspondence to: David van Klaveren, Department of Public Health, Erasmus MC, Dr. Molewaterplein 50, 3015 GE Rotterdam, The Netherlands.*
†*E-mail: d.vanklaveren.1@erasmusmc.nl*

Statistics in Medicine

Table I. Simulation study results; means (empirical standard errors) of concordance probability estimates in a development setting and three validation settings.

	Development $X \sim Unif[0,1]$ $\beta = 3$	Validation 1 $X \sim Unif[0,1]$ $\beta = 3$	Validation 2 $X \sim Unif[0.1,0.9]$ $\beta = 3$	Validation 3 $X \sim Unif[0.1,0.9]$ $\beta = 2$
Censoring (%)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
$\hat{\beta}$	3.015 (0.221)			
c-index	0.710 (0.013)	0.710 (0.013)	0.677 (0.014)	0.625 (0.015)
Uno	0.710 (0.013)	0.710 (0.013)	0.677 (0.014)	0.625 (0.015)
GH	0.710 (0.012)	0.710 (0.011)	0.678 (0.011)	0.678 (0.011)
Benchmark		0.711 (0.017)	0.678 (0.018)	0.678 (0.018)
Censoring (%)	50.4 (2.4)	50.7 (2.6)	50.6 (2.5)	64.2 (2.5)
$\hat{\beta}$	3.026 (0.295)			
c-index	0.721 (0.019)	0.720 (0.018)	0.683 (0.020)	0.628 (0.025)
Uno	0.717 (0.018)	0.716 (0.017)	0.681 (0.019)	0.629 (0.025)
GH	0.710 (0.015)	0.711 (0.015)	0.678 (0.014)	0.678 (0.014)
Benchmark		0.717 (0.021)	0.682 (0.021)	0.683 (0.021)

For each setting, 1000 replications of 400 patient profiles X were drawn from a uniform distribution (column heading). Survival times were generated by multiplication of $\exp(X\beta)$ with independent draws from the exponential distribution (β in column headings). Right-censoring times were drawn from a uniform distribution with support [0,c], where c was chosen to target 0% and 50% censoring. Concordance probability estimates were based on predictions $X\hat{\beta}$ with $\hat{\beta}$ estimated in the development data. The time-dependent Uno estimator was calculated at $\tau = 0.9c$. To obtain the Benchmark estimate, we calculated the predicted survival function for each patient in X based on the model fit in the development data. The predicted survival functions were used to sample 400 survival times. The Benchmark estimate was then calculated as the c-index in this new sample.

Following the same line of reasoning when patient data are clustered in clinical trials, we recommend using the within-cluster concordance probability. Our rationale is that between-trial heterogeneity is not exploitable in clinical practice.

We dispute the authors' conclusion in a head and neck cancer case study that external validation in a US population confirmed the performance of a prognostic model developed in a European population. Regardless of the GH overall concordance probability estimate based on cluster-specific predictions (0.625), we consider the Uno within-cluster probability estimates the most appropriate indicators of discriminative ability of the proposed prognostic model. These estimates were significantly lower in the US validation population (mean 0.488) than in the European development population (mean 0.615). Furthermore, these estimates were similar for the frailty model and the Cox model, but varied substantially across clusters, both in the European and in the US population. The difference between the GH estimates of the within-cluster concordance (0.570) and the between-cluster concordance (0.612) reflected substantially stronger heterogeneity between patients from different clusters than between patients within the same cluster.

In conclusion, for external validation in clinical practice, we recommend using nonparametric within-cluster concordance probability estimates (c-index or Uno), without using cluster-specific (conditional) predictions. The use of GH estimates is valuable for benchmark purposes. Between-cluster concordance probability estimates may be useful when between-cluster heterogeneity in case-mix is exploitable for guidance of decision making.

Acknowledgement

This work was supported by the Netherlands Organisation for Scientific Research (grant 917.11.383).

References

1. Mauguen A, Collette S, Pignon J-P, Rondeau V. Concordance measures in shared frailty models: application to clustered data in cancer prognosis. *Statistics in Medicine* 2013. DOI: 10.1002/sim.5852.

2. 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. *Statistics in Medicine* 2011; **30**:1105–1117.
3. Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 2005; **92**:965–970.
4. Van Oirbeek R, Lesaffre E. An application of Harrell's C-index to PH frailty models. *Statistics in Medicine* 2010; **29**:3160–3171.
5. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982; **247**:2543–2546.
6. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *American Journal of Epidemiology* 2010; **172**:971–980.
7. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine* 2000; **19**:453–473.

Statistics in Medicine

(wileyonlinelibrary.com) DOI: 10.1002/sim.6022 Published online in Wiley Online Library

Reply to 'Interpretation of concordance measures for clustered data'

We thank the authors of the letter to the editor [1] for their insightful comments on our paper on concordance measures in shared frailty models [2]. We intend here to answer some questions and to clarify related issues concerning the interpretation and the use of concordance measures in the context of clustered data.

Concerning the interpretation of the Gönen and Heller estimator, we agree that the interpretation should be done cautiously as it supposed that the Cox prognostic model used is correct, and the regression parameters are well estimated. This was previously mentioned in the discussion of our original paper [2]. We thank Van Klaveren *et al.* for bringing new input with simulation results.

In a second point, they suggested that using the validation data to estimate the frailties of new cluster in conditional predictions prevents for external validation and should be seen as a form of an internal validation. However, the frailties can be seen as unobservable covariates. Thus, in the same way than doing prediction validation in new dataset requires to use the covariate of patients from the new dataset it seems natural to derive new frailties for the new group using observed information from the new group. These new frailties are derived conditionally from the values of the covariates in the validation population but both the variance θ and the covariate effects β are estimated on the development population (see formula (1) of our article [2]). Only the covariates are taken from the validation population. Thus, if it is correct that marginal prediction can be used and can be useful, as shown in [3], we think that conditional predictions are also adequate to do external validation on an independent dataset. In the discussion, we stated that 'Making conditional prediction on new groups illustrates the ability of the proposed method to externally validate the prediction model.' This sentence illustrates the fact that it is possible (but not required) to investigate external validation using the prediction of the random effect on new groups. Finally, care must be taken in keeping same frailties to do some temporal validation, as frailties can evolve with time, even in a given cluster.

Considering the use of within or between-cluster measures only, we think it is always interesting to give both. The two give interesting information, and the comparison of the two measures can bring additional information, for example, on the impact of the frailty term on the prediction.

We take advantage of this response to clarify that, as specified by Van Oirbeek and Lesaffre [4] and in the section method of our paper, the within concordance is the same in the conditional and marginal level. Only the between-cluster measure differs, and even if it is not used to take some final decision, it is still interesting to display.

AUDREY MAUGUEN
INSERM U897
Epidémiologie-Biostatistique
Institut de Santé Publique, d'Epidémiologie et de Développement
146 rue Leo Saignat
Bordeaux 33076, France

SANDRA COLLETTE
European Organisation for Research and Treatment of Cancer
Avenue Emmanuel Mounier 83/11
Brussels 1200, Belgium

JEAN-PIERRE PIGNON
Institut Gustave-Roussy
Department of Biostatistics and Epidemiology
114 rue Edouard Vaillant
Villejuif 94805, France

VIRGINIE RONDEAU
INSERM U897
Epidémiologie-Biostatistique
Institut de Santé Publique, d'Epidémiologie et de Développement
146 rue Leo Saignat
Bordeaux 33076, France

References

1. van Klaveren D, Steyerberg E, Vergouwe Y. Interpretation of concordance measures for clustered data. *Statistics in Medicine* 2013. DOI: 10.1002/sim.5928.
2. Mauguen A, Collette S, Pignon J-P, Rondeau V. Concordance measures in shared frailty models: application to clustered data in cancer prognosis. *Statistics in Medicine* 2013. DOI: 10.1002/sim.5852.
3. Mauguen A, Rachet B, Mathoulin-Pélissier S, MacGrogan G, Laurent A, Rondeau V. Dynamic prediction of risk of death using history of cancer recurrences in joint frailty models. *Statistics in Medicine* 2013. DOI: 10.1002/sim.5980.
4. Van Oirbeek R, Lesaffre E. An application of Harrell's C-index to PH frailty models. *Statistics in Medicine* 2010; **29**:3160–3171.

Appendix D: Detailed calculation for the prediction 2 and 3

Prediction 2

We calculate the probability of death between t and $t+w$ considering at least J recurrences $P^2(t, t+w; \xi)$. We use the partial recurrence history of the patient $\mathcal{H}_i^{J,2}(t) = \{N_i^R(t) \geq J, X_{ik} | X_{ik} \leq t, X_{ik} > X_{i(k-1)}, \forall k = 1, \ldots, J \}$, with $X_{i0} = 0$.

Following the same steps as for the prediction 1 (Mauguen et al., 2013), we have:

$$
\begin{aligned}
&P^2(t, t+w; \xi) \\
&= P(D_i \leq t+w | D_i > t, \mathcal{H}_i^{J,2}(t), Z_{ij}^R, Z_i^D, \xi) \\
&= \int_0^\infty P(D_i \leq t+w | D_i > t, \mathcal{H}_i^{J,2}(t), Z_{ij}^R, Z_i^D, u_i, \xi) \times g(u_i | D_i > t, \mathcal{H}_i^{J,2}(t), Z_{ij}^R, Z_i^D, \xi) \mathrm{d}u_i
\end{aligned}
$$

First:

$$
P(D_i \leq t+w | D_i > t, Z_i^D, u_i, \xi) = \frac{S_i^D(t | Z_i^D, u_i, \xi) - S_i^D(t+w | Z_i^D, u_i, \xi)}{S_i^D(t | Z_i^D, u_i, \xi)}
$$

We have:

$$
\begin{aligned}
&g(u_i | D_i > t, \mathcal{H}_i^{J,2}(t), Z_{ij}^R, Z_i^D, \xi) \\
&= \frac{g(u_i, D_i > t, \mathcal{H}_i^{J,2}(t) | Z_{ij}^R, Z_i^D, \xi)}{P(D_i > t, \mathcal{H}_i^{J,2}(t) | Z_{ij}^R, Z_i^D, \xi)} \\
&= \frac{P(D_i > t, \mathcal{H}_i^{J,2}(t) | Z_{ij}^R, Z_i^D, \xi, u_i) g(u_i | Z_{ij}^R, Z_i^D, \xi)}{\int_0^\infty P(D_i > t, \mathcal{H}_i^{J,2}(t) | Z_{ij}^R, Z_i^D, \xi, u_i) g(u_i) \mathrm{d}u_i} \\
&= \frac{P(D_i > t | Z_i^D, \xi, u_i) P(\mathcal{H}_i^{J,2}(t) | Z_{ij}^R, \xi, u_i) g(u_i)}{\int_0^\infty P(D_i > t | Z_i^D, \xi, u_i) P(\mathcal{H}_i^{J,2}(t) | Z_{ij}^R, \xi, u_i) g(u_i) \mathrm{d}u_i} \\
&= \frac{S_i^D(t | Z_i^D, \xi, u_i) \prod_{k=1}^{J} \frac{\lambda_{ik}^R(X_{ik}/Z_{ij}^R, \xi, u_i) S_{ik}^R(X_{ik}/Z_{ij}^R, \xi, u_i)}{S_{i(k-1)}^R(X_{i(k-1)}/Z_{ij}^R, \xi, u_i)} g(u_i)}{\int_0^\infty S_i^D(t | Z_i^D, \xi, u_i) \prod_{k=1}^{J} \frac{\lambda_{ik}^R(X_{ik}/Z_{ij}^R, \xi, u_i) S_{ik}^R(X_{ik}/Z_{ij}^R, \xi, u_i)}{S_{i(k-1)}^R(X_{i(k-1)}/Z_{ij}^R, \xi, u_i)} g(u_i) \mathrm{d}u_i} \\
&= \frac{S_i^D(t | Z_i^D, \xi, u_i) \prod_{k=1}^{J} \lambda_{ik}^R(X_{ik}/Z_{ij}^R, \xi, u_i) S_{iJ}^R(X_{iJ}/Z_{ij}^R, \xi, u_i) g(u_i)}{\int_0^\infty S_i^D(t | Z_i^D, \xi, u_i) \prod_{k=1}^{J} \lambda_{ik}^R(X_{ik}/Z_{ij}^R, \xi, u_i) S_{iJ}^R(X_{iJ}/Z_{ij}^R, \xi, u_i) g(u_i) \mathrm{d}u_i}
\end{aligned}
$$

$$
\begin{aligned}
&g(u_i|D_i > t, \mathcal{H}_i^{J,2}(t), Z_{ij}^R, Z_i^D, \xi) \\
&= \frac{S_i^D(t|Z_i^D, \xi, u_i) \prod_{k=1}^{J} [\lambda_0^R(X_{ik}) u_i \exp(\beta' Z_{ij}^R)] S_{iJ}^R(X_{iJ}/Z_{ij}^R, \xi, u_i) g(u_i)}{\int_0^\infty S_i^D(t|Z_i^D, \xi, u_i) \prod_{k=1}^{J} [\lambda_0^R(X_{ik}) u_i \exp(\beta' Z_{ij}^R)] S_{iJ}^R(X_{iJ}/Z_{ij}^R, \xi, u_i) g(u_i) \mathrm{d}u_i} \\
&= \frac{S_i^D(t|Z_i^D, \xi, u_i)(u_i)^J S_{iJ}^R(X_{iJ}/Z_{ij}^R, \xi, u_i) g(u_i)}{\int_0^\infty S_i^D(t|Z_i^D, \xi, u_i)(u_i)^J S_{iJ}^R(X_{iJ}/Z_{ij}^R, \xi, u_i) g(u_i) \mathrm{d}u_i}
\end{aligned}
$$

And we deduce:

$$
\begin{aligned}
&P^2(t, t+w; \xi) \\
&= \int_0^\infty \frac{S_i^D(t|Z_i^D, u_i, \xi) - S_i^D(t+w|Z_i^D, u_i, \xi)}{S_i^D(t|Z_i^D, u_i, \xi)} \times \frac{S_i^D(t|Z_i^D, \xi, u_i)(u_i)^J S_{iJ}^R(X_{iJ}/Z_{ij}^R, \xi, u_i) g(u_i)}{\int_0^\infty S_i^D(t|Z_i^D, \xi, u_i)(u_i)^J S_{iJ}^R(X_{iJ}/Z_{ij}^R, \xi, u_i) g(u_i) \mathrm{d}u_i} \mathrm{d}u_i \\
&= \frac{\int_0^\infty [S_i^D(t|Z_i^D, u_i, \xi) - S_i^D(t+w|Z_i^D, u_i, \xi)](u_i)^J S_{iJ}^R(X_{iJ}/Z_{ij}^R, \xi, u_i) g(u_i) \mathrm{d}u_i}{\int_0^\infty S_i^D(t|Z_i^D, \xi, u_i)(u_i)^J S_{iJ}^R(X_{iJ}/Z_{ij}^R, \xi, u_i) g(u_i) \mathrm{d}u_i}
\end{aligned}
$$

Prediction 3

We calculate the probability of death between t and $t+w$ ignoring the previous times of recurrences. We have:

$$
\begin{aligned}
&P^3(t, t+w; \xi) \\
&= P(D_i \leq t+w | D_i > t, Z_i^D, \xi) \\
&= \int_0^\infty P(D_i \leq t+w | D_i > t, Z_i^D, u_i, \xi) \times g(u_i | D_i > t, Z_i^D, \xi) \mathrm{d}u_i
\end{aligned}
$$

Still:

$$
P(D_i \leq t+w | D_i > t, X_i^{(t)}, X_{i(J+1)} > t, Z_{ij}^R, Z_i^D, u_i, \xi) = \frac{S_i^D(t|Z_i^D, u_i, \xi) - S_i^D(t+w|Z_i^D, u_i, \xi)}{S_i^D(t|Z_i^D, u_i, \xi)}
$$

We have:

$$\begin{aligned} g(u_i|D_i > t, Z_i^D, \xi) &= \frac{g(u_i, D_i > t|Z_i^D, \xi)}{P(D_i > t|Z_i^D, \xi)} \\ &= \frac{P(D_i > t|Z_i^D, \xi, u_i)g(u_i|Z_i^D, \xi)}{\int_0^\infty P(D_i > t|Z_i^D, \xi, u_i)g(u_i)\mathrm{d}u_i} \\ &= \frac{S_i^D(t|Z_i^D, \xi, u_i)g(u_i)}{\int_0^\infty S_i^D(t|Z_i^D, \xi, u_i)g(u_i)\mathrm{d}u_i} \end{aligned}$$

We deduce from the two previous results:

$$\begin{aligned} &P^3(t, t+w; \xi) \\ &= \int_0^\infty \frac{S_i^D(t|Z_i^D, u_i, \xi) - S_i^D(t+w|Z_i^D, u_i, \xi)}{S_i^D(t|Z_i^D, u_i, \xi)} \times \frac{S_i^D(t|Z_i^D, \xi, u_i)g(u_i)}{\int_0^\infty S_i^D(t|Z_i^D, \xi, u_i)g(u_i)\mathrm{d}u_i}\mathrm{d}u_i \\ &= \frac{\int_0^\infty [S_i^D(t|Z_i^D, u_i, \xi) - S_i^D(t+w|Z_i^D, u_i, \xi)]g(u_i)\mathrm{d}u_i}{\int_0^\infty S_i^D(t|Z_i^D, \xi, u_i)g(u_i)\mathrm{d}u_i} \end{aligned}$$

Appendix E: Prognostic model developed on French series excluding the peritumoural vascular invasion

Table 3: Joint and Landmark Cox models estimations on the French cohort (n=1067 patients, 427 recurrent events) for the model without peritumoural vascular invasion

	Recurrent events		**Death**		**Death - Cox LM**	
Variable	HR	(95%CI)	HR	(95%CI)	HR	(95% CI)
Age						
$]40-55]$ *vs* >55	1.26	(0.96-1.64)	0.42	(0.18-0.97)	0.56	(0.41-0.76)
≤ 40 *vs* >55	2.30	(1.60-3.29)	1.00	(0.39-2.58)	0.54	(0.31-0.92)
Tumour size (>20 mm vs ≤ 20 mm)	1.81	(1.40-2.33)	4.28	(2.20-8.31)	1.20	(0.87-1.65)
Nodal involvement	1.98	(1.56-2.50)	5.84	(3.14-10.86)	1.97	(1.48-2.60)
Grade						
II vs I	2.16	(1.58-2.97)	7.05	(3.07-16.20)	1.07	(0.75-1.52)
III vs I	3.05	(2.14-4.34)	39.75	(15.57-101.44)	1.26	(0.84-1.88)
Recurrences before $t=5$ years						
One previous recurrence					7.22	(5.29-9.85)
Two previous recurrences					7.12	(3.19-15.91)
$\theta = var(u_i)$	1.06	(se=0.06)				
α	4.61	(se=0.29)				
LCV	1.19				0.93	

HR: Hazard ratio; CI: Confidence interval; LCV: Likelihood cross-validation criterion; LM: landmark
Cox Landmark at time $t=5$ years