



HAL
open science

Modélisation et prédiction conjointe de différents risques de progression de cancer à partir des mesures répétées de biomarqueurs

Loic Ferrer

► **To cite this version:**

Loic Ferrer. Modélisation et prédiction conjointe de différents risques de progression de cancer à partir des mesures répétées de biomarqueurs. Médecine humaine et pathologie. Université de Bordeaux, 2017. Français. NNT : 2017BORD0875 . tel-01690825

HAL Id: tel-01690825

<https://theses.hal.science/tel-01690825>

Submitted on 23 Jan 2018

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

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

Thèse présentée pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE BORDEAUX

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

Par Loïc FERRER

MODÉLISATION CONJOINTE ET PRÉDICTION
DES DIFFÉRENTS RISQUES DE PROGRESSION DE CANCER
À PARTIR DES MESURES RÉPÉTÉES DE BIOMARQUEURS

JOINT MODELLING AND PREDICTION
OF SEVERAL RISKS OF CANCER PROGRESSION
FROM REPEATED MEASUREMENTS OF BIOMARKERS

Sous la direction de Cécile PROUST-LIMA

Soutenue publiquement le 11 décembre 2017

Membres du jury

MATHOULIN-PÉLISSIER Simone	Pr, INSERM, Bordeaux	Présidente
LATOUCHE Aurélien	Pr, Institut Curie, Saint Cloud	Rapporteur
RIZOPOULOS Dimitris	Ass Pr, Erasmus UMC, Rotterdam	Rapporteur
DUPUY Jean-François	Pr, INSA, Rennes	Examineur
JOLY Pierre	MCF, Université de Bordeaux, Bordeaux	Examineur
PROUST-LIMA Cécile	CR, Inserm, Bordeaux	Directrice de thèse

Contents

Remerciements	7
Scientific production	11
Notations and abbreviations	15
Résumé substantiel	17
1 Introduction	27
1.1 Cancer	27
1.1.1 Definition and brief epidemiological context	27
1.1.2 Stages of the disease	27
1.1.3 Disease markers	28
1.1.4 Treatment strategies	29
1.2 Clinical objectives of the thesis	29
1.2.1 Transitions between disease stages	29
1.2.2 Dynamic personalized medicine	31
1.2.3 Various natures of longitudinal markers	32
1.3 Statistical issues	34
1.3.1 Joint modelling introduction	35
1.3.2 Extension to a multi-state process	35
1.3.3 Individual dynamic prediction tools	35
1.3.4 Extension to ordinal longitudinal markers	36
1.4 Studied data	36
1.5 Thesis structure	37
2 State of the art	39
2.1 Mixed-effects models for longitudinal data analysis	39
2.1.1 Linear mixed-effects model	39
2.1.2 Generalized linear mixed-effects model	43

2.1.3	Hypotheses on missing data	46
2.2	Models for event history data analysis	47
2.2.1	Survival data analysis	48
2.2.2	Multi-state formalism	53
2.2.3	Landmark approach	58
2.2.4	Frailty	59
2.3	Standard joint modelling	60
2.3.1	General structure of the joint modelling	61
2.3.2	Standard joint model with shared random effects	62
2.3.3	Joint model with discrete time-to-event data and exact likelihood inference	66
2.3.4	Extensions of joint models	67
3	Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer	69
3.1	Main article	70
3.2	Supplementary material	87
4	Score test for residual dependence between transition times in a joint model for a longitudinal marker and a multi-state process	103
4.1	Introduction	105
4.2	The joint multi-state frailty model	107
4.2.1	Formulation	108
4.2.2	Estimation	109
4.3	Score test for residual dependence between individual transition times . . .	110
4.3.1	Score statistic	110
4.3.2	Variance of the score statistic	111
4.3.3	Test statistic	111
4.4	Simulation study	112
4.4.1	Model for data generation	112
4.4.2	Model estimation	114
4.4.3	Type-I error	114
4.4.4	Power	115
4.5	Application on prostate cancer data	115
4.5.1	Initial model	116
4.5.2	Improved model	117

CONTENTS

4.5.3	Final model	117
4.5.4	Interpretation of the final model	118
4.6	Discussion	121
4.7	Software	122
4.8	References	123
4.9	Supplementary material	125
4.9.1	Formulation of the score statistic for a Gaussian frailty term in a joint model for a longitudinal process and a multi-state process . .	125
4.9.2	Simulation data generation	129
4.9.3	Complementary results for the application	130
5	Individual dynamic predictions using landmarking and joint modelling: validation of estimators and robustness assessment	133
5.1	Introduction	135
5.2	Prediction models	138
5.2.1	Definition of individual dynamic prediction	138
5.2.2	Joint model	139
5.2.3	Landmark cause-specific proportional hazards model	141
5.2.4	Landmark model based on pseudo-observations	143
5.2.5	Implementation	144
5.3	Motivating data	145
5.4	Simulation studies	145
5.4.1	Simulation study I: Validation of the estimators $\hat{\pi}_*^k(s, w; \hat{\theta})$	146
5.4.2	Simulation study II: Robustness to models hypotheses	147
5.5	Discussion	155
5.6	References	159
5.7	Supplementary material	163
5.7.1	Complete formulas of the estimators	163
5.7.2	Complementary results for Simulation II: robustness assessment . .	164
5.7.3	Simulation data generation	168
5.7.4	Computational time of the procedures	175
5.7.5	Example of R code	177
5.7.6	References	190

6	Reflection on the joint modelling of ordinal longitudinal data and time-to-event data	191
6.1	Introduction	192
6.2	Methods	193
6.2.1	Notations	193
6.2.2	Joint model	194
6.2.3	Estimation	196
6.2.4	Implementation	199
6.3	Discussion	199
7	Discussion	201
7.1	Joint multi-state modelling	201
7.2	Joint modelling of ordinal longitudinal data and survival data	204
7.3	Individual dynamic predictions	205
7.4	General conclusion	209
	Bibliography	210

Remerciements

La première personne que je tiens à remercier pour ce travail de thèse est bien évidemment ma directrice de thèse, Cécile Proust Lima. Tes connaissances, ton enthousiasme, ta pédagogie et ta bienveillance m'auront porté tout au long de ce travail et sont un exemple pour moi. J'espère sincèrement que l'on pourra continuer à travailler ensemble dans le futur.

Je suis très honoré de soumettre ce travail devant l'ensemble de mon jury de thèse. Simone Mathoulin-Pélissier, je vous remercie d'avoir accepté de présider cette soutenance de thèse. Aurélien Latouche and Dimitris Rizopoulos, I warmly thank you to accept to take part of my jury. Dimitris, this thesis work would not be the same without your impressive contributions on joint modelling. Aurélien, j'espère grandement que l'on pourra commencer à travailler ensemble bientôt. Enfin, Jean-François Dupuy et Pierre Joly, merci d'avoir accepté d'examiner ce travail de thèse.

Cette thèse a été conduite à travers plusieurs collaborations. Je tiens notamment à remercier Paul Sargos et Pierre Richaud, vos disponibilités et vos expertises sur les interprétations cliniques en cancer de la prostate m'auront été d'une grande aide en début de thèse. I warmly thank Tom Pickles and James Dignam for sending me the databases from which the thesis applications have been led. Hein Putter, thank you again for welcoming me in Leiden, I would also like to thank all the researchers, PhD students and postdocs of your research department for their welcome. Antoine Régnauld, je te remercie de m'avoir accueilli en tant que doctorant conseil dans ton équipe à MAPI Lyon. Cette expérience m'a été très enrichissante.

Je souhaite remercier l'ensemble des équipes Biostatistique et SISTM de l'ISPED. En particulier, un grand merci à Virginie pour m'avoir fait découvrir l'ISPED et pour ton encadrement lors de mon stage de Master 2. Tu m'as initié au monde de la recherche et m'as donné envie de réaliser cette thèse. Hélène, je tiens également à te remercier pour tes nombreux conseils et ta disponibilité. Pierre, je suis vraiment heureux que tu aies accepté

REMERCIEMENTS

de faire partie de mon jury de thèse. Enfin Daniel, merci pour ta disponibilité, notamment lorsque certaines réflexions complexes sont à mener. Merci à tous pour votre accueil et votre soutien.

Cher cluster Avakas, merci. Sans toi, cette thèse n'aurait pas pu être menée aussi loin. Je remercie également le Mésocentre de Calcul Intensif Aquitain (MCIA) pour m'avoir donné accès à toi.

Aurélie, c'était un vrai plaisir de faire partie de ton encadrement de stage de Master 2. Je tiens à remercier les enseignants de la Faculté des Sciences de Montpellier, en particulier Lionel Cucala, qui m'a fait découvrir les biostatistiques en fin d'études.

Merci aux personnes qui m'auront supporté tout au long de ces années passées à l'IS-PED. Vous étiez des collègues mais vous êtes devenus des amis : merci à Henri to Pau, mes co-thésardes Anaïs et Mathilde, Perrine et Chloé (désolé de ne pas vous avoir invitées aux soirées entre nouilles), Bobo la stat', Emilie la nordiste, Aïssatou (+ Mehdou), Alex et Robinou. Egalement un grand merci aux autres doctorants : Agnieszka, Bachirou, Rémi, Corentin, Camille et Maud, ainsi qu'à tous ceux que j'oublie par ailleurs. Enfin, merci aux "anciens" ispediens pour votre accueil qui m'aura donné envie de rester un petit peu plus que prévu à Bordeaux, notamment merci à Audrey, Paul, Jeremy (x2), Célia, Mbéry, Mélanie et Viviane.

Parce que je suis et resterai toujours sétois, je termine mes remerciements en pensant à ma ville et à mes amis sétois. Un énorme merci à toute ma famille : mon arrière grand papi, mes grands parents, mon cousin, mes taties et tontons, ma grande cousine, mon parrain, ma parraine et ma filleule. Joanna, merci de préférer les chemins tortueux aux lignes droites, merci pour ta force et ton soutien chaque jour. Enfin, je conclus mes remerciements en pensant à mon papa, ma maman, ma sœur et mon fils.

REMERCIEMENTS

Je dédie ce manuscript à mon fils, Matteo.

REMERCIEMENTS

Scientific production

Articles

Thesis publications

► [Ferrer L.](#), Rondeau V., Dignam J., Pickles T., Jacqmin-Gadda H. and Proust-Lima C. Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer. *Statistics in Medicine*, 35(22):3933–3948, 2016.

DOI: 10.1002/sim.6972

► [Ferrer L.](#), Putter H. and Proust-Lima C. Individual dynamic predictions using landmarking and joint modelling: validation of estimators and robustness assessment. *Submitted for publication*.

► [Ferrer L.](#), Rouanet A., Jacqmin-Gadda H. and Proust-Lima C. Score test for residual dependence between transition times in a joint model for a longitudinal marker and a multi-state process. *Submitted for publication*.

Related articles

► Król A., [Ferrer L.](#), Pignon J.-P., Proust-Lima C., Ducreux M., Bouché O., Michiels S., and Rondeau V. Joint model for left-censored longitudinal data, recurrent events and terminal event: Predictive abilities of tumor burden for cancer evolution with application to the FFCD 2000–05 trial. *Biometrics*, 72(3):907–916, 2016.

► Jacqmin-Gadda H., Proust-Lima C. and [Ferrer L.](#) Joint models for longitudinal and time to event data. Chapter in the book entitled *Handbook of Statistical Methods for Randomized Controlled Trials*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. *To appear*.

► Kaboré R., [Ferrer L.](#), Couchoud C., Harambat J. and Leffondré K. Dynamic prediction models for graft failure in pediatric kidney transplantation. *Submitted for publication*.

Communications

Oral communications at conferences

- ▶ Ferrer L., Rondeau V., Dignam J., Pickles T., Jacqmin-Gadda H. and Proust-Lima C. Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer. *International Biometric Society Channel Network Conference*, Nijmegen, the Netherlands, 2015.
- ▶ Ferrer L., Jacqmin-Gadda H. and Proust-Lima C. Modèles conjoints pour l'évaluation de marqueurs de substitution en essais cliniques : Illustration dans le cancer de la prostate. *9^{ème} Conférence Francophone d'Epidémiologie Clinique – 22^{èmes} journées des statisticiens de Centre de Lutte Contre le Cancer*, Montpellier, France, 2015.
- ▶ Ferrer L., Rouanet A., Jacqmin-Gadda H. and Proust-Lima C. Score test for residual correlation between multiple events in a joint model for longitudinal and multi-state data. *International Society for Clinical Biostatistics Conference*, Utrecht, the Netherlands, 2015.
- ▶ Ferrer L., Rondeau V., Dignam J., Pickles T., Jacqmin-Gadda H. and Proust-Lima C. Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer. *Sixièmes Rencontres des Jeunes Statisticiens*, Le Teich, France, 2015.
- ▶ Ferrer L., Putter H. and Proust-Lima C. Comparison of approaches for dynamic predictions in competing risks. *48^{èmes} Journées de Statistique de la Société Française de Statistique*, Montpellier, France, 2016.
- ▶ Ferrer L., Putter H. and Proust-Lima C. Comparison of approaches for dynamic predictions in presence of competing risks. *Population-based Time-to-event Analyses International Conference*, London, United Kingdom, 2016.
- ▶ Ferrer L., Król A., Rondeau V. and Proust-Lima C. Joint models for longitudinal measurements and multiple failure types after a first cancer. *Journées Scientifiques de l'IReSP*, Paris, France, 2016.
- ▶ Ferrer L., Putter H. and Proust-Lima C. Individual dynamic predictions using landmarking and joint modelling: validation of estimators and robustness assessment. *Journées GDR/SFB*, Bordeaux, France, 2017.

Invited talk at conference

► L. Ferrer and C. Proust-Lima, Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer. *Conference on Lifetime Data Science*, Storrs, United States of America, 2017.

Invited talks at seminars

► Ferrer L. and Proust-Lima C. Joint modelling of a longitudinal and a multi-state processes. *Survival lunch – Leiden University Medical Center*, Leiden, the Netherlands, 2015.

► Ferrer L., Putter H. and Proust-Lima C. Individual dynamic predictions using landmarking and joint modelling: validation of estimators and robustness assessment. *BSU Seminar – MRC Biostatistics Unit*, Cambridge, United Kingdom, 2017.

Written communication (poster) at conference

► L. Ferrer, H. Putter and C. Proust-Lima, Individual dynamic predictions from joint models and landmark models: robustness to misspecifications. *International Biometric Society Channel Network Conference*, Hasselt, Belgium, 2017.

Research visits

For the purpose of this thesis, I spent 3 periods of two weeks in the Survival Analysis team of the Leiden University Medical Center (LUMC) in Leiden (the Netherlands), directed by Hein Putter. This collaboration notably enabled the development of landmark techniques in the prediction setting.

Obtained grants

This thesis was supported by a joint grant from INSERM and Région Aquitaine. I also obtained two mobility grants in order to visit LUMC in Leiden: a grant from the Institut OpenHealth and a grant from the Réseau Franco-Néerlandais.

Notations and abbreviations

Notations

- T : Time-to-event variable
- Y : Longitudinal marker
- b : Random effects
- θ : Vector of parameters
- $\hat{\theta}$: Estimate of θ
- $\lambda(\cdot)$: Instantaneous hazard function / Transition intensity function
- $f_X(\cdot)$: Probability density function of a random variable X
- $L(\cdot)$: Likelihood function
- $l(\cdot)$: Log-likelihood function
- X^\top : X transpose
- $\mathbb{E}(X)$: Expectation of random variable X
- $\mathbb{1}\{\cdot\}$: Indicator function
- $\Pr(A)$: Probability of A
- $\Pr(A|B)$: Probability of A conditional on B
- $\mathcal{N}(\mu, \Sigma)$: Normal distribution (possibly multivariate) with mean μ and variance Σ
- $\Phi^{(n)}(a, b; \mu, \Sigma)$: Cumulative distribution function of a multivariate normal distribution of dimension n , truncated at a , computed at b ($a \leq b$), with mean μ and variance Σ

Abbreviations

BLUP: Best Linear Unbiased Estimator

NOTATIONS AND ABBREVIATIONS

CS PH: Cause-Specific Proportional Hazards

EM: Estimation-Maximisation algorithm

HRQoL: Health-Related Quality of Life

JM: Joint Modelling / Joint Model

LM: Landmarking / Landmark Model

MLE: Maximum Likelihood Estimation

MSEP: Mean Squared Error of Prediction

PV: Pseudo-Value

PSA: Prostate-Specific Antigen

RT: Radiation Therapy

Résumé substantiel

1 Introduction

La modélisation conjointe de données longitudinales et de données de survie est de plus en plus utilisée dans les essais cliniques en cancer. Elle permet de modéliser simultanément un processus longitudinal et un processus de survie tout en prenant en compte le lien entre ces deux processus.

En cancer de la prostate, la modélisation conjointe est très utile. L'antigène spécifique de la prostate (PSA), qui est une protéine sécrétée par la prostate, se trouve être surexprimé en présence du cancer. Ce marqueur tumoral longitudinal est couramment utilisé par les cliniciens afin de surveiller les patients atteints de cancer de la prostate localisé après un traitement (radiothérapie ou chirurgie) et de détecter la présence du cancer. Divers auteurs ont montré, à travers différents types de modèles mixtes, que la dynamique de ce marqueur biologique ainsi que le niveau pré-traitement de PSA et d'autres facteurs mesurant l'agressivité des cellules cancéreuses et l'étendue de la tumeur, étaient des facteurs de risque de progression et permettaient de prédire le risque de rechute clinique de façon dynamique (par exemple en utilisant les PSA collectés successivement pour adapter la prédiction au cours du temps) . En pratique, un patient peut avoir une succession d'événements de progressions cliniques avec par exemple une récurrence locale, suivie par une récurrence métastatique à distance puis le décès. Ainsi, au lieu d'un événement clinique unique, la progression du cancer de la prostate devrait être définie comme un processus multi-états avec un accent sur les transitions entre les états cliniques et l'impact de la dynamique du biomarqueur sur celles-ci. Ceci est essentiel pour comprendre et prédire avec précision l'évolution de la maladie, et c'est d'une importance particulière pour les cliniciens qui ont besoin de distinguer les différents types d'événements afin d'adapter correctement le traitement.

Un autre aspect clé dans la modélisation conjointe est la nature du marqueur longitudinal. Dans la littérature, ont principalement été proposés des développements pour un marqueur longitudinal Gaussien, ce qui est satisfaisant pour nombre de biomarqueurs issus de dosages sanguins tels que le PSA. Cependant, on recueille aussi fréquemment

des marqueurs non Gaussiens, et il n'est pas toujours possible de les transformer afin de satisfaire les hypothèses de normalité. C'est notamment le cas des échelles de qualité de vie contenant peu de modalités. Bien que ces marqueurs soient de plus en plus présents dans la recherche en cancer, leur analyse conjointe est limitée du fait en particulier de l'absence de logiciels s'adaptant à la nature ordinaire du marqueur.

2 Modélisation conjointe d'un processus longitudinal et d'un processus multi-états

Dans ce contexte, nous avons premièrement proposé un modèle conjoint avec effets aléatoires partagés pour un processus longitudinal Gaussien et un processus multi-états corrélés qui est divisé en deux sous-modèles : un sous-modèle linéaire mixte pour les données longitudinales, et un sous-modèle multi-états avec intensités proportionnelles pour chaque covariable pour le(s) temps de transition individuel(s), tous deux liés par des effets aléatoires partagés. Les paramètres de ce modèle sont estimés dans le cadre du maximum de vraisemblance via un algorithme EM couplé à un algorithme quasi-Newton en cas de convergence lente. Le modèle multi-états conjoint a été implémenté sous R, via la combinaison et l'extension de deux packages de référence (`mstate` pour modèles multi-états et `JM` pour modèles conjoints). L'implémentation est donc facile et efficace. Le programme d'estimation a été validé par une étude de simulation, soulignant les très bonnes performances du programme. Néanmoins, l'inférence statistique implique un calcul d'intégrale sur les effets aléatoires individuels qui induit des complexités numériques. L'étude de simulation a ainsi révélé la nécessité d'utiliser des quadratures Gaussiennes précises, en particulier lorsque la dimension des effets aléatoires augmente, afin d'obtenir des estimations précises. Nous avons ensuite appliqué le modèle multi-états conjoint à deux cohortes d'hommes traités par radiothérapie pour un cancer de la prostate localisé aux Etats-Unis et au Canada. L'application a confirmé que la dynamique des PSA via son niveau et sa pente courantes impactait fortement le risque instantané d'avoir une rechute clinique après la fin de la radiothérapie, mais n'était pas nécessairement associée à une évolution clinique de la maladie après un stade avancé. Par exemple, après une rechute à métastases, la dynamique de PSA n'impactait plus le risque de décéder. Ceci illustre que dans ces cancers avancés, d'autres marqueurs biologiques sont sans doute à privilégier. Ce travail a été publié dans *Statistics in Medicine* en 2016.

3 Score test pour une corrélation résiduelle entre temps de transition individuels

Lorsque l'on s'intéresse à un ensemble de progressions cliniques en santé, il peut arriver que certains individus soient plus à risque que d'autres d'avoir une succession d'événements, sans que ceci soit expliqué par les variables explicatives et marqueurs d'intérêt. En cancer de la prostate par exemple, pour deux patients avec les mêmes caractéristiques mesurées, l'un d'entre eux peut expérimenter plusieurs rechutes de la maladie, et l'autre aucune. Ce phénomène peut se traduire par la nécessité d'inclure un effet aléatoire individuel dans le modèle multi-états - on parle alors de terme de fragilité - qui capte cette hétérogénéité non expliquée par les variables explicatives incluses dans le modèle. Nous avons donc proposé un score test qui permet de tester l'inclusion d'un tel terme de fragilité Gaussien. Le score test a l'avantage de ne nécessiter que l'estimation du modèle multi-états conjoint "standard", c'est à dire qui n'inclut pas le terme de fragilité.

Une expression analytique de la statistique de score a été trouvée. Pour la variance, aucune expression n'a pu être obtenue et nous avons donc utilisé une approximation de la variance asymptotique corrigée par rapport aux paramètres de nuisance du modèle obtenue à l'aide d'une méthode de différence finie.

Une étude de simulation a été menée pour s'assurer des bonnes performances du score test. Les simulations réalisées ont révélé une estimation correcte de l'erreur de type-I et une bonne puissance sous couvert de collecter une information suffisante qui garantit d'atteindre la distribution asymptotique de la statistique de score. En particulier, le test ne peut pas être appliqué en présence d'un faible nombre d'individus ou de transitions par individu.

Le score test pour un terme de fragilité dans un modèle multi-états conjoint a été implémenté sous R, avec une fonction simple d'utilisation proposée aux utilisateurs qui utilise les fonctionnalités de JM. Une application a ensuite été menée à partir des données déjà étudiées avec des patients traités par radiothérapie pour un cancer prostate localisé. Celle-ci a confirmé l'utilité du score test afin de vérifier l'hypothèse Markovienne du processus multi-états, qui suppose notamment une indépendance entre les temps de transition individuels.

Le score test que nous avons proposé peut également être utilisé comme outil de qualité d'ajustement de modèle. Dans l'application il a notamment révélé la présence de sujets avec des caractéristiques extrêmes par rapport à la population générale. Il nous a ainsi permis d'obtenir un modèle plus performant et s'ajustant mieux aux données étudiées.

Le modèle final a confirmé que la dynamique des PSA avait un fort impact sur le risque de rechute clinique post-radiothérapie, mais n'était pas nécessairement associée à une évolution de la maladie après une première rechute. Ce travail a donné lieu à un article actuellement soumis pour publication dans un journal de biostatistique.

4 Pouvoir prédictif des modèles conjoints et modèles landmarks

Un autre pan de cette thèse concerne l'aspect prédictif de modèles prenant en compte des caractéristiques individuelles dynamiques. En effet, prédire précisément pour les sujets traités leur probabilité individuelle de progression telle qu'une rechute clinique dans les x prochaines années à partir des informations individuelles collectées jusqu'au temps de la prédiction, est devenu une question centrale dans le monitoring des patients traités pour un cancer. Des stratégies personnalisées de traitement peuvent ensuite être proposées selon les probabilités individuelles actualisées de chaque type de progression, ou le temps de la prochaine visite de suivi peut être optimisé.

Ces prédictions dynamiques individuelles peuvent être calculées à partir de deux principales approches de modélisation statistique : la modélisation conjointe et la modélisation par landmark. La modélisation conjointe analyse simultanément les mesures répétées du biomarqueur et les données de temps d'événement pour un ensemble de patients en les reliant par une fonction des effets aléatoires partagés. Cette approche a l'avantage de modéliser la progression de la maladie dans son ensemble, ce qui la rend très populaire chez les statisticiens. Mais elle repose souvent sur des hypothèses statistiques simplificatrices (e.g., proportionnalité des risques) et reste très complexe à estimer, si bien qu'elle est difficile à mettre en oeuvre pour réaliser de la prédiction dynamique. La modélisation par landmark consiste à ajuster des modèles de survie standards tenant compte de la dynamique du biomarqueur, en ne considérant que le sous-échantillon des patients à risque d'événement au temps de prédiction, et leur histoire des covariables jusqu'au temps de prédiction. Ces modèles induisent nettement moins de problèmes numériques et réduisent le possible biais d'estimation lié à l'hypothèse de proportionnalité des risques d'événement. Cependant, comme ils n'explorent pas complètement la corrélation entre le marqueur et le temps d'événement, ils peuvent donner des estimateurs peu précis. Ces deux approches diffèrent donc dans l'information utilisée, les hypothèses du modèle et la complexité des procédures computationnelles.

Motivés par la prédiction de progressions concurrentes du cancer de la prostate (re-

RÉSUMÉ SUBSTANTIEL

chute clinique ; décès non dû au cancer) à partir de l'histoire des PSA, nous avons comparé ces approches statistiques pour évaluer lesquelles sont les plus pertinentes en pratique pour développer et fournir des prédictions dynamiques de progression de cancer. Le coeur de ce travail a consisté en une étude de simulation approfondie car cela permet de confronter les résultats des différentes approches à la vérité simulée. Divers scénarios ont été considérés pour explorer les performances des modèles, notamment lorsque l'hypothèse de proportionnalité des risques était violée, le sous-modèle longitudinal était mal spécifié ou la structure d'association entre le processus longitudinal et le processus de survie était mal spécifiée. Les prédictions dynamiques individuelles dérivées de ces modèles ont été spécifiquement comparées en termes d'erreur de prédiction et de capacité discriminante.

En parallèle de cette étude comparative, nous avons aussi proposé une méthode permettant d'évaluer l'incertitude autour des prédictions dynamiques individuelles d'événement, alors que ce n'était pas encore possible avec l'approche landmark et que des techniques concurrentes avaient été proposées avec la modélisation conjointe. Les définitions des estimateurs proposés et du calcul de leur variabilité ont été formellement validés via l'étude de simulation. Les différents estimateurs ont globalement montré de bonnes performances en termes de biais et de taux de couverture mais il a été souligné que seul l'estimateur du modèle conjoint obtenu par intégration numérique autour des effets aléatoires était sans biais en cas d'information répétée faible. Les estimateurs des modèles landmark et l'estimateur du modèle conjoint basé sur une approximation de Laplace de l'intégrale nécessitent une information longitudinale individuelle collectée jusqu'au temps de prédiction suffisante pour être précis. Comme attendu, les estimateurs issus des modèles conjoints étaient bien plus efficaces que ceux issus des modèles landmarks, en particulier lorsque peu de sujets expérimentaient l'événement d'intérêt dans la fenêtre de prédiction.

La comparaison des modèles de prédiction sous divers scénarios a ensuite confirmé que le modèle conjoint était plus performant que les modèles landmarks en cas de bonne spécification. Lorsque la fonction d'association entre les deux processus était mal spécifiée, le modèle conjoint restait le plus performant mais les précisions des prédictions issues des deux approches étaient bien moindres. Cela souligne l'importance de bien définir la nature de la dépendance entre les processus lorsque l'on souhaite réaliser des prédictions individuelles. Dans le cas d'une forte violation de l'hypothèse de proportionnalité des risques, les estimateurs issus des modèles conjoints et landmarks avaient des performances comparables, supposant que la violation de cette hypothèse devait être extrême pour impacter significativement les prédictions issues du modèle conjoint. Enfin, les résultats ont révélé une sensibilité bien plus importante du modèle conjoint que les modèles landmarks à une

mauvaise spécification de la trajectoire du marqueur longitudinal. Ce résultat majeur met en évidence qu'il est impératif de spécifier avec soin la trajectoire temporelle du marqueur longitudinal lorsque l'on souhaite définir des prédictions dynamiques, sous peine de fournir des prédictions individuelles très imprécises.

Ce travail, mené en collaboration avec Hein Putter (LUMC, Leiden) lors de 3 mobilités a donné lieu à un manuscript qui est soumis dans un journal international de biostatistique. Les différentes techniques ont été implémentées sous R en modifiant et étendant des packages déjà existants, afin d'assurer une certaine praticité.

5 Modélisation conjointe de données longitudinales ordinales et de données de survie via une vraisemblance exacte

Un dernier travail de thèse a été initié sur la modélisation conjointe de données longitudinales ordinales et de données de survie via une vraisemblance exacte. Ce travail a été motivé par consultance faite en 2015 pour MAPI, une CRO Lyonnaise sur l'analyse de données de qualité de vie en essais cliniques sur le cancer.

La modélisation conjointe a principalement été proposée pour des données longitudinales Gaussiennes et des données de survie, et des programmes conviviaux sont aujourd'hui disponibles pour réaliser ces analyses. Dès lors que le marqueur est ordinal, comme lorsqu'on s'intéresse à l'évolution de la qualité de vie tronquée par la rechute de la maladie, bien que des modèles conjoints aient été proposés, aucun logiciel n'est disponible.

Pour pallier à ce manque, nous avons proposé une approche originale basée sur une vraisemblance exacte. Une étape cruciale dans l'estimation des modèles conjoints à effets aléatoires partagés est en effet la résolution d'une intégrale sur les effets aléatoires dans la vraisemblance. Celle-ci complique grandement les calculs et empêche l'inclusion de processus de lien plus complexes, tels que des processus auto-régressifs ou des mouvements browniens. Cette intégrale peut être évitée par l'utilisation de distributions Gaussiennes tronquées et de modèles de survie moins standards, des modèles probit séquentiels. Nous avons appliqué cette technique pour développer un modèle conjoint permettant de modéliser simultanément des données longitudinales ordinales et des données de survie via une vraisemblance exacte.

Le modèle et sa vraisemblance ont été écrits, et le programme d'estimation a été implémenté sous R de la manière la plus accessible et modulable possible, avec la pers-

pective d'en faire un package disponible sur le CRAN. Ce travail, qui n'est pas finalisé, a principalement été réalisé au cours d'un stage de Master II que j'ai coencadré avec Cécile Proust-Lima. L'étude de simulation visant à s'assurer des bonnes performances d'estimation du modèle est encore en cours de réalisation.

6 Discussion générale

Cette thèse a proposé plusieurs développements statistiques pour analyser deux types de données corrélées fréquemment observées dans la pratique clinique : des données répétées d'un marqueur longitudinal et des données d'histoire d'événements.

J'ai d'abord étendu le principe de la modélisation conjointe à l'analyse d'un marqueur longitudinal Gaussien et de données de temps de transition entre états caractérisées par un processus multi-états. Cela a été réalisé en étendant et combinant la modélisation multi-états, basée sur la théorie des processus de comptage, et la modélisation conjointe standard, qui s'applique pour un unique type d'événement. La modélisation multi-états conjointe non seulement distingue le type d'événement survenu chez l'individu, mais elle permet également de dépeindre les caractéristiques individuelles après la survenue de ce premier événement en s'intéressant à des intensités de transition entre états distincts.

Une limite de cette approche est qu'elle repose sur l'hypothèse de Markov. Or il peut subsister en pratique une corrélation entre temps de transition individuels non expliquée par les covariables du modèle. Pour capturer cette corrélation intra-individuelle résiduelle, nous avons proposé un test du score qui teste l'inclusion d'un terme de fragilité Gaussien dans un modèle multi-états conjoint. Ce test, qui a l'avantage de n'estimer le modèle que sous l'hypothèse nulle, a montré de bonnes performances dès lors que le nombre de sujets (et de transitions) est suffisant. Il fournit en pratique un outil de diagnostic pertinent dans les modèles conjoints multi-états et plus généralement dans les modèles multi-états, car au delà de l'hypothèse de fragilité, il teste plus généralement l'ajustement aux données.

Ces travaux, appliqués aux données de deux cohortes d'hommes traités par radiothérapie pour un cancer de la prostate localisé ont permis de quantifier l'effet de chaque facteur pronostique classique et de la dynamique des PSA sur chaque intensité de transition entre états cliniques. L'application a révélé que la dynamique des PSA était fortement associée au risque de rechute clinique suite à la radiothérapie, mais n'était plus forcément liée à l'évolution de la maladie dans des stades plus avancés. Par ailleurs, les autres facteurs pronostiques classiques, mesurant à la fin de la radiothérapie l'étendue de la tumeur, l'agressivité des cellules cancéreuses, et le niveau initial des PSA, n'étaient pas toujours

liés au risque de transition entre deux états cliniques, après ajustement sur la dynamique courante des PSA.

Dans une deuxième partie, je me suis centré sur le problème de prédiction individuelle dynamique d'incidences cumulées d'événements compétitifs à partir d'une information incluant des données répétées de biomarqueur. En effet, malgré de nombreuses contributions statistiques sur le sujet dans les dernières années, aucune n'avait formellement défini la quantité d'intérêt, validé ses estimateurs et comparé plusieurs approches de modélisation en simulations. J'ai comblé ce vide en approfondissant deux approches : la modélisation conjointe, qui définit correctement la corrélation entre les données répétées du marqueur longitudinal et la survenue d'événement, et la modélisation landmark, qui approche cette corrélation mais limite les biais liés aux hypothèses des modèles et propose des solutions inférentielles plus simples. Ce travail a notamment confirmé que les estimateurs issus de la modélisation conjointe étaient plus efficaces que ceux issus de l'approche landmark même si ces derniers fournissaient des estimations non biaisées. Il a aussi démontré la supériorité des capacités prédictives de l'approche par modélisation conjointe, sous couvert de spécifier avec grand soin la trajectoire temporelle du marqueur longitudinal et la structure de dépendance. Or ce sont deux aspects de la modélisation conjointe qui restent très peu discutés en pratique, les contributions se limitant le plus souvent à une évolution linéaire et une dépendance sur le niveau courant. Les recommandations issues de ce travail sont donc essentielles pour assurer des outils de prédiction dynamique les plus précis possibles.

Enfin, un dernier travail a concerné la modélisation conjointe de données répétées ordinales et de données de survie. En effet, peu de développements se sont focalisés sur ce type de données alors qu'il est de plus en plus présent en recherche en santé. Un modèle innovant a été proposé, qui permet en plus d'éviter les coûts computationnels fastidieux de la modélisation conjointe classique et assure l'inclusion de structures d'association plus sophistiquées entre les deux processus corrélés. Dans le cadre de la thèse, seuls le modèle et sa vraisemblance ont été écrits. Une étude de simulation a été initiée afin de valider les performances d'estimation du modèle conjoint développé mais elle n'a pas pu encore être finalisée. Une fois validée, la méthodologie trouvera des applications dans l'analyse de données de qualité de vie en cancer mais aussi bien au delà avec l'essor des données subjectives notamment de qualité de vie ou de psychométrie dans la recherche en épidémiologie.

RÉSUMÉ SUBSTANTIEL

En conclusion, ce travail fournit des développements statistiques pour analyser données longitudinales et données d'histoire d'événements lorsque toutes deux sont corrélées. Ces travaux ont été implémentés dans des logiciels classiques, et des recommandations ont été associées pour leur utilisation par la communauté large : clinique et statistique. Enfin, des applications ont été menées en cancérologie pour illustrer leur utilisation en pratique.

1 Introduction

1.1 Cancer

1.1.1 Definition and brief epidemiological context

Cancer is a term which indicates a collection of diseases characterised by an abnormal regulation of the cell cycle leading to a continual unregulated proliferation of cancer cells, with the possibility that these cells spread into any part of the human body [Cooper, 2000]. A cancer may start anywhere, invading normal tissues and forming tumors, which may threaten the survival of the infected tissues. When the tumor is benign, it does not spread to the other parts of the body and thus is not considered as cancerous. But when the tumor is malignant (and thus cancerous), some cells may invade surrounding tissues or even spread to distant tissues, the later process being referred to as metastasising. Finally, when vital organs are affected, the survival of the patient is threatened.

Cancer is the second leading cause of morbidity and mortality worldwide, with nearly 1 in 6 deaths in 2015, that is 8.8 million deaths [Ferlay et al., 2015]. In 2012 alone, 14.1 million new cases occurred, and the incidence rates projected an increase about 70% over the next two decades, that is 23.6 million new cases each year by 2030. A cancer may be of different type (defined from its location), with the lung, female breast, colorectal and prostate cancers in the ordered top 4 of the more common worldwide. Depending on the type of cancer, the vital status of the patient, and more generally his Health-Related Quality of Life (HRQoL), may be more or less threatened or impacted. Furthermore, the same disease may progress with several clinically identified evolutions, called stages, each stage inducing a very different prognosis for the patient.

1.1.2 Stages of the disease

The most basic way to distinguish the stages of the disease and define how it spreads over time is to distinguish them in four categories [Young Jr et al., 2001]:

- stage 0: in situ; the malignant cells are still in the place where they started and

- have not spread to nearby tissues;
- stage I: localised cancer; the tumor has not grown deeply into nearby tissues and remains limited to the organ of origin. It is often called early-stage cancer;
- stage II–III: regional spread; the tumor extends beyond the limits of the organ of origin;
- stage IV: distant spread; the cancer has spread to other areas of the body, and the tumor cells begun to grow at their new location(s). One also talks about remote, metastatic or advanced cancer.

Depending on the stage of the disease, the patient’s prognosis may widely vary. For example, in subjects with a prostate cancer diagnosed between 2005 and 2009 in England, the 5-year relative survival after diagnosis was about 96% for those diagnosed with stages I, II, III (loco-regional cancer) whereas it dropped to 69% for those with stage IV (distant cancer) at diagnosis [National Collaborating Centre for Cancer (UK), 2014]. Thus, according to the nature of the clinical state that the patients experience, their management may vary widely with adapted treatment strategies.

1.1.3 Disease markers

Whether to identify the presence of the disease, assess its stage, understand the disease evolution or predict it, a common solution is to use disease markers. The latter are substances or measurable parameters that allow to characterize the disease in a quantified way.

Several kinds of disease markers are used in the clinical practice. The biological markers (or biomarkers) are the most common because they consist in substances directly associated with the clinical state (or stage) of the disease. For example, the tumor markers, which are found in blood, urine, or body tissues, are widely used. The quantified expression of an antigen specific to a host organism, which may be over-expressed in presence of cancer, is a potential interesting tumor marker. One of the most famous of them is the prostate-specific antigen (PSA), which is secreted directly by the prostate and is found as over-expressed in the presence of the prostate cancer [Barry, 2001].

Other disease markers may be defined, as the tumor size, which directly measures the surface of the tumor or in a more accurate way its volume [Eisenhauer et al., 2009]. Other useful indicators are Health-Related Quality of Life (HRQoL) markers, which characterize latent domains associated to the disease such as the subject’s perception of his physical and mental health [Aaronson et al., 1993]. In current research, researchers try to identify new markers for specific or unspecific cancer which would be powerfully prognostic of the

INTRODUCTION

disease evolution. Recent researches notably targeted promising biomarkers from liquid biopsy of cancer, such as for example the detection and molecular characterisation of circulating tumor cells (CTC), potential powerful prognostic factor in metastatic breast, colon or prostate cancer [Allard et al., 2004].

1.1.4 Treatment strategies

To limit or annihilate the impact of the cancer on the patient's survival or his quality of life, several management strategies may be organised, according to the kind of disease and its stage.

First, the frequency of the monitoring is adapted. Clearly, after a recurrence, a patient with a very poor prognosis (high risk of distant recurrence for example) will be more actively monitored than if his prognosis was very good.

Then, adequate and effective treatments may be proposed according to the type of cancer and its stage, using specific treatment regimen. When the cancer is solid and isolated, a primary method of treatment is surgery. It may be used for example in localised prostate cancer, with the ablation of the prostate. To complete the treatment, or when the surgery is impossible because of the cancer location, as well as when the disease is in a more advanced stage, chemotherapy and/or radiotherapy can be proposed. Combinations of multiple treatment regimen are also possible, according to the disease's ability to resist to treatment, and/or for limiting the toxicity that it diffuses in the body. Nowadays, palliative care and psychosocial support are also proposed to patients, to improve their quality of life and even prolong their life [Temel et al., 2010].

1.2 Clinical objectives of the thesis

1.2.1 Transitions between disease stages

We earlier discussed that there are various possible managements in cancer, according to the kind of disease, its specific stage and its evolution process.

To give a basic example, let us focus on prostate cancer, major cancer with 1.1 million men worldwide diagnosed in 2012. A recent study of Hamdy et al. [2016] argued that in the treated subjects for a localised prostate cancer, the 10-year prostate-cancer-specific survival probability was very good (about 99%) whatever the used treatment strategy among active monitoring, surgery or radiotherapy. However, by investigating more in detail, authors showed that 12.4% of the subjects had a disease progression over the period,

including metastases. Actually, the incidence of disease progression was significantly higher in the active-monitoring group than in the radiotherapy and surgery groups, and the same conclusion was made when the focus was on the presence of metastases. Thus, authors finally concluded that a longer-term study would be crucial to evaluate the full evolution process of the disease.

Actually, these results confirm that in subjects treated for a localised prostate cancer, there is a distinction between the risks of loco-regional clinical relapse, remote clinical relapse or death due to cancer. In this example, grouping local and regional disease makes sense, leaving distant stage as a separate event. Indeed, the latter is the worse clinical consequence before death, and is clearly different from the other two. Note that in other cancers, one often groups local and regional diseases, since both have potential to be controlled by specific interventions such as radiotherapy, whereas distant metastasis is not impacted directly by radiotherapy. Figure 1.1 depicts the clinical progressions that we would like to focus on. Based on it, management in clinical practice would be very different depending on the type of relapse that the patient is likely to experience.

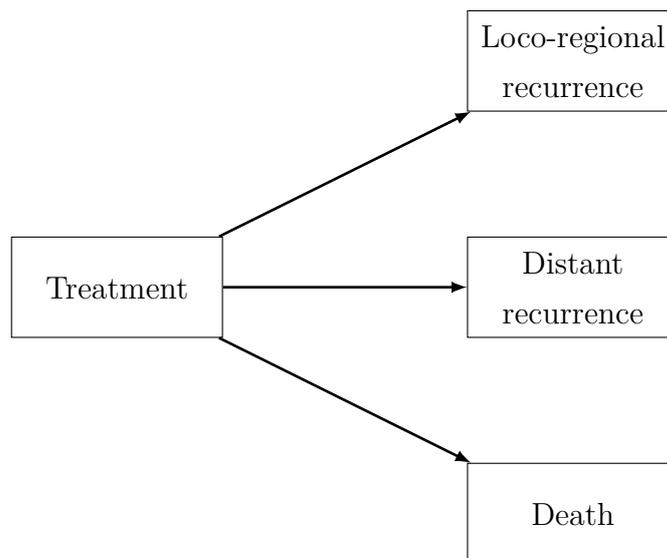


Figure 1.1 – Graphical representation of the possible first clinical recurrence in subjects treated for a localised prostate cancer.

In addition, the complete evolution process of the disease is not only depicted by one unique disease progression, but rather by multiple progressions that may occur successively. If we continue the example, it would be more appropriate to study the disease evolution process in its whole, by focusing on all the possible transitions that may experience the patient, as depicted in Figure 1.2. Here, after treatment, a patient may experience either a loco-regional recurrence, a distant recurrence or death. After a loco-

regional recurrence, he may experience either a distant recurrence or directly die. Finally, after a distant recurrence, a patient may only die. In all, 4 clinical states and 6 direct transitions are already considered in this basic example.

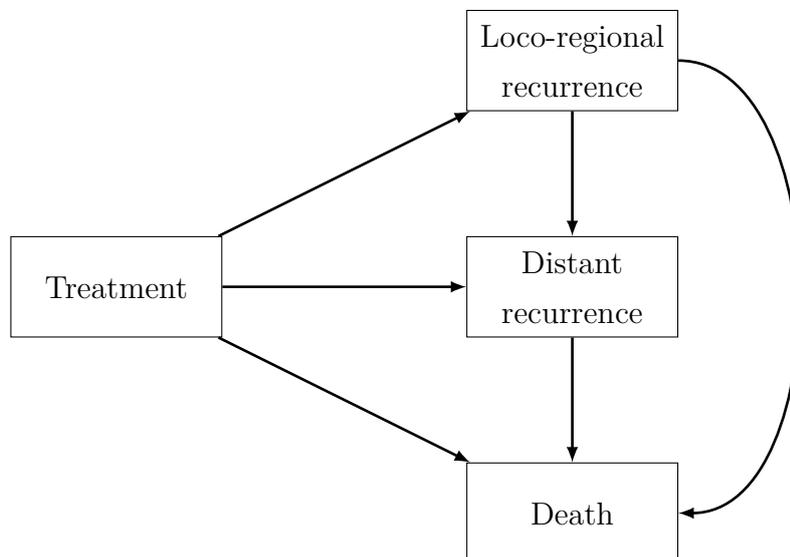


Figure 1.2 – Graphical representation of the possible transitions between clinical recurrences in subjects treated for a localised prostate cancer.

To understand the impact of the measured prognostic factors (such as disease marker data) on each stage of the disease, it is thus essential not only to distinguish the several natures of clinical progressions but also to focus on the identified transitions between them.

1.2.2 Dynamic personalized medicine

Another primordial aspect in clinical practice is the use of individualized predictions of each type of progression which can be updated at each follow-up visit of the subject or more generally at each new individual information. These individualized predictions are called *dynamic* because the considered information is dynamic: it could include for example a new measurement of a longitudinal marker or only the confirmation that the subject is still at risk of event [Proust-Lima and Blanche, 2015].

Figure 1.3 is an illustrative example of the principle of individual dynamic predictions in subjects with localised prostate cancer treated by radiotherapy (RT), from repeated measurements of a longitudinal biomarker and in the context of competing events. At the end of RT (baseline time), individual covariates are measured (in orange). Afterwards the subject is followed repeatedly by measuring the prostate specific antigen (PSA, green

crosses), famous prognostic factor of the risk of recurrence post-RT, where, without loss of generality, high PSA values induce a high risk of recurrence, and vice-versa. In these subjects, there is usually a drop in the PSA expression after the end of RT (due to the treatment effect), until reaching a certain nadir, and finally a more or less important PSA increase according to the disease evolution of the patient [Proust-Lima et al., 2008].

The predictions are computed from a landmark time s - time from which the prediction is made, and for a given horizon time w - window of prediction. At the top of the figure, the landmark time is $s = 1$ and we are interested in the cumulative incidences of recurrence (in blue) and death (in brown) in the next w years, from the individual information collected until $s = 1$. This information includes the one measured at baseline, four repeated PSA measures, and the fact that the subject is still free of clinical progression at $s = 1$. In the middle, the predictions are updated two years later ($s = 3$). Three supplementary PSA measures have been collected and we know that the subject did not experience any competing event in between. The same quantities of interest are updated at a later landmark time $s = 5$. In the figure, the variability around the estimated predictions is also depicted using intervals around the estimated landmark-specific cumulative incidences of each cause of event.

Such individualized dynamic predictions may notably be used to develop dynamic strategies in personalized medicine, for example either for adapting the treatment of each patient over time [Sène et al., 2016] or for planning the optimal time of his next screening visit [Rizopoulos et al., 2015]. These solutions might be expected to have manifest benefits. Of course, individualized treatment strategies might substantially improve the patients prognosis or/and their quality of life, but such individualized predictions might also be used in an economic perspective, for example by planning the optimal time of the next screening visit which not only minimises the probability of clinical event but also maximises the information gain brought by a new measure of the marker at this time point.

1.2.3 Various natures of longitudinal markers

In longitudinal health studies, one often observes continuous biomarkers such as the PSA data in prostate cancer or creatinine in acute kidney injury. However there are also more and more qualitative markers. For example, in HIV, one interesting marker is the detectability of the HIV viral load (binary: detectable/undetectable). In cancer, ordinal markers with only few modalities are also defined, such as the circulating tumor cell (CTC) biomarkers, or most of the summary scores built from Health-Related Quality

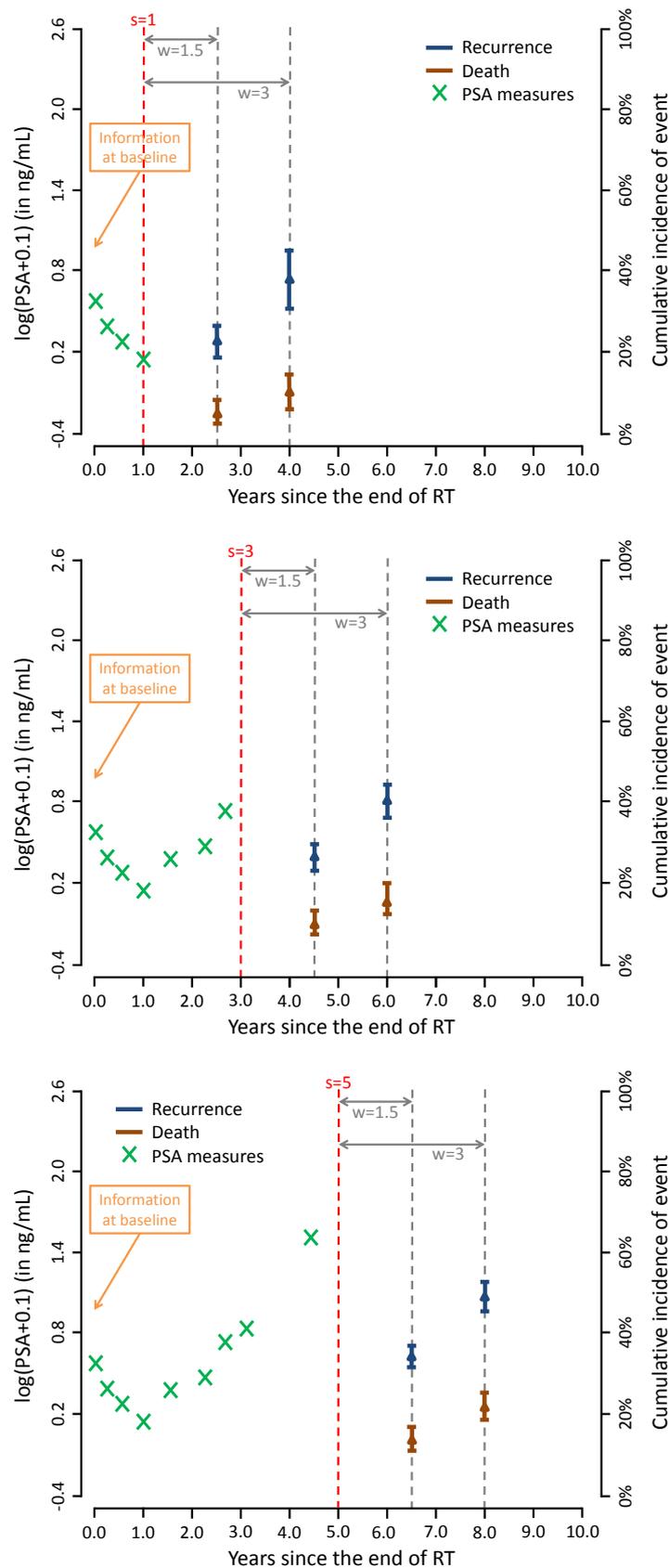


Figure 1.3 – Illustrative example in prostate cancer of individual cumulative incidences of two causes of event computed from three landmark times $s = 1$, $s = 3$ and $s = 5$ for two horizons $w = 1.5$ and $w = 3$.

of Life (HRQoL) questionnaires.

As for more standard continuous markers, one can be interested in modelling their trajectory over time or predicting their evolution while taking into account their association with the disease progression (characterised by the occurrence of clinical events).

1.3 Statistical issues

Multiple statistical methods have been developed in the literature to analyse either repeated measurements of a longitudinal marker or event history data. Nevertheless, the commonly used techniques to manage such data are no more valid when the two concomitant processes are correlated.

The mixed-effects models allow to model the repeated measurements of a longitudinal marker. These models are robust in the presence of missing data which are non-informative (or ignorable). This assumes that the missing data can be predicted from the individual covariates or the observations of the marker. When the missing data are informative (or non-ignorable), the model assumptions do not hold and the model may provide estimates off the mark [Little and Rubin, 2014; Rouanet et al., 2017]. These missing data mechanisms in longitudinal data analysis are detailed in Section 2.1.3.

Similarly, the event history models commonly used by statisticians, such as the Cox model with proportional hazards in survival data analysis [Cox, 1972], can only include time-dependent covariates assumed to be exogenous [Andersen and Gill, 1982; Fisher and Lin, 1999]. This assumption implies that the distribution of these covariates are not affected by the occurrence of the event of interest [Kalbfleisch and Prentice, 2011]. In case of violation of this assumption, that is in the presence of an endogenous covariate, the usual survival models may also provide non-accurate estimates. The endogenous and exogenous natures of the time-dependent covariates in event history models are introduced in Section 2.3.

These essential assumptions on both the missing data mechanisms of the longitudinal marker and the time-dependent covariates in event history models may be violated in many applications. For example, when we collect individual repeated biomarker data until the occurrence of a clinical event, clearly the missing data post-event may be informative and the repeated measurements of the biomarker may be endogenous.

1.3.1 Joint modelling introduction

To overcome the restrictive assumptions on the missing data mechanisms in usual mixed models for longitudinal data, and on the features of the time-dependent covariates in usual survival models for time-to-event data, the joint modelling approach has been introduced [Faucett and Thomas, 1996; Wulfsohn and Tsiatis, 1997]. The idea is to simultaneously model the longitudinal and survival processes, by linking them using a shared latent structure.

Joint models with shared random effects are the most common [Tsiatis and Davidian, 2004]. They take into account the correlation between the longitudinal and event history processes using a function of subject-specific random effects. However, the most classical joint model is limited to the joint analysis of a Gaussian longitudinal marker and survival data. More details are given in Section 2.3.

1.3.2 Extension to a multi-state process

Some authors extended the classical joint modelling framework to multiple time-to-event data, with more or less complex joint models [Elashoff et al., 2008; Dantan et al., 2011; Król et al., 2016]. But none of them developed a joint model for a longitudinal Gaussian marker and multi-state process, that would partition the disease states and the respective transitions between states.

We previously motivated that in many applications like cancer, the focus should be on transitions between clinical stages of the disease. Furthermore, other clinical events such as clinical interventions unplanned at baseline (new treatment for example) might be also included. With such a model, it would be possible to properly characterize the longitudinal marker evolution and quantify its impact on each transition intensity between the multiple clinical health states, while considering other non-endogenous prognostic factors.

1.3.3 Individual dynamic prediction tools

We previously argued that for each patient, management and treatment strategies differ widely according to his specific risk of each type of progression. However, the current prognostic tools provide an overall quantification of the risk of clinical event without distinction of the different types of progression [Maziarz et al., 2017; Goldstein et al., 2017]. Another crucial aspect is the longitudinal nature of many health studies, which allows to compute the individual predictions in a dynamic way, that is updated at each new subject's information [Proust-Lima and Blanche, 2015].

Several proposals have been proposed in the literature to compute the individual dynamic prediction of cumulative incidences of competing events from information including repeated biomarker data [Maziarz et al., 2017; Sène et al., 2016; Nicolaie et al., 2013b; Rizopoulos, 2011; van Houwelingen, 2007]. However, none formally defined the true probability of event to predict, validated its estimators and compared the different approaches of prediction using simulation studies. It is essential to fill the gap. The recommendations resulting from this work would be essential to ensure the most precise dynamic prediction tools possible in clinical practice.

1.3.4 Extension to ordinal longitudinal markers

Finally, we also discussed that in many applications the disease markers are discontinuous, and even ordinal with only few modalities. It was the case for example with most of HRQoL data in clinical cancer trials.

However, in the literature no software currently manages such data, and the solutions proposed in the literature rely on inference techniques that often lead to strong computational complexities [Li et al., 2010; Proust-Lima et al., 2016].

Thus, whether for providing correct estimates of the longitudinal evolution of the ordinal marker, or for properly characterizing the impact of its dynamics on the disease progression, it is necessary to have a model which takes into account the features of these data, while reducing the computational cost of the standard inference procedures.

1.4 Studied data

To deal with these clinical and statistical challenges, this thesis is based on data from two datasets comprising men treated for a clinically localised prostate cancer:

- data from the multi-center clinical trial RTOG 9406 (Radiation Therapy Oncology Group, USA) [Michalski et al., 2005], in which the data collection was conducted from 1994 to 2013;
- data from the cohort of the British Columbia Cancer Agency (BCCA) in Vancouver, Canada [Pickles et al., 2003], with subjects followed-up between 1994 and 2012.

In these two datasets, treatment was external beam radiotherapy and subjects were followed-up repeatedly since the end of treatment by measuring the prostate-specific antigen (PSA). They could experience several clinical events, such as loco-regional recurrence, distant recurrence or death (due or not to the prostate cancer). Additional therapy was

also possible with salvage hormonal therapy administration, which was unplanned at baseline but decided by clinicians according to physician observed signs in PSA or clinical signs, to prevent growth of potentially present sub-clinical cancer.

These data were essential to illustrate the purposes of the thesis. Thus, I thank warmly the data providers of these two datasets, and especially James Dignam and Tom Pickles.

1.5 Thesis structure

In biostatistics, the development of statistical models may have two main objectives: either for studying aetiology of a disease or for predicting a given quantity of medical interest. This thesis introduces statistical tools and software implementations for dealing with these two purposes in longitudinal cancer studies, based on two kinds of correlated data currently encountered: repeated measurements of a longitudinal marker and event history data.

Chapter 2 depicts a succinct state of the art of the statistical developments proposed in the literature for handling these two types of data.

Chapter 3 introduces a new joint model allowing to simultaneously study repeated measurements of a Gaussian longitudinal marker and individual times of transitions characterized by a multi-state process. The model is decomposed into a linear mixed model and a multi-state model with proportional intensities according to each covariate effect; it is estimated in the maximum likelihood framework.

Chapter 4 offers a complementary development for validating the estimates from the joint multi-state model: a score test for residual correlation between transition times in this joint multi-state model. The test statistic distribution is defined, with an analytic expression of the score statistic and an asymptotic expression of the variance corrected for the nuisance parameters which require numerical approximations.

These two developments are implemented in R, with easy-to-use functions for statisticians or clinicians. Simulation studies are performed to investigate the performances of both the estimation procedure of the joint multi-state model, and the type-I error and power of the score test. A detailed application is then carried out in subjects treated for a localised prostate cancer, by focusing on transitions between clinical progressions from PSA repeated measurements.

Chapter 5 focuses on the computation of individual dynamic predictions. It defines a true probability of event to predict and its estimators from two leading approaches: joint modelling and landmarking. Extensive simulation studies allow to validate these

definitions and compare the estimators in terms of bias, efficiency, prediction error and discriminatory power under several scenarios of well- or mis-specification of the models. Software developments are also provided for practical use.

Chapter 6 introduces preliminary results of an initiated work about joint modelling with shared random effects for an ordinal longitudinal marker and time-to-event data. An elegant inference solution is proposed with the use of exact likelihood.

Finally, Chapter 7 ends the manuscript with an overall conclusion which opens up both clinical and statistical prospects in cancer research.

2 State of the art

This chapter details the statistical methods on which the developments in this thesis rely. We first introduce mixed-effects models to deal with standard longitudinal data analysis in Section 2.1. Section 2.2 depicts methods for handling standard event history data, and Section 2.3 introduces the standard joint modelling which groups the two methodologies to analyse simultaneously the two kinds of data when they are correlated.

2.1 Mixed-effects models for longitudinal data analysis

For each subject $i \in \{1, \dots, N\}$, let us consider a marker $Y_i = (Y_{i1}, \dots, Y_{in_i})^\top$ which is observed at times $\{t_{ij}; j = 1, \dots, n_i\}$ with values $\{Y_{ij} = Y_i(t_{ij}); j = 1, \dots, n_i\}$, where n_i is the number of observed measures for subject i .

The central methodology to deal with this kind of data is the mixed-effects modelling. The main idea is that the expectation of the outcome is not only explained using population effects, but also using individual effects that captures both intra-subject correlation and heterogeneity over subjects which are unexplained by the covariates considered in the model.

2.1.1 Linear mixed-effects model

Let assume that marker Y is continuous and Gaussian. At each time point t_{ij} , marker value Y_{ij} is observed as a noisy measure of the true level Y_{ij}^* which remains unobserved. This latent marker value Y_{ij}^* is explained according to covariates at the population level with fixed effects β , and at the individual level with random effects b_i which depict the subject's deviation to his expected mean value, and at the same time capture the correlation between his repeated measurements:

$$\begin{aligned} Y_{ij} &= Y_{ij}^* + \epsilon_{ij} \\ &= X_{ij}^\top \beta + Z_{ij}^\top b_i + \epsilon_{ij}, \end{aligned} \tag{2.1}$$

where β is the p -vector of fixed effects associated with the design vector X_{ij} , b_i is the q -vector of random effects associated with the design vector Z_{ij} , ϵ_{ij} is the error term. The random effects b_i and the error terms $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{in_i})^\top$ are Gaussian, such that $b_i \sim \mathcal{N}(0, D)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I_{n_i})$, with I the identity matrix; b_i and ϵ_i are independent. This implies that two different observations Y_{ij} and $Y_{ij'}$ are assumed independent conditionally to b_i .

With its matrix formulation, the model becomes:

$$Y_i = X_i \beta + Z_i b_i + \epsilon_i, \quad (2.2)$$

where X_i and Z_i are matrices with row vectors X_{ij}^\top and Z_{ij}^\top , respectively.

Covariance structure for repeated measures

Model (2.2) was limited to the inclusion of random effects b_i which depict some individual deviations, classically associated with time functions, and an independent measurement error ϵ_i . With the general formulation of the linear mixed model (2.3), one can also include a residual correlation between serial measurements using ω_i [Laird and Ware, 1982; Jones and Boadi-Boateng, 1991]:

$$Y_i = X_i \beta + Z_i b_i + \omega_i + e_i, \quad (2.3)$$

where the residual error component is $\omega_i \sim \mathcal{N}(0, R_i)$ and the independent measurement error component is $e_i \sim \mathcal{N}(0, \sigma^2 I_{n_i})$. The three stochastic components b_i , ω_i and e_i , which represent random effects, serial correlation and measurement error respectively, are independent. To specify ω_i , standard choices are the use of autoregressive Gaussian process, also called AR(1) process, or Brownian motion, where both capture a residual temporal correlation between the individual repeated measurements [Diggle et al., 2002].

To simplify the formulations, the following is based on model (2.3) which does not consider any serial correlation term ω_i . But the same methodology applies when considering also stochastic deviations.

Estimation

The linear mixed model can be estimated using the maximum likelihood (ML) approach or the restricted maximum likelihood (REML) approach [Verbeke and Molenberghs, 2000].

STATE OF THE ART

Marker Y_i has a Gaussian distribution such that $Y_i \sim \mathcal{N}(X_i\beta, V_i = Z_i D Z_i^\top + \sigma^2 I_{n_i})$. The log-likelihood can be deduced:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^N \log f_Y(Y_i) \\ &= -\frac{1}{2} \sum_{i=1}^N \{n_i \log(2\pi) + \log(|V_i|) + (Y_i - X_i\beta)^\top V_i^{-1} (Y_i - X_i\beta)\}, \end{aligned} \quad (2.4)$$

where $\theta = (\beta^\top, D^\top, \sigma^2)^\top$ is the vector of model parameters, $|V_i|$ is the determinant of the matrix V_i .

In ML approach, a practical solution is to use the score equation for the fixed effects β :

$$\frac{\partial l(\theta)}{\partial \beta} = \sum_{i=1}^N X_i^\top V_i^{-1} (Y_i - X_i\beta) = 0.$$

For known parameters D and σ^2 , the estimate of β can be deduced:

$$\hat{\beta} = \left(\sum_{i=1}^N X_i^\top V_i^{-1} X_i \right)^{-1} \left(\sum_{i=1}^N X_i^\top V_i^{-1} Y_i \right).$$

In practice, parameters D and σ^2 are unknown. Thus one maximizes the log-likelihood on (D, σ^2) using numerical optimisation algorithms (as the Newton-Raphson algorithm for example), by replacing the fixed effects by their estimates from the score equation.

Based on the maximum likelihood theory, we can easily deduce the variance of the estimates $\hat{\beta}$:

$$\begin{aligned} \text{var}(\hat{\beta}) &= \left\{ -\mathbb{E} \left(\frac{\partial^2 l(\theta)}{\partial \beta \partial \beta^\top} \right) \right\}^{-1} \\ &= \left(\sum_{i=1}^N X_i^\top V_i^{-1} X_i \right)^{-1} \end{aligned}$$

and the estimate $\widehat{\text{var}}(\hat{\beta})$ substitutes V_i for $\widehat{V}_i = Z \widehat{D} Z^\top + \widehat{\sigma}^2 I_{n_i}$, with \widehat{D} and $\widehat{\sigma}^2$ the ML or REML estimates of D and σ^2 , respectively.

A draw back of the ML approach is that it does not take into account the loss in the degrees of freedom resulting from estimating fixed effects, and thus estimates \widehat{D} and $\widehat{\sigma}^2$ are biased. To obtain correct estimates \widehat{D} and $\widehat{\sigma}^2$, one can use the REML approach. Let us define vectors $Y = (Y_1^\top, \dots, Y_N^\top)^\top$, $b = (b_1^\top, \dots, b_N^\top)^\top$ and $\epsilon = (\epsilon_1^\top, \dots, \epsilon_N^\top)^\top$, each of length $n = \sum_i n_i$. One also denotes the (n, p) matrix $X = (X_1, \dots, X_N)^\top$ and the (n, q) matrix $Z = \text{diag}(Z_1, \dots, Z_N)$. Thus we can combine all the subject-specific regression

models (2.1) to one model: $Y = X\beta + Zb + \epsilon$. The REML idea is to take into account the dimension of the orthogonal vectorial space of X by considering the likelihood of the transformed data $L^\top Y$, with L a $(n, n - p)$ full-rank matrix such as $L^\top X = 0$. Thus one has $L^\top Y \sim \mathcal{N}(L^\top X\beta = 0, L^\top \text{var}(Y)L)$ and the associated likelihood does not depend on β . Up to a constant, the restricted likelihood is [Harville, 1974]

$$l_{\text{REML}}(D, \sigma^2) = l(\hat{\beta}, D, \sigma^2) - \frac{1}{2} \log \left| \sum_{i=1}^N X_i^\top V_i^{-1} X_i \right|. \quad (2.5)$$

The variance of the parameter estimates is obtained in the same way as in ML approach, by replacing the ML estimates of D and σ^2 by the REML estimates.

When the focus is on the variance parameters and the number of fixed effects is large or/and the sample size is small, one prefers to use the REML approach. However, the REML technique does not allow to compare models with different fixed effects, and ML and REML techniques give similar estimates when the sample size is large so that ML approach is often adopted.

Note that another possible solution, less common in software because computationally more costly, is to directly optimise the likelihood (2.4) according to the complete vector of parameters θ . The variance of $\hat{\theta}$ is then deduced using numerical computation of the Fisher information matrix according to the whole vector of parameters θ , i.e. using $\text{var}(\hat{\theta}) = \left\{ -\mathbb{E} \left(\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^\top} \right) \right\}^{-1}$.

Predictions

Predictions represent here future observed or unobserved values of Y for a (possibly new) subject, or fitted values at observation times for a subject used in the model estimation. From linear mixed-effects models, two kinds of predictions are studied: predictions marginal or conditional to the random effects (and more generally to the random deviation).

Let us denote $\hat{\theta}$ the complete vector of coefficients estimated from the linear mixed model. The marginal predictions $\hat{Y}_i^M = X_i \hat{\beta}$ depict the average predictions of Y_i for subjects with the same covariates as subject i , whereas the individual predictions conditional to the random effects represent the specific predictions for a given subject i , including his individual deviations to the mean trajectory in the population with the same characteristics:

$$\hat{Y}_i^{SS} = X_i \hat{\beta} + Z_i \hat{b}_i$$

with $\widehat{\beta}$ the estimated fixed effects and \widehat{b}_i the predicted individual random effects or empirical Bayesian estimates which are deduced as follows.

As the joint distribution (Y_i, b_i) is such that

$$\begin{pmatrix} b_i \\ Y_i \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D & DZ_i^\top \\ Z_i D & V_i \end{pmatrix} \right),$$

the conditional expectation of the random effects given the observations is $\mathbb{E}(b_i|Y_i) = DZ_i^\top V_i^{-1}(Y_i - X_i\beta)$ and we denote \widehat{b}_i the Best Linear Unbiased Predictor (BLUP) of b_i :

$$\widehat{b}_i = \mathbb{E}(b_i|Y_i; \widehat{\theta}) = \widehat{D}Z_i^\top \widehat{V}_i^{-1}(Y_i - X_i\widehat{\beta}).$$

2.1.2 Generalized linear mixed-effects model

The linear mixed model allows to study the longitudinal evolution only of a Gaussian marker. The generalized linear mixed-effects model combines the generalized linear model and the linear mixed model to generalize the type of marker's distribution to any distribution of the exponential family [Fitzmaurice et al., 2012], that is with a conditional density of the form

$$f_Y(Y_{ij}|b_i; \psi_{ij}, \phi) = \exp \left[\frac{Y_{ij}\psi_{ij}(b_i) - c\{\psi_{ij}(b_i)\}}{a(\phi)} - d(Y_{ij}, \phi) \right], \quad (2.6)$$

where $c(\cdot)$, $d(\cdot)$ and $a(\cdot)$ are known functions, $\psi_{ij}(b_i)$ and ϕ denote the natural (also called canonical) and dispersion parameters in the exponential family, respectively.

In the generalized linear mixed-effects model, one directly models the expectation of the observed measures (and thus the error term disappears), and this quantity is related to the covariate effects using a link function g :

$$g\{\mathbb{E}(Y_{ij}|b_i)\} = X_{ij}^\top \beta + Z_{ij}^\top b_i, \quad (2.7)$$

where β is the vector of fixed effects and b_i is the vector of random effects for subject i which is usually Gaussian with $b_i \sim \mathcal{N}(0, D)$. Note however, that in some cases, the Gaussian distribution can be replaced by another distribution that simplifies computational aspects [Molenberghs et al., 2007; Verbeke et al., 2014].

The model is estimated using the maximum likelihood approach. The observations Y_{ij} are independent conditionally to the random effects b_i , i.e. $f_Y(Y_i|b_i; \theta) = \prod_{j=1}^{n_i} f_Y(Y_{ij}|b_i; \theta)$ and the observed likelihood is obtained by integrating over the individual random effects

distribution:

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^N f_Y(Y_i; \theta) \\
&= \prod_{i=1}^N \int_{\mathbb{R}^q} f_Y(Y_i|b_i; \theta) f_b(b_i; \theta) db_i \\
&= \prod_{i=1}^N \int_{\mathbb{R}^q} \prod_{j=1}^{n_i} f_Y(Y_{ij}|b_i; \theta) f_b(b_i; \theta) db_i,
\end{aligned} \tag{2.8}$$

where $f_b(b_i; \theta)$ is the density of the random effects.

In most cases, the integral over the random effects does not have a close form and has to be achieved using numerical integration (with Gaussian quadratures or Monte-Carlo techniques for example). The log-likelihood is maximised using optimisation algorithms such as, for example, the EM algorithm, the Newton-Raphson algorithm (or derivatives) or the Marquardt algorithm. The estimated variance of the parameter estimates can be obtained by solving the observed Fisher information matrix at $\hat{\theta}$.

Among all the kinds of data from the exponential family that can manage the generalized linear mixed-effects models, we only focus in the following on two: binary data (e.g. detectability/undetectability of the HIV viral load) and ordinal data (e.g. summary score from HRQoL questionnaire) which are useful for the thesis developments.

Model for repeated binary observations

For binary repeated data ($Y \in \{0, 1\}$), one models the probability that each observation equals 1: $\mathbb{E}(Y_{ij}|b_i) = \Pr(Y_{ij} = 1|b_i)$. A common solution is to consider that the link function g is a logistic function ($g(x) = \text{logit}(x)$) [Stiratelli et al., 1984]. The logistic mixed model is written as

$$\text{logit}\{\Pr(Y_{ij} = 1|b_i)\} = X_{ij}^\top \beta + Z_{ij}^\top b_i$$

and then

$$\Pr(Y_{ij} = 1|b_i) = \frac{\exp(X_{ij}^\top \beta + Z_{ij}^\top b_i)}{1 + \exp(X_{ij}^\top \beta + Z_{ij}^\top b_i)}.$$

Another solution developed in the literature is to assume that there is a latent Gaussian variable Y_{ij}^* such that $Y_{ij} = \mathbb{1}\{Y_{ij}^* \geq 0\}$. Then we can model Y^* using a linear mixed model: $Y_{ij}^* = X_{ij}^\top \beta + Z_{ij}^\top b_i + \epsilon_{ij}$ with $b_i \sim \mathcal{N}(0, D)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. The model becomes

$$\begin{aligned}
\Pr(Y_{ij} = 1|b_i) &= \Pr(Y_{ij}^* \geq 0|b_i) \\
&= \Pr(\epsilon_{ij} \leq X_{ij}^\top \beta + Z_{ij}^\top b_i) \\
&= \Phi(X_{ij}^\top \beta + Z_{ij}^\top b_i),
\end{aligned}$$

STATE OF THE ART

and thus the link function g is a probit function ($g(x) = \Phi^{-1}(x)$, where $\Phi(\cdot)$ is a probability density function of a standard Gaussian distribution). This model is called a probit mixed model.

In practice, these two models are very close since the normal and logistic distributions are very similar except for extreme values [Hahn and Soyer, 2005]. The choice of one or another usually depends on the application domain.

Model for repeated ordinal observations

Let us now consider that Y takes ordinal values in the finite space $\{0, 1, \dots, M\}$. The same technique as for the probit mixed model for binary observations applies to the ordinal case. Henceforth we consider that there is a latent Gaussian variable Y^* such that observing the marker at a certain value is equivalent to have its associated latent variable in a given interval:

$$Y_{ij} = m \quad \Leftrightarrow \quad \alpha_m < Y_{ij}^* \leq \alpha_{m+1}$$

for $m \in \{0, \dots, M\}$, where coefficients $\{\alpha_m; m = 1, \dots, M\}$ are unknown thresholds, $\alpha_0 = -\infty$ and $\alpha_{M+1} = +\infty$.

One thus models the probability to have the marker's observation under a certain threshold using a probit function,

$$\Pr(Y_{ij} \leq m | b_i) = \Phi(\alpha_{m+1} - X_{ij}^\top \beta - Z_{ij}^\top b_i)$$

for $m \in \{0, \dots, M-1\}$. This model is called cumulative probit mixed model.

To ensure identifiability of the model, two constraints have to be applied. In addition to the variance of the error term σ^2 which is here fixed to 1, one can choose for example to consider that the fixed intercept β_0 is null or to fix a threshold, for instance $\alpha_1 = 0$.

The model is classically estimated in the maximum likelihood framework. The likelihood is the same as in equation (2.8), with the conditional density of the longitudinal outcome:

$$\begin{aligned} f_Y(Y_{ij} | b_i; \theta) &= \Pr(Y_{ij} = m | b_i; \theta) \\ &= \Pr(\alpha_m < Y_{ij}^* \leq \alpha_{m+1} | b_i; \theta) \\ &= \Phi(\alpha_m - X_{ij}^\top \beta - Z_{ij}^\top b_i, \alpha_{m+1} - X_{ij}^\top \beta - Z_{ij}^\top b_i) \end{aligned}$$

where $\Phi(a, b)$ is the cumulative distribution function of a standard Gaussian variable truncated at a and computed at b .

Note that by using a logistic link function and by considering that the error terms follow a logistic distribution, the same technique can be used to define the cumulative logistic mixed model.

2.1.3 Hypotheses on missing data

In longitudinal data analysis, two types of missing data can arise: intermittent missing data, that is when a subject misses a visit but comes back at the next one, or missing data due to drop-out, that is when a subject leaves the study. Little and Rubin [2014] proposed to classify the missing data mechanisms in 3 types:

- MCAR (Missing Completely At Random): the individual probability to have a missing data is conditional to the individual covariates but is independent from the observations and missing observations of the marker; for example if the subject no longer comes at visits because he moved to the other side of the planet;
- MAR (Missing At Random): the individual probability to have a missing data is conditional to the individual covariates as well as the observed marker values, but is independent of the unobserved measures; for example when the subject has a major deterioration of its marker values between two consequent visits and decides to stop coming at the next follow-up visits;
- MNAR (Missing Not At Random): the individual probability to have a missing data is conditional to the individual covariates as well as the observed and unobserved marker values; for example when the subject no longer comes because he experienced a major deterioration since his last visit which was not predictable from the past observations.

The estimates obtained using ML (or REML) are unbiased in presence of MCAR or MAR data, since the missing data can be predicted by the model's information; one talks about *ignorable* missing data. However, these inferences techniques can lead to biased estimates in case of MNAR data, and one talks about *non-ignorable* missing data [Fitzmaurice et al., 1995; Molenberghs and Verbeke, 2001; Kurland and Heagerty, 2005; Rouanet et al., 2017].

In that case, other techniques have to be considered [Little, 1993; Michiels et al., 2002; Diggle and Kenward, 1994; Henderson et al., 2000]. Among them, joint models that will be detailed in Section 2.3, simultaneously model the marker and the missing data mechanism.

2.2 Models for event history data analysis

This thesis also focuses on methods for the analysis of event history data. The analysis of event history data consists in studying time(s) of individual transition(s) between states. The specificity of such data is that the quantity of interest is not always observable because of truncation and censoring which can prevent the observation of the event time. Methods to analyse these kinds of data have been characterized through the counting process theory developed by Andersen et al. [1993].

The censoring mechanisms are traditionally classified into three categories: right censoring, left censoring and interval censoring. In this thesis we only focus on the former, right censoring, that occurs when the subject experiences the event after its last follow-up time, for example because he is lost to follow-up before the end of the study without this explained by his condition or because he is still alive when the study ends. As in mixed-effects models, censoring can be distinguished into two types: informative censoring (similar to MNAR missing data), which assumes that the censoring mechanism is related to the prognosis of the patient unexplained by the covariates in the model, and non-informative censoring (similar to MCAR missing data), which stipulates that the censoring mechanism is conditional to the individual covariates but independent of the unexplained prognosis of the subject. In this thesis we make the task easier by focusing on non-informative right censored data only.

Truncation constitutes another cause of unobservation. It assumes that the event is observed conditionally to another event. This work only considers left truncation, which is the most common in practice: the subject is observable only if he is at risk at the inclusion in the study.

In time-to-event analyses, several choices of time origin can be chosen, such as the subject's birth (the time is then the age of the subject) or the study entry (the considered time is the time spent since the inclusion). These imply different interpretations of the results [Andersen, 2017; Joly et al., 2013]. In this thesis, the time is considered as the time since the inclusion in the study. This is common in clinical studies where the focus is on the subject's prognosis after the administration of a treatment. In the presence of multiple events, the calendar time will be considered, that assumes that the baseline time is always the study entry time, contrary to the gap time which defines the time as the

time spent in the current state [Putter et al., 2007; Duchateau et al., 2003].

In the following, we first introduce models for handling survival data in Section 2.2.1 and generalize them to the multi-state data framework in Section 2.2.2. We finally end with an introduction to landmark approach and frailty models in Sections 2.2.3 and 2.2.4, respectively.

2.2.1 Survival data analysis



Figure 2.1 – Representation of the allowed transition between states in the survival setting.

The simplest case of time-to-event analysis is the survival analysis: the subject is in a initial state and we are interested in his transition time leading to another state. This example is illustrated in Figure 2.1. In clinical studies, typically state 0 is defined as healthy and state 1 is defined as death or recurrence of the disease. We observe $T_i^* = \min(T_i, C_i)$ with T_i the true event time and C_i the right censoring time. The indicator of event is denoted δ_i with $\delta_i = \mathbb{1}\{T_i \leq C_i\}$ which equals either 1 if the subject had the event of interest or 0 if he was censored.

The distribution of survival data can be depicted with several interesting functions depending on the objective, as for example the cumulative distribution function, the survival function or the hazard function. First, we introduce the probability density function

$$f_i(t) = \lim_{dt \rightarrow 0} \frac{\Pr(t \leq T_i < t + dt)}{dt}.$$

The cumulative distribution function can be derived:

$$F_i(t) = \Pr(T_i \leq t) = \int_0^t f_i(u) du$$

and the survival function, which is the probability that the event does not occur until time t , can be expressed as

$$S_i(t) = \Pr(T_i > t) = 1 - F_i(t).$$

The hazard function is defined as the instantaneous probability of event given that the subject is still at risk:

$$\lambda_i(t) = \lim_{dt \rightarrow 0} \frac{\Pr(t \leq T_i < t + dt | T_i \geq t)}{dt} = \frac{f_i(t)}{S_i(t)}$$

and we can deduce the cumulative hazard function

$$\Lambda_i(t) = \int_0^t \lambda_i(u) \, du.$$

Finally, the relationship between survival and hazard functions is such as:

$$S_i(t) = \exp(-\Lambda_i(t)) = \exp\left(-\int_0^t \lambda_i(u) \, du\right).$$

To estimate these quantities on data, two approaches can be used. The non-parametric estimators describe these quantities of interest in a crude way while the regression models (notably the proportional hazards models) describe these quantities as functions of explanatory variables.

Non-parametric estimators

Let us denote $Y(t)$ the number of subjects at risk at time t and $N(t)$ the number of observed survival times until t . In the following, the introduced non-parametric estimators omit the subscript i because they are not subject-specific, but rather depict the quantity of interest in the whole population.

A famous non-parametric estimator of the survival function is the Kaplan-Meier estimator [Kaplan and Meier, 1958]:

$$\hat{S}_{KM}(t) = \prod_{u \leq t} \left(1 - \frac{dN(u)}{Y(u)}\right) \quad (2.9)$$

where $dN(t)$ is the number of events at time t . The variance of $\hat{S}_{KM}(t)$ can be estimated using the Greenwood formula [Greenwood, 1926]:

$$\widehat{\text{var}}\{\hat{S}_{KM}(t)\} = \{\hat{S}_{KM}(t)\}^2 \sum_{u \leq t} \frac{dN(u)}{Y(u)(Y(u) - dN(u))}$$

and using the asymptotic normal distribution of $\hat{S}_{KM}(t)$, the 95% confidence interval can be deduced as

$$\left[\hat{S}_{KM}(t) \pm q_{0.025} \sqrt{\widehat{\text{var}}\{\hat{S}_{KM}(t)\}} \right]$$

where $q_{0.025}$ is the 2.5th percentile of a standard normal distribution. This interval can however include values outside the interval $[0, 1]$, notably when the sample size is small. A solution is to use a log-transformation to define $\log(\hat{S}_{KM}(t))$ and its variance $\widehat{\text{var}}\{\log \hat{S}_{KM}(t)\} = \sum_{u \leq t} \frac{dN(u)}{Y(u)(Y(u) - dN(u))}$. To be absolutely assured that the estimated confidence interval remains in $[0, 1]$, another solution is to apply a log-log transformation $\log(-\log(S(t)))$ and define the associated variance using the delta-method to deduce a correct confidence interval.

An alternative to the Kaplan-Meier estimator is the Nelson-Aalen estimator [Nelson, 1969; Aalen, 1978] which defines the estimator of the cumulative hazard function and its variance as

$$\widehat{\Lambda}_{NA}(t) = \sum_{u \leq t} \frac{dN(u)}{Y(u)} \quad \text{and} \quad \widehat{\text{var}}\{\widehat{\Lambda}_{NA}(t)\} = \sum_{u \leq t} \frac{dN(u)}{Y(u)^2}. \quad (2.10)$$

The estimator of the survival function can then be deduced as $\widehat{S}_{NA}(t) = \exp(-\widehat{\Lambda}_{NA}(t))$. The latter is a particular case of the Aalen-Johansen estimator (2.18) which will be further introduced to deal with multi-state data.

The Kaplan-Meier and Nelson-Aalen methods are asymptotically equivalent to estimate the survival function. Many authors discussed about their differences but it is not the goal here [Colosimo et al., 2002].

Proportional hazards model

In health studies, a major objective is to understand the occurrence of an event of interest, such as a clinical event, from given explanatory covariates. To reach this goal, a widely used method is the proportional hazards (PH) models.

The proportional hazards model is based on the individual instantaneous hazard of event which is explained according to covariates through a linear predictor:

$$\begin{aligned} \lambda_i(t) &= \lim_{dt \rightarrow 0} \frac{\Pr(t < T_i \leq t + dt | T_i > t)}{dt} \\ &= \lambda_0(t) \exp\{X_i^\top \gamma\} \end{aligned} \quad (2.11)$$

where X_i is a vector of covariates measured at baseline, γ the associated vector of regression parameters and $\lambda_0(t)$ is the baseline hazard.

The main assumption of this model is the proportionality of hazards according to the levels of each covariate. The proportional hazards assumption facilitates interpretation. With only one covariate, the hazard ratio can be defined as $\frac{\lambda_i(t|X_i = x + 1)}{\lambda_i(t|X_i = x)} = \exp(\gamma)$, and the relation $\lambda_i(t|X_i = x + 1) = \exp(\gamma) \times \lambda_i(t|X_i = x)$ can be defined. It links the two hazards only using the coefficient of proportionality $\exp(\gamma)$ which is invariant over time. When the model includes $p > 1$ covariates, the hazard ratio is defined conditionally to the other covariates: $\frac{\lambda_i(t|X_{i,p} = x + 1, X_{i,-p} = y)}{\lambda_i(t|X_{i,p} = x, X_{i,-p} = y)} = \exp(\gamma_p)$.

In parametric proportional hazards models, $\lambda_0(t)$ has a parametric expression. Common choices of distribution include exponential, Weibull, Gompertz, Gamma, piecewise

STATE OF THE ART

constant functions or spline functions. For example, with the exponential distribution, the baseline hazard is constant over time: $\lambda_0(t) = \lambda_0$. With the Weibull function, the baseline hazard progresses over time, although monotonous: $\lambda_0(t) = \rho\nu^\rho t^{\rho-1}$, where ρ and ν are unknown parameters.

As the subjects are independent, the likelihood $L(\theta)$ is the product of the individual contributions to the likelihood $L_i(\theta)$, θ the vector of unknown parameters to estimate. If subject i is censored at T_i^* , $L_i(\theta)$ equals the survival probability at the observed time. If subject i has the event at T_i^* , $L_i(\theta)$ equals the survival probability at the observed time multiplied by the hazard at this time $\lambda_i(T_i^*)$:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^N L_i(\theta) \\ &= \prod_{i=1}^N \lambda_i(T_i^*; \theta)^{\delta_i} S(T_i^*; \theta) \end{aligned}$$

Usually, the log-likelihood $l(\theta) = \log L(\theta)$ is maximized using iterative optimization algorithms.

However, the specification of the parametric baseline hazard is not always obvious in practice, and a misspecification of $\lambda_0(t)$ may lead to biased estimates [Struthers and Kalbfleisch, 1986]. To deal with this, a semi-parametric proportional hazards model, which avoids the specification of the baseline hazard, has been proposed in the literature [Cox, 1972]. Over the years, this model, called the Cox proportional hazards model, has become the standard regression model in survival analysis. It explains the individual instantaneous hazard of event according to a baseline hazard and individual covariates, as previously:

$$\begin{aligned} \lambda_i(t) &= \lim_{dt \rightarrow 0} \frac{\Pr(t < T_i \leq t + dt | T_i > t)}{dt} \\ &= \lambda_0(t) \exp\{X_i^\top \gamma\}. \end{aligned}$$

However, the baseline hazard $\lambda_0(t)$ is now kept unspecified.

The Cox PH model can be estimated using the partial likelihood [Breslow, 1972]:

$$\begin{aligned} pL(\gamma) &= \prod_{i=1}^N pL_i(\gamma) \\ &= \prod_{i=1}^N \int_0^\infty \frac{\exp(X_i^\top \gamma)}{\sum_j Y_j(t) \exp(X_j^\top \gamma)} dJ_i(t) \end{aligned}$$

where $Y_j(t) = \mathbb{1}\{T_j^* \geq t\}$ and $J_i(t) = \mathbb{1}\{T_i^* \leq t, \delta_i = 1\}$. The estimator of the cumulative baseline hazard (called Breslow's estimator) can be deduced:

$$\widehat{\Lambda}_0(t) = \int_0^t \widehat{\Pi}^{(0)}(\widehat{\gamma}, u)^{-1} d\overline{J}(u) \quad (2.12)$$

where $\widehat{\Pi}^{(0)}(\widehat{\gamma}, u) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{T_i^* \geq u\} \exp\{X_i^\top \widehat{\gamma}\}$ and $\overline{J}(u) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{T_i^* \leq u, \delta_i = 1\}$.

Pseudo-value approach

The models previously introduced are built on the specification of hazard functions which relate individual instantaneous hazard and individual covariates. It makes them particularly relevant for aetiology purposes.

However, consider now that we are interested for each subject i in another quantity of interest, such as his individual probability of event $\pi_i(t) = \Pr(T_i \leq t)$. With the aim to model or predict $\pi_i(t)$, some authors introduced models based on the direct modelling of a cumulative probability of event, with for example the Fine-Gray model [Fine and Gray, 1999] or the binomial regression models [Scheike et al., 2008] (both in the presence of competing events), or the pseudo-value approach [Andersen and Pohar Perme, 2010]. Only the latter is developed here.

The idea of the pseudo-value (also called pseudo-observation) approach is that the probability of event is the expectation of an indicator of event: $\pi_i(t) = \mathbb{E}\{\mathbb{1}(T_i \leq t)\}$. However this indicator of event is not always observed in practice because of censoring. So it can be replaced by the pseudo-observation $\widehat{\mu}_i(t)$ of $\mathbb{1}(T_i \leq t)$ which is defined as the jackknife estimator of the non-parametric estimator $\widehat{F}(t)$ of the quantity of interest $\pi_i(t)$:

$$\widehat{\mu}_i(t) = N\widehat{F}(t) - (N-1)\widehat{F}_{(-i)}(t) \quad (2.13)$$

where $\widehat{F}(t) = 1 - \widehat{S}(t)$ with $\widehat{S}(t)$ the Kaplan-Meier estimator of the survival function on the whole sample (2.9), and $\widehat{F}_{(-i)}(t)$ is the same quantity estimated when the subject i is eliminated from the sample.

Each individual pseudo-observation $\widehat{\mu}_i(t)$ is then related to individual covariates X_i using a generalized linear model with a link function g :

$$g[\mathbb{E}\{\widehat{\mu}_i(t)\}] = X_i^\top \gamma \quad (2.14)$$

where the link function g may be chosen for interpretation purpose or inference facilities. In practice the identity, log or clog functions are the most commonly used.

The model is then estimated using generalized estimating equations (GEE) [Liang and Zeger, 1986].

The pseudo-observation approach allows to quantify direct effects of covariates on probabilities of event and has the advantage to avoid the computation of integrals over time when interested in probabilities of event, which is frequently encountered in prediction setting.

2.2.2 Multi-state formalism

Survival analysis allows the study of a first event time. When this event has possibly multiple distinct causes (e.g., a treatment failure defined as disease recurrence or death), one talks about survival analysis with competing risks [Prentice et al., 1978]. But time-to-event data analysis usually comprises a set of times of interest, not only one. Examples include recurrent events (e.g., when we focus on a treatment for repeated asthma crisis in asthmatic patients) or more generally multi-state data (e.g., when we focus on a disease evolution process depicted through multiple possible transitions between disease stages). The theory of multi-state processes unifies the time-to-event data analysis [Meira-Machado et al., 2009; Andersen and Keiding, 2002; Hougaard, 1999]. In the following, I detail this theory by relying on the book of Andersen et al. [1993], which introduced the counting process theory for dealing with event history processes.

For each subject $i \in \{1, \dots, N\}$, let us consider multi-state process $E_i = \{E_i(t), T_{i0} \leq t \leq C_i\}$ which is observed between the left truncation time T_{i0} and the right censoring time C_i , with values in the finite space $S = \{0, 1, \dots, M\}$, where $E_i(t)$ denotes the state occupied by subject i at time t .

The survival process is the simplest case of multi-state processes (see Figure 2.1): one considers $M = 1$, subjects are at risk in the transient state 0, which assumes that $\Pr(E_i(t) = 0 | E_i(s) = 0) \neq 1$ for $s < t$, and they can transit to the absorbing state 1, which satisfies $\Pr(E_i(t) = 1 | E_i(s) = 1) = 1$. Note that when the state is not absorbing, it is transient.

When we distinguish several types of event, as depicted in Figure 2.2 with $M = 2$ causes of event, we obtain the competing risks setting, where state 0 is the transient state and states $1, \dots, M$ are absorbing states. The multi-state theory can also handle subsequent events that occur after a first one. In that case, the focus is on time(s) of transition, as illustrated in Figure 2.3 which represents the possible transitions in an

unidirectional illness-death process (state 0: healthy, state 1: illness, state 2: death). In this case, a subject can experience several transient states before the absorbing state.

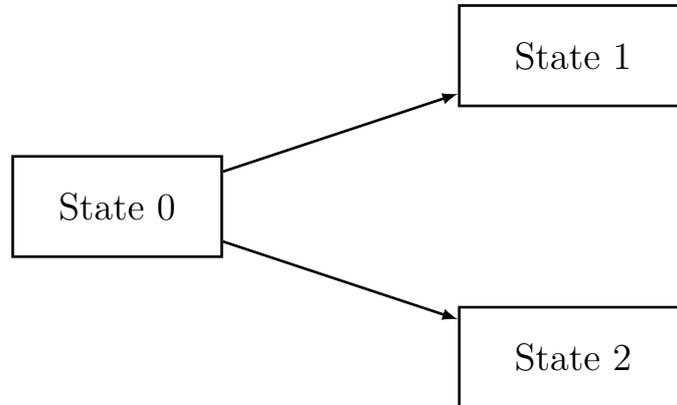


Figure 2.2 – Illustrative example of a competing risks setting with one transient state and two absorbing states.

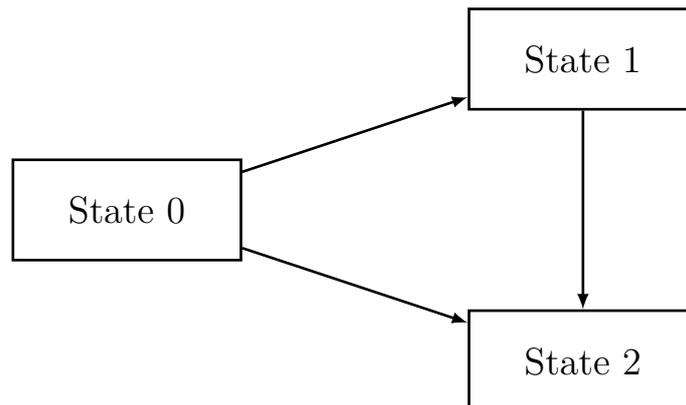


Figure 2.3 – Illustrative example of a multi-state process with two transient states and an absorbing state, model also known as illness-death model.

In the following we assume that process E_i is Markovian and non-homogeneous. The Markovian assumption ensures that the future of the process depends only of the current state and not of the past state(s) (i.e. $\Pr(E_i(t+u) = k | E_i(t) = h, E_i(s), s < t) = \Pr(E_i(t+u) = k | E_i(t) = h), \forall (h, k) \in S^2, \forall u \geq 0$) while the non-homogeneous assumption means that the evolution of the process varies with time.

As in survival analysis, the multi-state data analysis is mainly based on several functions which represent key quantities of interest, such as the transition intensities or the

transition probabilities.

The intensity of transition from state $h \in S$ to state $k \in S$ at time t is defined as

$$\lambda_{hk}^i(t) = \lim_{dt \rightarrow 0} \frac{\Pr(E_i(t+dt) = k | E_i(t) = h)}{dt} \quad (2.15)$$

with $\lambda^i(t) = \{\lambda_{hk}^i(t)\}$ the $(M+1) \times (M+1)$ matrix of individual transition intensities.

The matrix of individual cumulative transition intensities, noted $\Lambda^i(t)$, is composed of non-diagonal elements

$$\Lambda_{hk}^i(t) = \int_0^t \lambda_{hk}^i(u) du, \forall h \neq k,$$

and diagonal elements $\Lambda_{hh}^i(t) = -\sum_{k \neq h} \Lambda_{hk}^i(t)$.

The transition probability defines the probability that a subject in state h at time s occupies state k at a later time t :

$$P_{hk}^i(s, t) = \Pr(E_i(t) = k | E_i(s) = h), \text{ with } s \leq t. \quad (2.16)$$

We then call $P^i(s, t) = \{P_{hk}^i(s, t)\}$ the matrix of transition probabilities, which satisfies the Chapman-Kolmogorov equation:

$$P^i(s, t) = P^i(s, u)P^i(u, t), \text{ with } 0 \leq s \leq u \leq t.$$

$P^i(s, t)$ is the unique solution of the Kolmogorov forward differential equations:

$$\begin{aligned} P^i(s, s) &= I, \\ \frac{\partial}{\partial t} P^i(s, t) &= P^i(s, t) \lambda^i(t). \end{aligned}$$

This result is based on the system of differential equations for all the possible states $(h, k, l) \in \mathcal{S}^3$: $\frac{\partial P_{hl}^i(s, t)}{\partial t} = \sum_k P_{hk}^i(s, t) \lambda_{hk}^i(t)$; $P_{hk}(s, s) = \delta_{hk}$, with $\delta_{hk} = 1$ if $h = k$ and 0 otherwise.

Like in survival analysis, both non-parametric estimators of these quantities of interest and regression models have been proposed.

Non-parametric estimators

Let $N_{hk}(t)$ be the number of direct observed transitions from state h to state k up to time t , and $Y_h(t)$ the number of individuals in state h just before time t . The non-parametric estimator of the cumulative intensities, called $\Lambda^*(t)$ has elements $\Lambda_{hk}^*(t)$ estimated through the Nelson-Aalen estimator:

$$\hat{\Lambda}_{hk}^*(t) = \int_0^t \frac{dN_{hk}(u)}{Y_h(u)}, h \neq k, \quad (2.17)$$

and $\Lambda_{hh}^*(t) = -\sum_{k \neq h} \Lambda_{hk}^*(t)$. The solution to the Kolmogorov equations (2.17) permits to express the non-parametric estimate of the transition probabilities $\hat{P}^*(s, t)$ using the product-integral:

$$\hat{P}^*(s, t) = \prod_{(s, t]} (\mathbf{I} + d\hat{\Lambda}^*(u)). \quad (2.18)$$

where $\hat{\Lambda}^*(u)$ is the non-parametric estimate of the cumulative transition intensities at time u with $d\hat{\Lambda}_{hh}^*(u) \geq -1$ for all u , and \mathbf{I} is the $(M+1) \times (M+1)$ identity matrix. This estimated matrix of transition probabilities is called Aalen-Johansen estimator.

Let now $s < T_1^* < \dots < T_{m^*}^* \leq t$ be the ordered times of observed direct transitions between s and t for all the individuals. It can be deduced from (2.18):

$$\hat{P}^*(s, t) = \prod_{l=1}^{m^*} (\mathbf{I} + \Delta\hat{\Lambda}^*(T_l^*)), \quad (2.19)$$

where $\Delta\hat{\Lambda}^*(T_l^*) = \hat{\Lambda}^*(T_l^*) - \hat{\Lambda}^*(T_{l-1}^*)$ and $\Delta\hat{\Lambda}_{hh}^*(T_l^*) \geq -1$ for all T_l^* .

For example, using an unidirectional illness-death process (with possible transitions depicted in Figure 2.3),

$$\mathbf{I} + \Delta\hat{\Lambda}^*(T_l^*) = \begin{pmatrix} 1 - \frac{\Delta N_0(T_l^*)}{Y_0(T_l^*)} & \frac{\Delta N_{01}(T_l^*)}{Y_0(T_l^*)} \\ 0 & 1 - \frac{\Delta N_1(T_l^*)}{Y_1(T_l^*)} \\ 0 & 1 \end{pmatrix},$$

with $N_h(T_l^*) = \sum_{k \neq h} N_{hk}(T_l^*)$ and $\Delta N_{hk}(T_l^*) = N_{hk}(T_l^*) - N_{hk}(T_{l-1}^*)$.

Estimation of the non-parametric estimator covariance

The covariance matrix of the Aalen-Johansen estimator of transition probabilities can be estimated by the Greenwood-type estimator:

$$\widehat{\text{cov}}(\hat{P}^*(s, t)) = \int_s^t \hat{P}^*(u, t)^\top \otimes \hat{P}^*(s, u-) \widehat{\text{cov}}(d\hat{\Lambda}^*(u)) \hat{P}^*(u, t) \otimes \hat{P}^*(s, u-)^\top$$

where \otimes denotes the Kronecker product.

Andersen et al. [1993] simplified the computations in (2.20) using the recursion formula:

$$\begin{aligned} \widehat{\text{cov}}(\hat{P}^*(s, t)) &= \{(\mathbf{I} + \Delta\hat{\Lambda}^*(t))^\top \otimes \mathbf{I}\} \widehat{\text{cov}}(\hat{P}^*(s, t-)) \{(\mathbf{I} + \Delta\hat{\Lambda}^*(t)) \otimes \mathbf{I}\} + \\ &\quad \{\mathbf{I} \otimes \hat{P}^*(s, t-)\} \widehat{\text{cov}}(\Delta\hat{\Lambda}^*(t)) \{\mathbf{I} \otimes \hat{P}^*(s, t-)\}, \end{aligned}$$

where

$$\widehat{\text{cov}}(\Delta\widehat{\Lambda}_{hk}^*(t), \Delta\widehat{\Lambda}_{h'k'}^*(t)) = \begin{cases} \frac{(Y_h(t) - \Delta N_h(t))\Delta N_h(t)}{Y_h(t)^3}, & \text{for } h = k = h' = k', \\ -\frac{(Y_h(t) - \Delta N_h(t))\Delta N_{hk'}(t)}{Y_h(t)^3}, & \text{for } h = k = h' \neq k', \\ -\frac{(\delta_{kk'}Y_h(t) - \Delta N_{hk}(t))\Delta N_{hk'}(t)}{Y_h(t)^3}, & \text{for } h = h', h \neq k, h \neq k', \\ 0, & \text{for } h \neq h', \end{cases}$$

with $\delta_{kk'}$ the Kronecker delta. Note that $\widehat{\text{cov}}(\widehat{P}^*(s, t))$ and $\widehat{\text{cov}}(\Delta\widehat{\Lambda}^*(t))$ are two $(M + 1)^2 \times (M + 1)^2$ covariance matrices.

These results may be used to construct the 95% pointwise confidence intervals for the elements of the Aalen-Johansen estimator:

$$\exp \left\{ \log \widehat{P}_{hk}^*(s, t) \pm 1.96 \frac{\sqrt{\widehat{\text{var}}(\widehat{P}_{hk}^*(s, t))}}{\widehat{P}_{hk}^*(s, t)} \right\}.$$

(Semi-) Parametric estimator

In the multi-state framework, one can also model individual transition intensities of event $\lambda_{hk}^i(t)$ according to covariates, using for instance proportional intensity models:

$$\lambda_{hk}^i(t) = \lambda_{0,hk}(t) \exp\{X_{hk,i}^\top \gamma_{hk}\}, \quad (2.20)$$

with $(h, k) \in \mathcal{S}^2$ and $X_{hk,i}$ the vector of covariates measured at baseline associated with the intensity of transition $h \rightarrow k$ through the vector of parameters γ_{hk} . The baseline transition intensity $\lambda_{0,hk}(t)$ can be parametric or non-parametric, by extending the formulas developed in the survival context and by stratifying on each transition $h \rightarrow k$ [De Wreede et al., 2010]. Other authors proposed to approximate the baseline transitions by splines and use a penalised likelihood which penalises the second derivative of $\lambda_{hk}^i(t)$ using a loess parameter [Joly and Commenges, 1999] to avoid local variations and discontinuity in $\lambda_{0,hk}(t)$.

In the parametric setting, the parameters are then estimated in the maximum likelihood framework, while the semi-parametric setting generalises the Cox model estimation by stratifying on the transitions. Once the vector of estimated parameters $\widehat{\theta}$ obtained, we are able to estimate for each subject his individual transition probability $\widehat{P}^i(s, t|\widehat{\theta})$:

$$\begin{aligned} \widehat{P}_{hh}^i(s, t|\widehat{\theta}) &= \exp \left\{ \widehat{\Lambda}_{hh}^i(t|\widehat{\theta}) - \widehat{\Lambda}_{hh}^i(s|\widehat{\theta}) \right\}, \\ \widehat{P}_{hk}^i(s, t|\widehat{\theta}) &= \int_s^t \widehat{P}_{hh}^i(s, u|\widehat{\theta}) \widehat{\lambda}_{hk}^i(u|\widehat{\theta}) \widehat{P}_{kk}^i(u, t|\widehat{\theta}) du, h \neq k. \end{aligned} \quad (2.21)$$

In practice, when the state space S is large, these integrals become too complex numerically, and one can approximate $P^i(s, t|\hat{\theta})$ through the product-integral:

$$\hat{P}^i(s, t|\hat{\theta}) = \prod_{(s,t]} (\mathbb{I} + d\hat{\Lambda}^i(u)), \quad (2.22)$$

with $d\hat{\Lambda}_{hh}^i(u) \geq -1$ for all u .

Pseudo-value approach

Although most works in multi-state models rely on proportional intensity models, other types of models can also be used similarly as in survival analysis. For instance, the pseudo-value approach can be used to directly model transition probabilities. This is achieved by generalizing the survival analysis setting (see Section 2.2.1) which only focused on the modelling of cumulative probability functions for one type of event.

As an example, when interested in the transition probability $P_{hk}^i(s, t) = \Pr(E_i(t) = k | E_i(s) = h)$, the non-parametric estimator defined in the pseudo-value formulation (2.13) is obtained from the Aalen-Johansen estimator (2.18).

2.2.3 Landmark approach

We claimed earlier that the (parametric or semi-parametric) proportional hazards models assumed invariant effects over time, which is called the proportional hazards (PH) assumption. van Houwelingen [2014] argued that these models are not robust to the violation of the PH assumption, that is in presence of time-varying effects of covariates. To avoid biases and complex inclusion of time-varying effects, he proposed to use a simple solution: landmarking. The idea is to consider landmark times, that is a series of times after inclusion from which analyses are to be done, and fit a separate model at each landmark time point by selecting only the individuals still at risk at the landmark point. These models also assume PH but only after a certain landmark time. In addition, when the interest is in a short time frame as in prediction for instance, an administrative censoring may also be used to further reduce the possible bias due to the PH assumption. Finally, each landmark-specific effects can be combined to define time-dependent effects.

In the non-parametric estimators and the (semi-)parametric models we defined, another phenomenon well-discussed in the literature can arise and seriously trouble the results of the study: *the immortal time bias* [Sylvestre et al., 2006; Suissa, 2007; Giobbie-Hurder et al., 2013]. It occurs when, because of exposure definition, the outcome under

study could not occur. A famous example is the paper of Anderson et al. [1983], which debates on the effect of response to chemotherapy on survival. The naive way to run the analysis would be to consider two groups, a "responder" group and a "non-responder" group, and compare survival between the two groups. However, a potential responder cannot belong to the responder group if he dies before his time of response. Thus, the individuals in the responder group are considered as immortal for some time: immortal time bias. Actually, this example illustrates that one cannot make the analysis using groups based on their future composition, because the latter is unknown. Still, a possible alternative is to perform the analysis by only considering the present responses in subjects at risk at the landmark point within the so called landmark approach.

van Houwelingen [2007] explained that the same remark can be made in the presence of time-dependent covariates, when the clinical interest is on the prediction of survival given the covariate history (which may vary over time) up to a dynamic landmark point. A solution would be to adjust the survival model on predicted marker values at the prediction time from a separate mixed model [Tsiatis et al., 1995]. However, death stops the collection of the marker measurements and the further marker values certainly correspond to healthiest subjects so that immortal time bias is transposable to the case of longitudinal data. Since the linear mixed model based on ML (or REML) estimation does not take into account such informative missing data, a solution is to define a joint probability model for the covariate process and the survival process [Wulfsohn and Tsiatis, 1997], as further detailed in Section 2.3. However, such a modelling may lead to complex inference. An alternative [Albert and Chib, 2001] is to perform a sequential analysis at each event time (considered as a landmark time), by considering only the subjects still at risk and only their information available up to the landmark point.

2.2.4 Frailty

In presence of independent or clustered time-to-event data, the inclusion of frailty may be essential. The frailty is a random term which has a multiplicative effect on the hazard. It usually follows an exponential distribution such as a log-normal distribution or a Gamma distribution. When the frailty is included, one talks about frailty model. It may be included for several reasons [Hougaard, 1995; Nielsen et al., 1992].

When time-to-event data are clustered, as for example in the presence of members of a same family who share the same genetic background, or in the presence of groups

of patients treated in different hospitals, the inclusion of a frailty term may be essential [Liquet et al., 2012]. Indeed time-to-event data may remain dependent given the model's covariates. In that case, the frailty captures this residual correlation. One then talks about shared frailty model, because the elements of a same cluster share the same frailty. In this model, the event times are assumed independent conditionally to the frailty.

Clustered data also occur when we are interested in recurrent events in a same subject (here the cluster is the subject). In that case, the frailty captures the residual within-subject correlation, that is the correlation between the multiple event times of the same subject which is unexplained by the model's covariates. Such a model has been also extended to joint frailty models which take into account both recurrent event data and a terminal event [Liu et al., 2004; Rondeau et al., 2007]. In these models the frailty acts simultaneously as a multiplicative effect on both the hazard to experience the recurrent event and the hazard to experience the terminal event, and thus takes into account the residual correlation between the two processes.

Clustered data may also be considered in more general multi-state data, such as individual transition times. Putter and van Houwelingen [2011] argued that the inclusion of frailties in such models allows to capture both lack of fit of the model, heterogeneity in the data, and residual correlation in the transition times data. Thus, they claimed that the interpretation of the frailty effect is not always obvious. Finally, the authors stipulated that identifiability problems could occur in such models due to the small number of transitions.

Frailties have also been proposed in the extreme case of one unique (independent) event time [Hougaard, 1986], but in that case the identifiability is theoretically impossible without strong assumptions [Putter and van Houwelingen, 2011].

2.3 Standard joint modelling

The parametric and semi-parametric models developed in Section 2.2 have been introduced with covariates that did not vary with time. One could consider in addition the inclusion of time-dependent covariates. However, in these models the time varying covariates are assumed to be exogenous (or external).

The notion of external or exogenous versus internal or endogenous variable can be explained as follows [Kalbfleisch and Prentice, 2011]. Let us introduce times s and t such that $0 < s \leq t$, $Y(t)$ a time-dependent covariate observed at t , $\mathcal{Y}(s) = \{Y(u); 0 < u \leq s\}$

the history of the covariate until s and T the event time. A time-dependent covariate is exogenous when its distribution is not affected by the occurrence of the event, which can be expressed as

$$f_Y(Y(t)|\mathcal{Y}(s), T \geq s) = f_Y(Y(t)|\mathcal{Y}(s), T = s) \quad \text{or} \quad \lambda(s|\mathcal{Y}(s)) = \lambda(s|\mathcal{Y}(t)).$$

Exogenous time-dependent covariates are typically covariates not related with the subject, such as the season of the year, the air pollution, etc.

In contrast, endogenous time-dependent covariates are covariates that are related to the individual process, typically such as individual repeated measurements of a biomarker.

In addition to the issue of endogeneity, covariates included in standard event history models are assumed to be measured at each event time and without error, which is unrealistic for most measures.

For longitudinal marker analyses, we earlier mentioned that informative dropout may lead to incorrect estimates in mixed-effects models. To deal with MNAR, multiple approaches have been proposed in the literature such as pattern mixture models [Little, 1993], selection models [Diggle and Kenward, 1994] or joint models [Wulfsohn and Tsiatis, 1997]. Nowadays, pattern mixture models and selection models are used in practice for sensibility analysis. In this thesis, only the joint modelling approach, which considers the joint distribution of the longitudinal repeated measurements of the marker and the time-to-event data, is considered.

Whether for event history data analysis including time-dependent endogenous covariates, or for longitudinal data analysis with informative dropout, the joint modelling has considerably developed over the past years [Tsiatis and Davidian, 2004; Rizopoulos, 2010; Lawrence Gould et al., 2015]. Nowadays, it has become a key method for the analysis of longitudinal and event history correlated processes.

2.3.1 General structure of the joint modelling

The main idea of a joint model is to link two processes: a longitudinal process (biomarker measurements for example) and an event process (time-to-event data for example) using a shared latent structure, as illustrated in Figure 2.4.

In the literature, two kinds of joint models have been developed: joint models with shared latent classes and joint models with shared random effects. In the former the population is assumed to be heterogeneous and divided into homogeneous latent classes

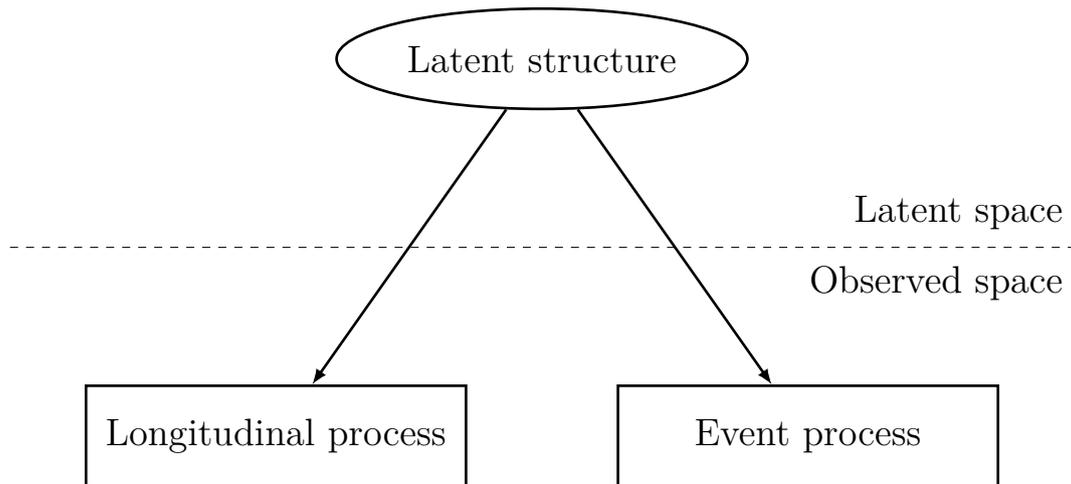


Figure 2.4 – Graphical general representation of a joint model for a longitudinal process and an event process. Ovals represent latent variables while squares represent observed variables

which are characterized by a specific mean trajectory of the longitudinal process and a specific risk of event [Proust-Lima et al., 2014]. In the latter the population is assumed to be homogeneous, and the link between the longitudinal and event processes is captured by a function of the individual random effects that also explain the variability in the longitudinal trajectories over subjects [Tsiatis and Davidian, 2004]. The latter is by far the most famous one in the statistical community, and we mainly focus on it in this thesis.

2.3.2 Standard joint model with shared random effects

The joint modelling was initially proposed in a two-stage estimation, which allowed to take into account the fact that the observations were intermittently observed and with error [Tsiatis et al., 1995]. However such a technique did not properly take into account the correlation between the two processes. To eliminate this bias, the joint model has been primarily developed by focusing on full joint likelihood approaches [Wulfsohn and Tsiatis, 1997].

Formulation

Let us consider the standard setting where the longitudinal marker follows a Gaussian distribution and the event history process is a survival process.

The classical joint model with shared random effects is decomposed into two sub-models: a linear mixed model for the repeated measurements of the biomarker Y_i and

a proportional hazards model for the individual hazard of event $\lambda_i(\cdot)$, both linked by a function of shared random effects b_i :

$$\begin{cases} Y_i(t) &= Y_i^*(t) + \epsilon_i(t) \\ &= X_i^L(t)^\top \beta + Z_i(t)^\top b_i + \epsilon_i(t) \\ \lambda_i(t) &= \lambda_0(t) \exp\{X_i^E{}^\top \gamma + W_i(t|b_i)^\top \eta\} \end{cases} \quad (2.23)$$

where $X_i^L(t)$ and $Z_i(t)$ are vectors of possibly time-dependent covariates associated with the p -vector of fixed effects β and the q -vector of individual random effects b_i , $b_i \sim \mathcal{N}(0, D)$. The error terms are such that $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{in_i})^\top \sim \mathcal{N}(0, \sigma^2 I_{n_i})$, where I is the identity matrix; ϵ_i and b_i are independent.

In the survival sub-part, $\lambda_0(t)$ is the baseline hazard. Most of the time, the baseline hazard is parametric (e.g., Weibull, piecewise constant or a small number of B-splines). It is rarely kept unspecified; indeed, the partial likelihood (like in the Cox model) cannot be employed and the full likelihood has to be defined. As a solution, one might consider a piecewise constant function with jumps at each event time but this would induce too many parameters and lead to computational problems [van der Vaart, 1998; Hsieh et al., 2006].

The vector of prognostic factors X_i^E is associated with the vector of coefficients γ while the multivariate function of marker dynamics $W_i(t|b_i)$ is associated with the vector of coefficients η . Usual specifications for $W_i(t|b_i)$ are the true current level of the marker $W_i(t|b_i) = Y_i^*(t)$ or the true current slope of the marker $W_i(t|b_i) = \partial Y_i^*(t)/\partial t$ or both $W_i(t|b_i) = (Y_i^*(t), \partial Y_i^*(t)/\partial t)^\top$ [Ye et al., 2008; Yu et al., 2008; Rizopoulos, 2012b]. But any function could be considered. For instance, authors introduced the area under the longitudinal trajectory up to time t , $W_i(t|b_i) = \int_0^t Y_i^*(u) du$ [Abrahamowicz et al., 2006; Brown, 2009; Rizopoulos, 2012b] while others considered directly the vector of random effects $W_i(t|b_i) = b_i$ [Follmann and Wu, 1995; Henderson et al., 2000; Vonesh et al., 2006].

Interpretation

This joint model formulation has the advantage to keep the same type of interpretation of the linear mixed-effects model for the longitudinal sub-part and of the proportional hazard model for the event sub-part as the separate models except that parameters γ are henceforth also adjusted on $W_i(t|b_i)$. Parameters η also use the proportional hazards assumption. For example, if we consider the more common case where the association function is on the true current level of the marker (i.e. $W_i(b_i, t) = Y_i^*(t)$), then $\exp(\eta)$ gives the hazard ratio for a 1-unit increase in the predictor $Y_i^*(t)$ at the current time, given covariates X_i^E .

Estimation

There exist different software to fit joint models with shared random effects [Rizopoulos, 2010; Philipson et al., 2012; Rizopoulos, 2016; Crowther et al., 2013; Zhang et al., 2016]. More and more are based on bayesian inference, mainly to avoid the possible complex inference in joint models that use maximum likelihood approach. We rely in this thesis on the R package **JM** [Rizopoulos, 2010] where estimation is done in the maximum likelihood framework.

In a joint model, the individual contribution to the likelihood exploits the independence between the time-to-event data (T_i, δ_i) and the longitudinal data Y_i , and the joint observed likelihood is

$$L(\theta) = \prod_{i=1}^N \int_{\mathbb{R}^q} f_Y(Y_i|b_i; \theta) f_E(T_i, \delta_i|b_i; \theta) f_b(b_i; \theta) db_i \quad (2.24)$$

with

$$f_Y(Y_i|b_i; \theta) = \frac{1}{(2\pi\sigma^2)^{n_i/2}} \exp\left(-\frac{\|Y_i - X_i^L\beta - Z_i b_i\|^2}{2\sigma^2}\right)$$

where $\|x\|$ denotes the Euclidean norm of vector x , X_i^L and Z_i are the matrices of covariates with row vectors $X_i^L(t_{ij})^\top$ and $Z_i(t_{ij})^\top$, $j = 1, \dots, n_i$, respectively;

$$f_E(T_i, \delta_i|b_i; \theta) = \lambda_i(t|b_i; \theta)^{\delta_i} \exp\left(-\int_0^{T_i} \lambda_i(u|b_i; \theta) du\right)$$

and

$$f_b(b_i; \theta) = \frac{1}{(2\pi)^{q/2} |D|^{1/2}} \exp\left(-\frac{b_i^\top D^{-1} b_i}{2}\right).$$

JM package maximizes the log-likelihood

$$l(\theta) = \sum_{i=1}^N \int_{\mathbb{R}^q} \{\log f_Y(Y_i|b_i; \theta) + \log f_E(T_i, \delta_i|b_i; \theta) + \log f_b(b_i; \theta)\} db_i,$$

and the optimization is performed using the expectation-maximization (EM) algorithm coupled with a quasi-Newton algorithm in case of slow convergence.

The log-likelihood of classical joint models with shared random effects involves integrals over time and random effects that can be achieved using numeric integration algorithms (Monte-Carlo or Gaussian quadratures for example). In **JM**, the integral over time is approximated using a Gauss-Kronrod quadrature and the computation of the integral

over the individual random effects is achieved using a Gauss-Hermite quadrature called pseudo-adaptive Gauss-Hermite quadrature.

The quadrature approximates the integral using a weighted sum of function values, at specified points within the domain of integration; the Gaussian quadrature is based on the use of polynomial functions. The idea of the adaptive quadrature is to recentre and rescale the integrand in the integral in order to obtain a better approximation of the integral [Lesaffre and Spiessens, 2001]. By using the same notations and explanations as the book of Rizopoulos [2012b], we call $A(\cdot)$ a function of the random effects, and define the adaptive Gauss-Hermite rule as:

$$\mathbb{E}\{A(\theta, b_i)|T_i, \delta_i, Y_i; \theta\} \approx 2^{q/2}|\widehat{B}_i|^{-1} \sum_{t_1=1}^K \dots \sum_{t_q=1}^K \pi_t A(\theta, \widehat{r}_t) f(\widehat{r}_t|T_i, \delta_i, Y_i, \theta) \exp(-\|b_t\|^2),$$

where K is the number of quadrature points and $\widehat{r}_t = \widehat{b}_i + \sqrt{2}\widehat{B}_i^{-1}b_t$ uses the prespecified abscissas $b_t = (b_{t_1}, \dots, b_{t_q})^\top$ with corresponding prespecified weights π_t . The empirical Bayes estimate \widehat{b}_i is deduced from the joint model with $\widehat{\theta}$ the corresponding estimated vector of parameters, using:

$$\widehat{b}_i = \arg \max_b \{\log L(\widehat{\theta})\} = \arg \max_b \{\log f(T_i, \delta_i, Y_i; \widehat{\theta})\},$$

and their covariance matrix \widehat{H}_i^{-1} is such as:

$$\widehat{H}_i = - \frac{\partial^2 \log f(T_i, \delta_i, Y_i, b; \widehat{\theta})}{\partial b \partial b^\top} \Big|_{b=\widehat{b}_i}.$$

We can then deduce \widehat{B}_i the Choleski factor of \widehat{H}_i .

However the location of the quadrature points using this adaptive Gauss-Hermite rule requires the computation of the mode and the variance of the posterior distribution of the random effects at each step of the optimisation algorithm. To avoid the computational cost of rescaling quadrature points at each iteration, Rizopoulos [2012b] proposed to use the pseudo-adaptive Gauss-Hermite rule which uses the estimates of the previously fitted linear mixed sub-model of the joint model (in the independence setting) to roughly recentre and rescale the subject-specific integrands. The formula is the same as previously, except that \widetilde{B}_i , \widetilde{r}_t and \widetilde{b}_i replace \widehat{B}_i , \widehat{r}_t and \widehat{b}_i , respectively. Henceforth, one has $\widetilde{r}_t = \widetilde{b}_i + \sqrt{2}\widetilde{B}_i^{-1}b_t$, and the empirical Bayes estimates \widetilde{b}_i are deduced from a first fitted linear mixed model with $\widetilde{\theta}_y$ the corresponding estimated vector of parameters, using:

$$\widetilde{b}_i = \arg \max_b \{\log f(Y_i, b; \widetilde{\theta}_y)\},$$

and their covariance matrix \widetilde{H}_i^{-1} is such as:

$$\begin{aligned}\widetilde{H}_i &= -\frac{\partial^2 \log f(Y_i, b; \tilde{\theta}_y)}{\partial b \partial b^\top} \Big|_{b=\tilde{b}_i} \\ &= \frac{Z_i^\top Z_i}{\tilde{\sigma}^2} + \widetilde{D}^{-1}.\end{aligned}$$

We can then deduce \widetilde{B}_i the Choleski factor of \widetilde{H}_i .

2.3.3 Joint model with discrete time-to-event data and exact likelihood inference

As mentioned earlier, the main difficulty in joint models using shared random effects comes from the computation of the likelihood that involves an integral over time and the aforementioned random effects. In practice, this complexity considerably restricts the possible choices of the association structure between the longitudinal and survival processes.

Barrett et al. [2015] proposed an elegant calculation trick which avoids to approximate the multidimensional integral with numeric methods such as Gaussian quadratures. Based on the results on truncated Gaussian distributions developed in Arnold [2009], the idea is to obtain a closed-form of the integral over the random effects, using the relation

$$\int \Phi^{(n)}(\nu_1 + \Omega z, \nu_2 + \Omega z; 0, V) \phi^{(q)}(z; 0, I_q) dz = \Phi^{(n)}(\nu_1, \nu_2; 0, V + \Omega \Omega^\top) \quad (2.25)$$

where $\Phi^{(n)}(a, b; 0, \Sigma)$ is a n -multivariate normal distribution function, expressed at vector b and truncated at vector a ($a \leq b$), with mean zero and variance Σ , and $\phi^{(q)}(b; 0, I_q)$ is a q -dimensional standard normal density function expressed at q -vector b ; ν_1, ν_2 are two vectors of size n ; Ω and V are matrices with dimensions (n, q) and (n, n) , respectively. Please note that in the initial paper of Arnold [2009], there is a small mistake since the relation $\Phi^{(n)}(a, b; 0, \Sigma) = \Phi^{(n)}(-\infty, b; 0, \Sigma) - \Phi^{(n)}(-\infty, a; 0, \Sigma)$ for any vectors a, b and matrix Σ is no more valid when $n > 1$.

To enter the Gaussian space and apply this formula, Barrett et al. [2015] use a discretization of the timescale for the time-to-event data. They replace the hazard function by the probability to have the event of interest in the time interval given that the subject was free of the event in the previous interval and model this probability by a sequential probit model. The integrand of the likelihood (2.24) thus becomes a product of Gaussian distribution function and density function, and formula (2.25) provides a close form likelihood.

2.3.4 Extensions of joint models

The joint models we introduced apply for repeated measurements of a Gaussian longitudinal marker and survival data, both correlated.

In the literature, many developments and applications exist concerning the extension of the classical joint model to multiple markers and multiple time-to-event data. For example, Chi and Ibrahim [2006] introduced joint models for multivariate longitudinal and multivariate survival data. Elashoff et al. [2008] proposed a joint model to deal with competing risks data, while Andrinopoulou et al. [2014] considered the competing risks setting in presence of two correlated longitudinal markers. Król et al. [2016] even introduced a complex trivariate joint model for left-censored longitudinal data, recurrent events and a terminal event. Finally, Dantan et al. [2011] introduced a joint model with latent state for longitudinal data and illness-death data.

However, none proposed a joint model for a longitudinal process and a multi-state process, although potentially essential in the full modelling of a disease process evolution.

3 Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer

Motivated by the (possibly successive) clinical transitions in prostate cancer with repeated measurements of PSA, we first developed a new joint model with shared random effects for a longitudinal process and a multi-state process.

This model allows to simultaneously analyse repeated measures of a Gaussian longitudinal marker and times of transitions between states characterized by a multi-state process. This is achieved by combining two modelling approaches: the multi-state modelling, based on the counting process theory [Andersen et al., 1993], and the standard joint modelling [Rizopoulos, 2010], which applies for a single event. The joint multi-state model is decomposed into two submodels: a linear mixed model for the Gaussian repeated measurements of the longitudinal marker and a multi-state model with proportional intensities (between each covariate level) for the individual transition times, both linked by a function of the individual shared random effects. The model thus allows not only to distinguish the type of event occurring for the subject, but also to depict the individual characteristics after the occurrence of the first event, by focusing on transition intensities between distinct states.

The observed likelihood is maximised to provide the estimates of the model parameters. This likelihood involves integrals over time and individual random effects which are approximated with numerical integration algorithms. Extensions of the previous solutions based on the use of pseudo-adaptive Gaussian quadratures [Rizopoulos, 2012a] are proposed to accurately and efficiently compute such quantities. The optimization is performed using an EM algorithm coupled to a quasi-Newton algorithm in case of slow convergence. The model estimation is implemented in R by combining and extending two packages of reference in the statistical community: `mstate` [De Wreede et al., 2010]

for preparing data in a multi-state setting, and JM [Rizopoulos, 2010] for estimating joint models with shared random effects for a Gaussian longitudinal marker and time-to-event data. The implementation is thus easy and effective, because it keeps the syntax and the features of JM while adapting to the multi-state framework.

A simulation study is carried out to check the performances of the estimation program over several scenarii of dimensions of the random effects and number of transitions.

The model is then applied on the two cohorts introduced in Chapter 1 and including men with a localised prostate cancer and treated by radiotherapy .

This work has been published in Statistics in Medicine [Ferrer et al., 2016].

3.1 Main article

Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer

Loïc Ferrer,^{a,*†} Virginie Rondeau,^a James Dignam,^b
Tom Pickles,^c H el ene Jacqmin-Gadda^a and C ecile Proust-Lima^a

Joint modelling of longitudinal and survival data is increasingly used in clinical trials on cancer. In prostate cancer for example, these models permit to account for the link between longitudinal measures of prostate-specific antigen (PSA) and time of clinical recurrence when studying the risk of relapse. In practice, multiple types of relapse may occur successively. Distinguishing these transitions between health states would allow to evaluate, for example, how PSA trajectory and classical covariates impact the risk of dying after a distant recurrence post-radiotherapy, or to predict the risk of one specific type of clinical recurrence post-radiotherapy, from the PSA history. In this context, we present a joint model for a longitudinal process and a multi-state process, which is divided into two sub-models: a linear mixed sub-model for longitudinal data and a multi-state sub-model with proportional hazards for transition times, both linked by a function of shared random effects. Parameters of this joint multi-state model are estimated within the maximum likelihood framework using an EM algorithm coupled with a quasi-Newton algorithm in case of slow convergence. It is implemented under R, by combining and extending `mstate` and `JM` packages. The estimation program is validated by simulations and applied on pooled data from two cohorts of men with localized prostate cancer. Thanks to the classical covariates available at baseline and the repeated PSA measurements, we are able to assess the biomarker's trajectory, define the risks of transitions between health states and quantify the impact of the PSA dynamics on each transition intensity. Copyright   2016 John Wiley & Sons, Ltd.

Keywords: joint modelling; longitudinal process; multi-state process; prostate cancer; R; shared random effects

1. Introduction

In longitudinal health studies, marker data are usually collected at repeated measurement times until the occurrence of an event such as disease relapse or death, with the objective to study the link between these two correlated processes or use the information brought by the marker's dynamics to explain or predict the time to event. In such analyses, the repeated measurements of the marker should not be considered as a standard time-dependent covariate in a survival model [1, 2] because the marker is an internal outcome measured with error and at discrete times whereas the Cox model assumes that the exact values of the explanatory variables are known for all the individuals at risk at each event time. To counteract these weaknesses, the two processes can be modelled jointly [3, 4]. The principle is to define two sub-models (one mixed sub-model for the longitudinal data and one survival sub-model for the time-to-event data) and use a common latent structure to link them. The shared random effect models, notably developed by Tsiatis and Davidian [5], are the most popular joint models. They usually assume that a function of the random effects from the linear mixed model is included as covariate in the survival model. This function can be any underlying features of the marker dynamics.

^aINSERM U1219, ISPED, Universit e de Bordeaux, Bordeaux, France

^bDepartment of Public Health Sciences, University of Chicago and NRG Oncology, Chicago, IL, U.S.A.

^cDepartment of Radiation Oncology, University of British Columbia, Vancouver, Canada

*Correspondence to: Lo ic Ferrer, INSERM U1219, ISPED, Universit e de Bordeaux, 146 rue L eo Saignat, 33076 Bordeaux Cedex, France.

†E-mail: loic.ferrer@inserm.fr

The joint modelling method is very useful in prostate cancer. The prostate-specific antigen (PSA), which is a protein secreted by the prostate, is found to be over-expressed in the presence of prostate cancer. This blood-based longitudinal tumour marker is commonly used by clinicians to monitor patients with localized prostate cancer following treatment (radiation therapy or surgery) in order to detect sub-clinical presence of disease. Proust-Lima *et al.* [6], Taylor *et al.* [7] and Yu *et al.* [8] showed, through various types of joint models, that the dynamics of this biomarker, along with the pre-treatment PSA level and other factors measuring the aggressiveness of cancer cells and the extent of the tumour, were risk factors for progression and permitted one to dynamically predict (i.e. using PSA to adapt prediction over time) the risk of clinical relapse.

In practice, a patient may experience a succession of clinical progression events with, for example, a local recurrence, followed by a distant metastatic recurrence and then death. So instead of the occurrence of a single clinical event, the progression of prostate cancer should be defined as a multi-state process with a focus on the transitions between clinical states and the impact of the biomarker dynamics on it. This is essential to understand and predict accurately the course of the disease, and it is of particular relevance for the clinicians that need to distinguish the different types of events in order to properly adapt the treatment.

Some authors already extended the classical joint modelling framework to multiple time-to-event data. Chi and Ibrahim [9] proposed a joint model for multivariate longitudinal data and multivariate survival data. Liu and Huang [10] and Kim *et al.* [11] looked into the simultaneous study of three correlated outcomes: longitudinal data, times of recurrent events and time of terminal event. Elashoff *et al.* [12] and Rizopoulos [13] extended the joint model to competing risks data, which allows to characterize the cause of survival event. Dantan *et al.* [14] developed a joint model with latent state for longitudinal data and illness-death data. Tom and Farewell [15] proposed a complex multi-state model that combined an intermittently observed longitudinal categorical process and a multi-state process. Recently, Andrinopoulou *et al.* [16] studied simultaneously two longitudinal markers and competing events. However, the joint study of Gaussian longitudinal data and multi-state data has never been proposed and implemented. Thus, we introduce a joint model with shared random effects for repeated measurements of a longitudinal marker and times of transitions between multiple states. It consists in a linear mixed model and a multi-state model with transition-specific proportional intensities, both linked by shared random effects.

The computational aspect is the main obstacle in the development of joint models with shared random effects. As explained by Gould *et al.* [17], the R package JM, developed by Rizopoulos [18], has enabled many advances in the use of joint modelling, particularly through efficient numerical integrations. On the other hand, the R package mstate, developed by De Wreede *et al.* [19], provides estimation of multi-state models. In the present work, we combine and adapt these two packages in order to estimate joint multi-state models. Thus, the implementation is easy and effective. Through the adaptation of `jointModel()` function of JM package, our approach uses the maximum likelihood approach, which is performed using the expectation-maximization (EM) algorithm coupled with a quasi-Newton algorithm in case of slow convergence. The software advantage is that it keeps the features, syntax and outputs of JM.

The paper is organized as follows. Section 2 presents the joint model for longitudinal and multi-state processes. Estimation and implementation procedures are detailed in Section 3 and validated by simulations in Section 4. The model is applied to two cohorts of men with prostate cancer in Section 5, and a brief discussion is finally given in Section 6.

2. Joint multi-state model

2.1. Notations

For each individual i , a longitudinal process and a multi-state process are observed. Let $\{E_i(t), t \geq 0\}$ be the multi-state process where $E_i(t)$ denotes the occupied state by subject i at time t and takes values in the finite state space $S = \{0, 1, \dots, M\}$. It is assumed that the multi-state process is continuous and observed between the left truncation time (time of entry in the study) T_{i0} and the right censoring time C_i , so that the observed process is $E_i = \{E_i(t), T_{i0} \leq t \leq C_i\}$. We further consider that E_i is a non-homogeneous Markov process. The Markov property ensures that the future of the process depends only on the present state and not on the past state, that is, $\Pr(E_i(t+u) = k | E_i(t) = h, \{E_i(s), s < t\}) = \Pr(E_i(t+u) = k | E_i(t) = h)$, $\forall h, k \in S, \forall u \geq 0$ [19], and the non-homogeneous property guarantees that the time since T_{i0} impacts the future evolution of the process. Let us consider $T_i = (T_{i1}, T_{i2}, \dots, T_{im_i})^\top$

the vector of the $m_i \geq 1$ observed time(s) for individual i , with $T_{ir} < T_{i(r+1)}, \forall r \in \{0, \dots, m_i - 1\}$, and where \top denotes the transpose operator. If the last observed state for subject i ($E_i(T_{im_i})$) is absorbing, that is, it is impossible to leave it once entered (typically death), we observe m_i direct transition(s). Otherwise, T_{im_i} equals C_i the right censoring time, and we observe $m_i - 1$ direct transition(s). We define by $\delta_i = (\delta_{i1}, \dots, \delta_{im_i})^\top$ the vector of observed transition indicator(s), with $\delta_{i(r+1)}$ equals 1 if a direct transition is observed at time $T_{i(r+1)}$ (i.e. $E_i(T_{ir}) \neq E_i(T_{i(r+1)})$) and 0 otherwise, $\forall r \in \{0, \dots, m_i - 1\}$. For each subject i , we also observe $Y_i = (Y_{i1}, \dots, Y_{in_i})^\top$ the vector of n_i measures of the marker collected at times t_{i1}, \dots, t_{in_i} , with $t_{in_i} \leq T_{im_i}$.

2.2. Joint multi-state model formulation

The joint multi-state model is decomposed into two sub-models: a linear mixed sub-model for the longitudinal data (repeated measurements of the biomarker) and a multi-state model with transition-specific proportional intensities for the event history data (transition and censoring times), both linked by a function of the shared random effects.

2.2.1. Longitudinal sub-model. To model the trajectory of the longitudinal marker, we use a linear mixed model. Under Gaussian assumptions, we assume that Y_{ij} the observed measure of the marker at time point t_{ij} is a noisy measure of the true level $Y_i^*(t_{ij})$. This non-observed level $Y_i^*(t_{ij})$ is explained according to time and covariates with fixed effects β at the population level, and random effects b_i that take into account the correlation between repeated measures of the same individual:

$$\begin{aligned} Y_{ij} &= Y_i^*(t_{ij}) + \epsilon_{ij} \\ &= X_i^L(t_{ij})^\top \beta + Z_i(t_{ij})^\top b_i + \epsilon_{ij}, \end{aligned} \tag{1}$$

with $X_i^L(t_{ij})$ and $Z_i(t_{ij})$ the vectors of possibly time-dependent covariates associated with the p -vector of fixed effects β and the q -vector of random effects b_i , $b_i \sim \mathcal{N}(0, D)$, respectively. Note that $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{in_i})^\top \sim \mathcal{N}(0, \sigma^2 I_{n_i})$ where I is the identity matrix; ϵ_i and b_i are independent.

2.2.2. Multi-state sub-model. To model the transition times, we use a Markov multi-state model with proportional hazards that takes into account the marker's dynamics through the shared random effects b_i . Thus, for the transition from state $h \in S$ to state $k \in S$, the transition intensity at time t takes the following form:

$$\begin{aligned} \lambda_{hk}^i(t|b_i) &= \lim_{dt \rightarrow 0} \frac{\Pr(E_i(t+dt) = k | E_i(t) = h; b_i)}{dt} \\ &= \lambda_{hk,0}(t) \exp \left\{ X_{hk,i}^S \top \gamma_{hk} + W_{hk,i}(b_i, t)^\top \eta_{hk} \right\}, \end{aligned} \tag{2}$$

with $\lambda_{hk,0}(\cdot)$ the parametric baseline intensity (Weibull, piecewise constant or B-splines for example) and $X_{hk,i}^S$ the vector of prognostic factors associated with the r -vector of coefficients γ_{hk} . The multivariate function $W_{hk,i}(b_i, t)$ defines the dependence structure between the longitudinal and multi-state processes. We can choose $W_{hk,i}(b_i, t) = Y_i^*(t)$ (the true current level of the marker), or $W_{hk,i}(b_i, t) = \partial Y_i^*(t) / \partial t$ (the true current slope), $W_{hk,i}(b_i, t) = (Y_i^*(t), \partial Y_i^*(t) / \partial t)^\top$ (both), or any other function of the random effects in the context under study. Thus, the s -vector of coefficients η_{hk} quantifies the impact of the longitudinal marker's dynamics on the transition intensity between the states h and k .

3. Estimation

3.1. Likelihood

The parameters of this joint model are estimated in the maximum likelihood framework. Because the longitudinal and multi-state processes are independent conditionally on the random effects, the complete observed likelihood is obtained through the product of the individual contributions to the likelihood for the N individuals as follows:

$$L(\theta) = \prod_{i=1}^N \int_{\mathbb{R}^q} f_Y(Y_i|b_i; \theta) f_E(E_i|b_i; \theta) f_b(b_i; \theta) db_i, \quad (3)$$

where θ is the vector of all the parameters contained in (1) and (2), and $f(\cdot)$ is a probability density function.

In the longitudinal part, described by the linear mixed model (1), the conditional longitudinal outcomes are such that

$$f_Y(Y_i|b_i; \theta) = \frac{1}{(2\pi\sigma^2)^{n_i/2}} \exp\left(-\frac{\|Y_i - X_i^L \beta - Z_i b_i\|^2}{2\sigma^2}\right), \quad (4)$$

where $\|x\|$ denotes the Euclidean norm of vector x , X_i^L is the matrix of covariates with row vectors $X_i^L(t_{ij})^T$, $j = 1, \dots, n_i$, and likewise $Z_i = \{Z_i(t_{ij})\}$.

For the multi-state part, let $P_{hk}^i(s, t)$ be the transition probability from state h to state k between times s and t for individual i , that is, $P_{hk}^i(s, t) = \Pr(E_i(t) = k | E_i(s) = h)$. For each $r \in \{0, \dots, m_i - 1\}$, the continuity and Markov assumptions imply that individual i remains in state $E_i(T_{ir})$ between times T_{ir} and $T_{i(r+1)}$ with probability $P_{E_i(T_{ir}), E_i(T_{ir})}^i(T_{ir}, T_{i(r+1)} | b_i)$, and transits to state $E_i(T_{i(r+1)})$ with intensity $\lambda_{E_i(T_{ir}), E_i(T_{i(r+1)})}^i(T_{i(r+1)} | b_i)$ when $T_{i(r+1)}$ is an observed transition time. By conditioning on $E_i(T_{i0})$, this translates in the individual contribution to the likelihood:

$$\begin{aligned} f_E(E_i|b_i; \theta) &= \prod_{r=0}^{m_i-1} \left\{ P_{E_i(T_{ir}), E_i(T_{ir})}^i(T_{ir}, T_{i(r+1)} | b_i) \lambda_{E_i(T_{ir}), E_i(T_{i(r+1)})}^i(T_{i(r+1)} | b_i)^{\delta_{i(r+1)}} \right\} \\ &= \prod_{r=0}^{m_i-1} \left\{ \exp\left(\int_{T_{ir}}^{T_{i(r+1)}} \lambda_{E_i(T_{ir}), E_i(T_{ir})}^i(u | b_i) du\right) \lambda_{E_i(T_{ir}), E_i(T_{i(r+1)})}^i(T_{i(r+1)} | b_i)^{\delta_{i(r+1)}} \right\} \end{aligned} \quad (5)$$

with $\lambda_{hh}^i(t) = -\sum_{k, k \neq h} \lambda_{hk}^i(t)$. The possible delayed entry is accounted for by conditioning on $E_i(T_{i0})$.

Finally, the random effects b_i follow a multivariate Gaussian distribution such that

$$f_b(b_i; \theta) = \frac{1}{(2\pi)^{q/2} \det(D)^{1/2}} \exp\left(-\frac{b_i^T D^{-1} b_i}{2}\right). \quad (6)$$

3.2. Implementation

The joint multi-state model has been implemented under R, via the combination of two well-known packages: `mstate` for multi-state models and `JM` for joint models with shared random effects. To fit semi-parametric Markov multi-state models, `mstate` prepares the database for multi-state analysis, more specifically by defining each patient's history as a series of rows, one for each transition at risk for each individual (in contrast with only one data record (row) per individual in a classical survival analysis). By stratifying on the transition type, the standard `coxph()` function of the R package `survival` can then be used to fit transition-specific Cox models. With standard longitudinal and time-to-event data, `JM` package initializes the values of the parameters with function `lme()` (`nlme` package) for the longitudinal sub-model and `coxph()` (`survival` package) for the survival sub-model. Then, function `jointModel()` carries out the estimation procedure.

So by replacing the standard call to `coxph()` by the call to `coxph()` on the data prepared with `mstate`, an extended `jointModel()` function, called `JMstateModel()`, can carry out the estimation procedure of the joint model for longitudinal and multi-state data. The implementation procedure thus includes four steps:

- `lme()` function (`nlme` package) to initialize the parameters of the longitudinal sub-model;
- `mstate` package functions `mstate` and `expand.covs()` to prepare the multi-state data;
- `coxph()` function (`survival` package) applied to the prepared data to initialize the parameters of the multi-state sub-model; and
- `JMstateModel()` function to estimate all the parameters of the joint multi-state model.

A detailed example is given in Web Appendix A, and full detailed examples are available on <https://github.com/LoicFerrer/JMstateModel/>.

3.3. Algorithm

`JMstateModel()` function computes and maximizes the joint log-likelihood extended to handle multi-state data using integration and optimization algorithms available in `JM` package. Thus, the procedure combines an EM algorithm coupled with a quasi-Newton algorithm if the convergence is not achieved. Furthermore, the integral with respect to time in (5) and the integral with respect to the random effects in (3) do not have an analytical solution. These integrals are approached by numerical integration. The integrals over time are approximated using Gauss–Kronrod quadratures, and the integral over the random effects using pseudo-adaptive Gauss–Hermite quadratures. Inference is provided by asymptotic properties for maximum likelihood estimators. The variance–covariance matrix of the parameter estimates is based on the inverse of the Hessian matrix. Details on the optimization procedure, the EM algorithm and the numerical integrations can be found in Rizopoulos [13].

The main difficulty with the inference comes from the numerical approximation of the integral over the random effects, especially when the dimension of the random effects increases. The pseudo-adaptive Gauss–Hermite quadrature proposed by Rizopoulos [20] centers the integral using the posterior distribution of the random effects, derived from the initial linear mixed model. This reduces the required number of quadrature points compared with the standard Gaussian quadrature and avoids the intensive computations of the adaptive quadrature. We went one step further by repeating this procedure: the joint model can be estimated once using the pseudo-adaptive technique, and it can then be reestimated by starting from the previously estimated parameters and centering the integral on the predicted random effects derived from the joint model rather than on the linear mixed model. We expect the integral to be more accurate while using a relatively small number of quadrature points. In the remainder, the technique is referred to as the multi-step pseudo-adaptive Gauss–Hermite rule. More details are in Web Appendix B.

4. Simulation study

The estimation procedure was validated in a simulation study.

4.1. Data generation

In one specific replicate, the longitudinal and multi-state data were generated for each subject $i = 1, \dots, 500$, according to the joint multi-state model defined as follows:

$$\left\{ \begin{array}{l} Y_{ij} = Y_i^*(t_{ij}) + \epsilon_{ij} \\ \quad = (\beta_0 + \beta_{0,X}X_i + b_{i0}) + \\ \quad (\beta_1 + \beta_{1,X}X_i + b_{i1}) \times ((1 + t_{ij})^{-1.2} - 1) + \\ \quad (\beta_2 + \beta_{2,X}X_i + b_{i2}) \times t_{ij} + \epsilon_{ij}, \\ \lambda_{hk}^i(t|b_i) = \lambda_{hk,0}(t) \exp \{ \gamma_{hk}X_i + \eta_{hk,level} Y_i^*(t) + \eta_{hk,slope} \partial Y_i^*(t) / \partial t \}, \end{array} \right. \quad (7)$$

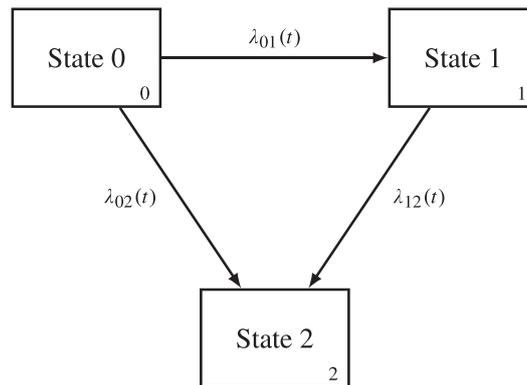
where the multi-state process that included three states ($h, k \in \{0, 1, 2\}$) and three transitions is described in Figure 1.

The same shape of trajectory as in the application was assumed with $((1 + t)^{-1.2} - 1)$ for a short-term drop and t for a long-term linear trend. First, X_i and $b_i = (b_{i0}, b_{i1}, b_{i2})^\top$ were generated according to normal distributions with mean 2.04 and variance 0.5, and mean vector

$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ and variance-covariance matrix

$\begin{pmatrix} 0.363 & 0.345 & 0.011 \\ 0.345 & 1.742 & 0.310 \\ 0.011 & 0.310 & 0.173 \end{pmatrix}$, respectively. The times of measurements were $t_{ij} = 0, 0.33, 0.67, \dots, 16.33$, and

ϵ_{ij} was generated from a normal distribution with mean zero and variance 0.074. The log baseline intensities were linear combinations of cubic B-splines with the same knot vector $(0.004, 7.458, 18.201)^\top$ for the three transitions, and the vectors of spline coefficients $(-5.537, -4.373, -4.541, -7.524, -5.205)^\top$ for transition $0 \rightarrow 1$, $(-5.231, -4.122, -3.815, -1.495, -0.887)^\top$ for transition $0 \rightarrow 2$ and $(-2.157, -2.491, -2.175, -0.975, -0.472)^\top$ for transition $1 \rightarrow 2$. Parameters values and knot locations were chosen according to the application data described in Section 5.



$$\bar{\psi}_{\text{sim}} = \begin{pmatrix} 198 & 149 & 153 \\ 0 & 56 & 93 \\ 0 & 0 & 246 \end{pmatrix}$$

Figure 1. Simulated multi-state process. Arrows indicate the directions of the possible transitions. $\lambda_{hk}(t)$ characterizes the intensity of transition between states h and k at time t . The matrix $\bar{\psi}_{\text{sim}}$ has size $(3, 3)$ and is composed of elements $\bar{\psi}_{\text{sim},(h+1)(k+1)}$, $h, k \in \{0, 1, 2\}$, where $\bar{\psi}_{\text{sim},(h+1)(k+1)}$ is the average number of observed direct transitions $h \rightarrow k$ over the 500 replicates. The diagonal elements $\bar{\psi}_{\text{sim},(h+1)(h+1)}$ denote the average number of patients who were censored in state h . Note that the sum of elements of a row $(h + 1)$ of $\bar{\psi}_{\text{sim}}$ corresponds to the average number of patients who experienced the state h .

The procedure described in Beyersmann *et al.* [21] and Crowther and Lambert [22] was used to generate the vector of observed times $T_i = (T_{i1}, \dots, T_{im_i})^\top$. For each individual i , the censoring time C_i was generated from a uniform distribution on $[1, 25]$, and the vector of true transition times $T_i^* = (T_{i,01}^*, T_{i,02}^*, T_{i,12}^*)^\top$ was generated according to the following procedure: (i) three random numbers $u_{i,01}$, $u_{i,02}$ and $u_{i,12}$ were generated from three independent standard uniform distributions; (ii) $T_{i,01}^*$ and $T_{i,02}^*$ were generated by solving $\int_0^{T_{i,0k}^*} \lambda_{0k}^i(v_{0k}|b_i) dv_{0k} + \log(u_{i,0k}) = 0$, for $k = 1, 2$, through the Brent's univariate root-finding method [23]; (iii) then, the true transition time $T_{i,12}^*$ was generated by solving $\int_{T_{i,01}^*}^{T_{i,12}^*} \lambda_{12}^i(v_{12}|b_i) dv_{12} + \log(u_{i,12}) = 0$. Finally, by comparing T_i^* and C_i , the vector T_i , which characterizes the multi-state process, was deduced.

The longitudinal measurements, generated from the linear mixed sub-model, were truncated at T_{i1} the first observed time of the multi-state process.

4.2. Estimated model

The model defined in (7) was used for the estimation with $b_i \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D_{11} & D_{12} & D_{13} \\ D_{12} & D_{22} & D_{23} \\ D_{13} & D_{23} & D_{33} \end{pmatrix}\right)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. The log baseline intensities were approximated by a linear combination of cubic-splines with one internal knot placed at the median of the observed transition times.

4.3. Simulation results

The simulations results were obtained through 500 replicates of 500 individuals. Each joint multi-state model was estimated using 3 and 9 pseudo-adaptive Gauss–Hermite quadrature points and a two-step pseudo-adaptive Gauss–Hermite quadrature using 9 quadrature points at each step. The simulation results are presented in Table I.

These results were very satisfying with unbiased estimates and correct 95% coverage rates. They showed, however, the need to use a certain number of Gauss–Hermite quadrature points to approximate the integral over the random effects. Indeed, the use of 3 Gauss–Hermite quadrature points using the pseudo-adaptive Gauss–Hermite rule induced poor coverage rates of the parameters associated with the long time effect in the longitudinal sub-part. The underestimation of the variance parameters was almost

Table 1. Simulation results according to 3 and 9 quadrature points using the pseudo-adaptive Gauss-Hermite rule (called one-step), and 9–9 quadrature points using the two-step adaptive Gauss-Hermite rule (called two-step). For each scenario, the statistics are (from left to right): mean, mean standard error, standard deviation, relative bias (in percentage) and coverage rate (in percentage).

True value	3 Gauss-Hermite quadrature points (one-step)					9 Gauss-Hermite quadrature points (one-step)					9–9 Gauss-Hermite quadrature points (two-step)				
	Mean	StdErr	StdDev	Rel. bias	Cov. rate	Mean	StdErr	StdDev	Rel. bias	Cov. rate	Mean	StdErr	StdDev	Rel. bias	Cov. rate
<i>Longitudinal process</i>															
β_0	-0.254	0.087	0.091	-0.5	95.4	-0.253	0.088	0.091	-0.7	95.4	-0.252	0.088	0.091	-1.0	95.6
$\beta_{0,X}$	0.797	0.040	0.043	-0.2	94.2	0.797	0.040	0.043	-0.2	94.6	0.797	0.041	0.043	-0.2	95.0
β_1	0.950	0.175	0.198	0.2	91.0	0.951	0.189	0.197	0.3	94.4	0.954	0.196	0.196	0.6	95.4
$\beta_{1,X}$	0.894	0.081	0.093	-1.2	90.4	0.902	0.087	0.093	-0.3	92.8	0.903	0.091	0.092	-0.2	94.4
β_2	-0.088	0.022	0.059	-3.7	56.0	-0.081	0.045	0.060	-7.6	83.2	-0.084	0.059	0.060	-5.0	95.2
$\beta_{2,X}$	0.198	0.010	0.028	-4.2	49.0	0.202	0.018	0.028	-2.6	79.6	0.204	0.027	0.028	-1.4	93.0
$\log(\sigma)$	-1.300	0.007	0.008	0.0	93.2	-1.300	0.007	0.008	0.0	93.6	-1.300	0.007	0.008	0.0	93.6
<i>Multi-state process</i>															
$\gamma_{01,X}$	0.199	0.130	0.132	1.2	94.6	0.199	0.130	0.133	0.9	94.0	0.199	0.130	0.132	0.7	94.4
$\gamma_{02,X}$	0.182	0.124	0.119	7.5	95.8	0.182	0.124	0.119	7.4	95.4	0.182	0.124	0.119	7.3	95.8
$\gamma_{12,X}$	-0.232	0.168	0.187	-4.7	91.8	-0.235	0.168	0.188	-3.5	91.4	-0.234	0.168	0.188	-3.6	92.4
$\eta_{01,level1}$	0.419	0.097	0.100	0.9	93.6	0.419	0.097	0.100	-0.1	93.2	0.418	0.097	0.100	-0.4	93.2
$\eta_{02,level1}$	-0.091	0.052	0.056	7.1	93.2	-0.099	0.052	0.056	8.3	93.8	-0.099	0.052	0.056	8.5	93.4
$\eta_{12,level1}$	0.046	0.088	0.092	16.5	93.8	0.052	0.087	0.091	12.2	94.4	0.052	0.087	0.092	12.3	94.2
$\eta_{01,slope}$	2.919	0.453	0.458	-0.3	94.6	2.940	0.455	0.460	0.7	94.6	2.952	0.455	0.464	1.1	94.6
$\eta_{02,slope}$	1.142	0.457	0.458	3.6	94.0	1.197	0.458	0.461	4.8	93.0	1.199	0.458	0.461	4.9	93.8
$\eta_{12,slope}$	0.134	0.836	0.857	-32.8	95.0	0.109	0.834	0.851	-18.4	95.0	0.113	0.834	0.852	-15.4	94.8
<i>Random effects</i>															
D_{11}	0.360	0.026	0.025	-0.8	95.0	0.360	0.026	0.025	-0.8	95.2	0.360	0.026	0.025	-0.8	95.0
D_{12}	0.342	0.046	0.045	-1.0	95.4	0.342	0.046	0.045	-0.9	94.8	0.342	0.046	0.045	-0.9	95.0
D_{13}	0.011	0.013	0.013	-0.8	95.2	0.011	0.013	0.013	-0.7	95.0	0.011	0.013	0.013	-0.3	95.4
D_{22}	1.742	0.131	0.133	-0.8	94.4	1.732	0.132	0.133	-0.6	94.2	1.732	0.132	0.133	-0.6	94.2
D_{23}	0.310	0.033	0.033	-0.7	94.4	0.309	0.033	0.033	-0.4	94.2	0.309	0.034	0.033	-0.3	94.4
D_{33}	0.173	0.012	0.013	-1.0	93.0	0.171	0.013	0.013	-0.7	93.8	0.172	0.013	0.013	-0.6	94.2

corrected using 9 quadrature points in the pseudo-adaptive Gauss–Hermite rule, and finally, the estimated Hessian was good with 9 and 9 quadrature points using the two-step pseudo-adaptive Gauss–Hermite rule. Overall, these results confirmed the good performances of the implemented procedure. To further investigate whether the technique could be applied to more complex multi-state data, we ran another simulation study with five states and 10 transitions as in the application. In this second simulation, the longitudinal part was simplified by assuming a linear trajectory over time. Although some direct transitions did not have a lot of information, the coverage rates of the multi-state model parameters were good. Results are detailed in Web Appendix C.

5. Application

We analysed data from patients with a localized prostate cancer treated by external beam radiotherapy. The analysis aimed to explore the link between PSA dynamics and transition intensities between clinical states, as well as to describe PSA repeated measurements and times of transitions between health states.

5.1. Data description

Our study focuses on 1474 men with a clinically localized prostate cancer and treated by external beam radiotherapy (EBRT): 629 patients come from the multi-center clinical trial RTOG 9406 (Radiation Therapy Oncology Group, USA) in which data collection has been conducted from 1994 to 2013 [24], and 845 patients come from the cohort of the British Columbia Cancer Agency (BCCA) in Vancouver, Canada [25] with examinations between 1994 and 2012 (Table II). During his follow-up, a patient can possibly go through several states defined as local recurrence, distant recurrence, initiation of hormonal therapy (HT) and death, due or not to prostate cancer. The initiation of salvage HT, which is an additional treatment prompted by physician observed signs in PSA or clinical signs, is designed to prevent growth of potentially present sub-clinical cancer. This intervention is not planned at diagnosis or initiated by any precise rule but is rather based on a mutual agreement between the clinician and his patient. Thus, it is treated as a disease state transition representing failure of the initial treatment to satisfactorily control the disease. Furthermore, as recommended in Proust-Lima *et al.* [6], we only considered the local relapses that took place 3 years or later after radiation, or within 3 years of EBRT when the last PSA value was >2 ng/mL. PSA data were collected at regular visits, for a median number of 10 PSA measurements per patient. Note that PSA data were collected between the end of EBRT and the occurrence of the first event (first

Table II. Description of the two cohorts.			
Cohort	RTOG 9406	BCCA	Pooled
Study period	1994–2013	1994–2012	
Number of patients	629	845	1474
Number of PSA measures per patient	13 (4, 23)	9 (3, 15)	10 (3, 21)
iPSA*	2.0 (1.0, 3.0)	2.1 (0.6, 3.3)	2.1 (0.8, 3.1)
Clinical T-stage			
1	355 (56.4%)	184 (21.8%)	539 (36.6%)
2	261 (41.5%)	514 (60.8%)	775 (52.6%)
3–4	13 (2.1%)	147 (17.4%)	160 (10.9%)
Gleason score			
2–6	424 (67.4%)	605 (71.6%)	1029 (69.8%)
7	167 (26.6%)	189 (22.4%)	356 (24.2%)
8–10	38 (6.0%)	51 (6.0%)	89 (6.0%)
Mean time of first event [†]	9.8 (2.3, 15.9)	7.7 (1.9, 14.1)	8.2 (2.0, 15.0)
Mean time of last contact [‡]	11.6 (2.9, 16.7)	9.0 (3.4, 14.8)	9.7 (3.1, 15.9)

Continuous data: Median (5th and 95th percentiles).

Categorical data: Amount (percentage).

Times are in years since the end of external beam radiotherapy.

* Pre-therapy PSA value (ng/ml) in the $\log(. + 0.1)$ scale.

[†] Minimum between the time of first transition and the time of censoring.

[‡] Minimum between the time of death and the time of censoring.

BCCA, British Columbia Cancer Agency; PSA, prostate-specific antigen; RTOG, Radiation Therapy Oncology Group.

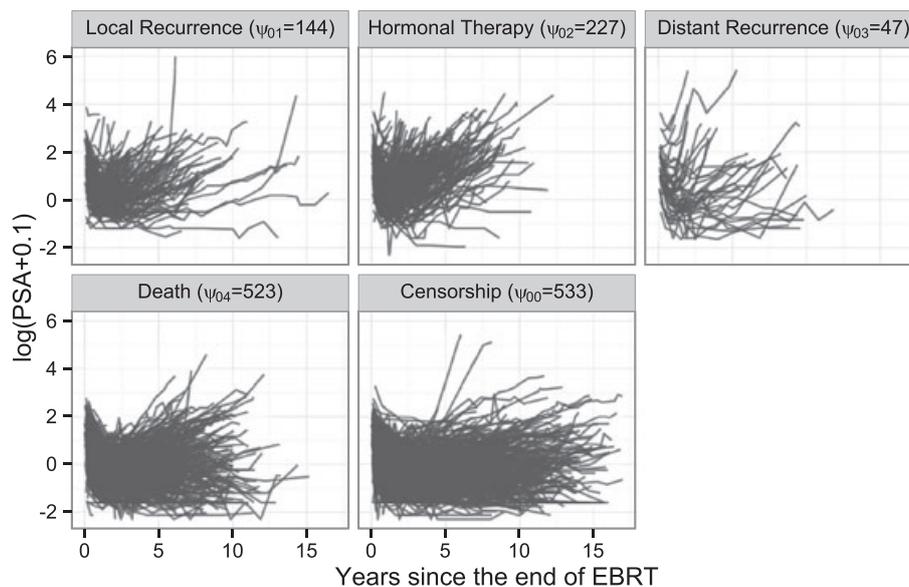


Figure 2. Individual trajectories of $\log(\text{PSA} + 0.1)$ after the end of EBRT and according to the type of first relapse in the two cohorts ($N = 1474$). ψ_{0k} is detailed in Figure 3.

clinical recurrence, HT, death or censorship). Subjects with only one PSA measure were excluded, and subjects who had an event in the first year after EBRT were excluded to prevent the inclusion of patients with substantial residual initial tumors. As shown in Table II, three baseline factors were considered: the pre-therapy level of PSA in the log scale (iPSA), the T-stage category, which characterizes the tumour size (three categories were considered: 2; 3–4 versus 1 in reference), and the Gleason score category, which measures the aggressiveness of cancer cells (three categories: 7; 8–10 versus 2–6 in reference). In the models, a cohort covariate was also considered coded as 1 for RTOG 9406 and -1 for BCCA.

The PSA individual trajectories collected between the end of EBRT and the occurrence of the first event are depicted in Figure 2. Overall, this longitudinal process is biphasic, with a decrease in the level of PSA in the first years following the end of EBRT, and a subsequent stabilization or linear rise thereafter. According to the type of first relapse, the biomarker's long-term increase may have different intensities (see 'Hormonal Therapy' and 'Censorship' for example).

The multi-state data are depicted through the transitions between the five states and the corresponding amount of observed direct transitions in Figure 3.

From the end of EBRT (state 0), a patient can experience either a transition to a localized recurrence (state 1), an HT (state 2), a distant recurrence (state 3) or death (absorbing state 4). After a localized recurrence (state 1), a patient may initiate an HT (state 2) or experience either a distant recurrence (state 3) or die (state 4). After initiation of HT, a patient may only experience a distant recurrence or die, and finally, after a distant recurrence, a patient may only die. In total, 144 subjects had a local recurrence; 317 men initiated an HT including 90 after a local recurrence; 90 men had a distant recurrence including 10 directly after a local recurrence and 33 after an HT initiation. In total, 802 patients died including 523 who did not have another recorded progression of the cancer before. Among the 672 men who were censored during the follow-up, 533 were censored before experiencing any clinical progression.

5.2. Specification of the joint model

The joint multi-state model being a complex model, a step-by-step procedure was carried out to specify the joint model. The specifications of the longitudinal and multi-state sub-models were based on two separate analyses, that is assuming independence between the two processes. Covariate selection was made using univariate or multivariate Wald tests.

5.2.1. Longitudinal sub-model specification. The biphasic shape of log-PSA was described in a linear mixed model with two functions of time according to previous works [6]: $f_1(t) = (1 + t)^\alpha - 1$ and $f_2(t) = (t)^{1+\nu}/(1 + t)^\nu$, where α and ν were estimated by profile likelihood ($\alpha = -1.2, \nu = 0$). Thus, these two functions depicted the short-term drop in the level of log-PSA after EBRT and the long-term linear

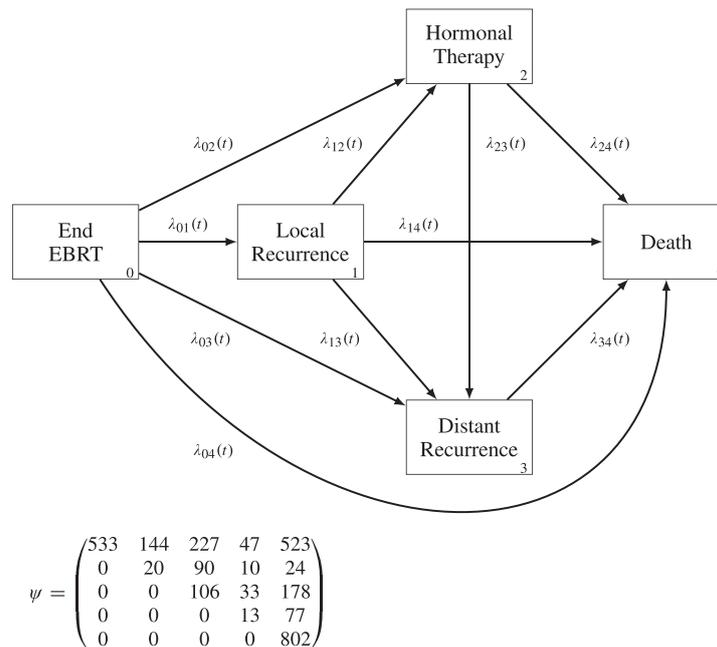


Figure 3. Multi-state representation of the clinical progressions in prostate cancer. Arrows indicate the directions of the possible transitions ($N = 1474$). $\lambda_{hk}(t)$ characterizes the intensity of transition between states h and k at time t . Matrix ψ has size $(5, 5)$ and is composed of elements $\psi_{(h+1)(k+1)}$, $h, k \in \{0, 1, 2, 3, 4\}$, where $\psi_{(h+1)(k+1)}$ is the number of observed direct transitions $h \rightarrow k$. Diagonal elements $\psi_{(h+1)(h+1)}$ denote the number of patients who were censored in state h . Note that the sum of elements of one row ($h + 1$) of ψ corresponds to the number of patients who entered state h .

increase of log-PSA, respectively. By denoting $Y_{ij} = \log(\text{PSA}_i(t_{ij}) + 0.1)$ the log-measure of PSA for the individual i at time t_{ij} – the natural logarithm transformation is performed to obtain a Gaussian shape for the longitudinal response – the linear mixed sub-model took the following form:

$$\begin{aligned} Y_{ij} &= Y_i^*(t_{ij}) + \epsilon_{ij} \\ &= (\beta_0 + X_i^{L0 \top} \beta_{0,\text{cov}} + b_{i0}) + \\ &\quad (\beta_1 + X_i^{L1 \top} \beta_{1,\text{cov}} + b_{i1}) \times f_1(t_{ij}) + \\ &\quad (\beta_2 + X_i^{L2 \top} \beta_{2,\text{cov}} + b_{i2}) \times f_2(t_{ij}) + \epsilon_{ij}, \end{aligned}$$

with $b_i = (b_{i0}, b_{i1}, b_{i2})^\top \sim \mathcal{N}(0, D)$, D unstructured, and $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{in_i})^\top \sim \mathcal{N}(0, \sigma^2 I_{n_i})$. The covariates X_i^{L0} , X_i^{L1} and X_i^{L2} were sub-vectors of the baseline prognostic factors obtained using a backward stepwise procedure. For the sake of brevity, we will speak about PSA dynamics and biomarker's current level/slope when referring actually to the dynamics of $\log(\text{PSA} + 0.1)$ and the current level/slope of $Y_i^*(t)$, respectively.

5.2.2. Multi-state sub-model specification. In the multi-state sub-part, the determination of prognostic factors and proportionality between baseline intensities was also made by considering no link between the two processes ($\eta = 0$) and unspecified baseline intensities (i.e. using a standard semi-parametric multi-state model). The full sub-model considered transition-specific baseline intensities and transition-specific effects of baseline prognostic factors. To reduce the excessive number of parameters to be estimated, proportional baseline intensities were first assumed for some transitions. Clinically, it made sense to consider proportional baseline intensities for transitions leading to local recurrence or HT: $\lambda_{01,0}(t) = \exp(-\zeta_{02})\lambda_{02,0}(t) = \exp(-\zeta_{12})\lambda_{12,0}(t)$; and for the transitions leading to distant recurrence: $\lambda_{03,0}(t) = \exp(-\zeta_{13})\lambda_{13,0}(t) = \exp(-\zeta_{23})\lambda_{23,0}(t)$. These assumptions were confirmed by the data. We could not make the same assumption for all transitions leading to death because the proportional hazards assumption was not verified. Instead, we chose $\lambda_{14,0}(t) = \exp(-\zeta_{24})\lambda_{24,0}(t)$, and $\lambda_{04,0}(t)$ was stratified on the cohort. This procedure reduced the number of baseline intensities to six. A second step consisted in selecting

the prognostic factors. Factors with an associated p -value > 0.5 were removed, and common covariate effects on several transitions were considered using multivariate Wald tests. For example, the baseline T-stage category had the same effect on transition intensities $0 \rightarrow 1$, $0 \rightarrow 3$ and $2 \rightarrow 3$. Finally, prognostic factors and the log-coefficients of proportionality between baseline intensities with p -value < 0.1 were selected by using a backward stepwise procedure.

5.2.3. Joint multi-state model specification. In the joint model, log baseline intensities approximated by linear combinations of cubic B-splines with three internal knots replaced the unspecified ones. Note that the first knot was placed at 1 year to take into account the null risk of recurrence before 1 year in these data. The dependence function $W_{hk,i}(b_i, t)$ was the same for all the transitions $h \rightarrow k$ and was determined using Wald tests. It resulted that the combination of the true current level and the true current slope of the biomarker fitted at best the relationship between PSA dynamics and the instantaneous risk to transit between health states. Thus, the multi-state sub-model was as follows:

$$\lambda_{hk}^i(t|b_i) = \lambda_{hk,0}(t) \exp \left\{ X_{hk,i}^S \top \gamma_{hk} + \left(\begin{array}{c} Y_i^*(t) \\ \partial Y_i^*(t)/\partial t \end{array} \right)^\top \left(\begin{array}{c} \eta_{hk,level} \\ \eta_{hk,slope} \end{array} \right) \right\},$$

The relations between $\lambda_{hk,0}(t)$ and the final $X_{hk,i}^S$, for $h, k \in \{0, \dots, 4\}$ are indicated in Section 5.2.2 and in Table III. Note that the covariates that were removed of the joint model specification are not in Table III.

5.3. Results

The parameter estimates of the joint multi-state model are presented in Table III. These parameters were those selected according to the procedure described previously. The inference was performed using 9 and 9 quadrature points with the two-step Gauss–Hermite quadrature rule. The parameters of the baseline intensities are not shown here for clarity.

The estimated regression parameters in the longitudinal sub-part confirmed that pre-treatment PSA level was associated with the initial PSA level and the biphasic PSA trajectory; T-stage value was associated both with the short-term and the long-term dynamics, while Gleason score was only associated with the long-term trajectory. Higher values of these covariates measured at baseline corresponded to higher long-term PSA levels. The cohort effect indicated a significant difference between the two cohorts only for the long-term PSA evolution, with a greater long term increase of PSA in Vancouver.

For the multi-state process, the model showed that an advanced initial stage was not always associated with the intensities of transitions between health states after adjustment for the PSA dynamics. In particular, the Gleason score had significant effects on two transition intensities only. Moreover, having a high PSA value at baseline was significantly associated with a higher instantaneous risk to directly experience HT initiation or death after EBRT, but reduced the intensities of transitions leading to distant recurrence or death after a previous event. A poor (i.e. higher) T-stage category at baseline had globally a deleterious effect on the clinical endpoints. For the transitions from end of EBRT or HT initiation to distant recurrence, a patient with a Gleason score of 7 at baseline had a $2.60 = \exp(0.954)$ (95% CI = 1.60–4.20) higher hazard to transit than a patient with a Gleason score < 7 . The cohort was significantly associated with the intensities of transitions leading to death after clinical recurrence or HT initiation – and the direct transition leading to local recurrence after end of BRT. The instantaneous risk to experience these transitions was higher in BCCA. The cohort effect was also significant, with higher intensities in RTOG 9406, for the direct transitions from local recurrence or HT initiation to distant recurrence.

Regarding the association parameters between PSA dynamics (current level and current slope) and clinical progressions, remind that PSA data were collected until the occurrence of the first event. This has an impact on the interpretation of these association parameters. Indeed, because of the focus on the biomarker trajectory before the first event, posterior marker values were extrapolated according to this basal PSA trajectory. We found highly significant deleterious effects of the PSA dynamics on the intensities of transitions from the initial state to all the types of progression (local recurrence, HT or distant recurrence). For example, after adjustment for covariates and for the true slope of the biomarker, an increase of one unit of the true biomarker's level (log PSA without error measurement) induced a $1.43 = \exp(0.358)$ (95% CI = 1.45–1.89) higher risk to experience a local recurrence. These results were expected: in patients with localized prostate cancer and treated by radiotherapy, a persistently high PSA level or/and a strong increase of PSA leads to higher hazard to experience a clinical recurrence or an

Table III. Parameter estimates, standard errors and *p*-values in the joint multi-state model on the pooled data (*N* = 1474).

	Longitudinal process				Multi-state process		
	Value	StdErr	<i>p</i> -value		Value	StdErr	<i>p</i> -value
β_0	-0.26	0.06	< 0.001	$\gamma_{02,iPSA}$	0.35	0.08	< 0.001
$\beta_{0,iPSA}$	0.80	0.03	< 0.001	$\gamma_{04,iPSA}$	0.25	0.08	0.001
$\beta_{0,cohort}$	-0.01	0.02	0.541	$\gamma_{(13,14,23,24,34),iPSA}$	-0.25	0.08	0.001
β_1	0.70	0.14	< 0.001	$\gamma_{(01,03,23),tstage2}$	0.92	0.18	< 0.001
$\beta_{1,iPSA}$	0.89	0.06	< 0.001	$\gamma_{(01,03,23),tstage3-4}$	0.76	0.23	0.001
$\beta_{1,tstage2}$	0.38	0.08	< 0.001	$\gamma_{(12,14,34),tstage2}$	-0.11	0.25	0.659
$\beta_{1,tstage3-4}$	0.47	0.13	< 0.001	$\gamma_{(12,14,34),tstage3-4}$	0.33	0.30	0.271
$\beta_{1,cohort}$	-0.04	0.04	0.346	$\gamma_{(03,23)gleason7}$	0.95	0.25	< 0.001
β_2	-0.19	0.04	< 0.001	$\gamma_{(03,23)gleason8-10}$	0.07	0.43	0.873
$\beta_{2,iPSA}$	0.19	0.02	< 0.001	$\gamma_{(01,14,24,34),cohort}$	-0.42	0.06	< 0.001
$\beta_{2,tstage2}$	0.14	0.02	< 0.001	$\gamma_{(13,23),cohort}$	0.88	0.17	< 0.001
$\beta_{2,tstage3-4}$	0.26	0.04	< 0.001	$\zeta_{(12,13)}$	4.19	0.38	< 0.001
$\beta_{2,gleason7}$	0.07	0.02	< 0.001	ζ_{23}	3.08	0.53	< 0.001
$\beta_{2,gleason8-10}$	0.22	0.04	< 0.001	$\eta_{01,level}$	0.36	0.09	< 0.001
$\beta_{2,cohort}$	-0.06	0.01	< 0.001	$\eta_{02,level}$	0.50	0.07	< 0.001
$\log(\sigma)$	-1.30	0.01		$\eta_{03,level}$	0.42	0.12	< 0.001
				$\eta_{04,level}$	-0.15	0.05	0.005
D_{11}	0.37	0.02		$\eta_{12,level}$	-0.17	0.10	0.095
D_{12}	0.01	0.01		$\eta_{13,level}$	-0.43	0.20	0.033
D_{13}	0.35	0.03		$\eta_{14,level}$	0.10	0.14	0.456
D_{22}	0.14	0.01		$\eta_{23,level}$	-0.17	0.10	0.081
D_{23}	0.25	0.02		$\eta_{24,level}$	0.05	0.05	0.346
D_{33}	1.70	0.09		$\eta_{34,level}$	0.02	0.08	0.813
				$\eta_{01,slope}$	2.63	0.31	< 0.001
				$\eta_{02,slope}$	3.11	0.25	< 0.001
				$\eta_{03,slope}$	2.68	0.55	< 0.001
				$\eta_{04,slope}$	0.92	0.34	0.007
				$\eta_{12,slope}$	2.16	0.63	0.001
				$\eta_{13,slope}$	3.44	0.83	< 0.001
				$\eta_{14,slope}$	-0.22	1.27	0.864
				$\eta_{23,slope}$	1.13	0.68	0.099
				$\eta_{24,slope}$	0.21	0.52	0.692
				$\eta_{34,slope}$	-0.56	0.78	0.472

D_{ij} denotes the *ij*-element of the random effect covariance matrix. $\gamma_{(hk,h'k'),x}$ denotes the common effect of covariate X on the intensities of transitions $h \rightarrow k$ and $h' \rightarrow k'$, i.e. $\gamma_{(hk,h'k'),x} = \gamma_{hk,x} = \gamma_{h'k',x}$. Similarly, $\zeta_{(12,13)} = \zeta_{12} = \zeta_{13}$.

additional therapy. In contrast, for the direct transition leading to death after radiotherapy, we found a deleterious effect of the current slope and a protective effect of the current level of the biomarker: at a given moment in the initial state, for two patients with the same baseline characteristics and the same slope of log PSA, the one with higher PSA value will be less likely to directly die. In this studied population, an important cause of direct death is induced by comorbidities, because most of death from prostate cancer experienced a documented disease progression before. From the local recurrence, there was a large deleterious effect of the current slope of the biomarker for the intensities of transitions leading to the HT or the distant recurrence, and there was a borderline significant protective effect of the current level for the intensity of transition leading to the distant recurrence. From the HT or the distant recurrence, there was no significant effect of the PSA dynamics on the hazard to change state. This was also clinically sensible, as it reflects that progression in these advanced stages is not linked anymore to PSA increase. In practice, criteria other than PSA are considered in this phase of the disease, such as the PCWG2 criteria [26]. Moreover, deaths in patients with HT might be explained by cardiac toxicity because of HT.

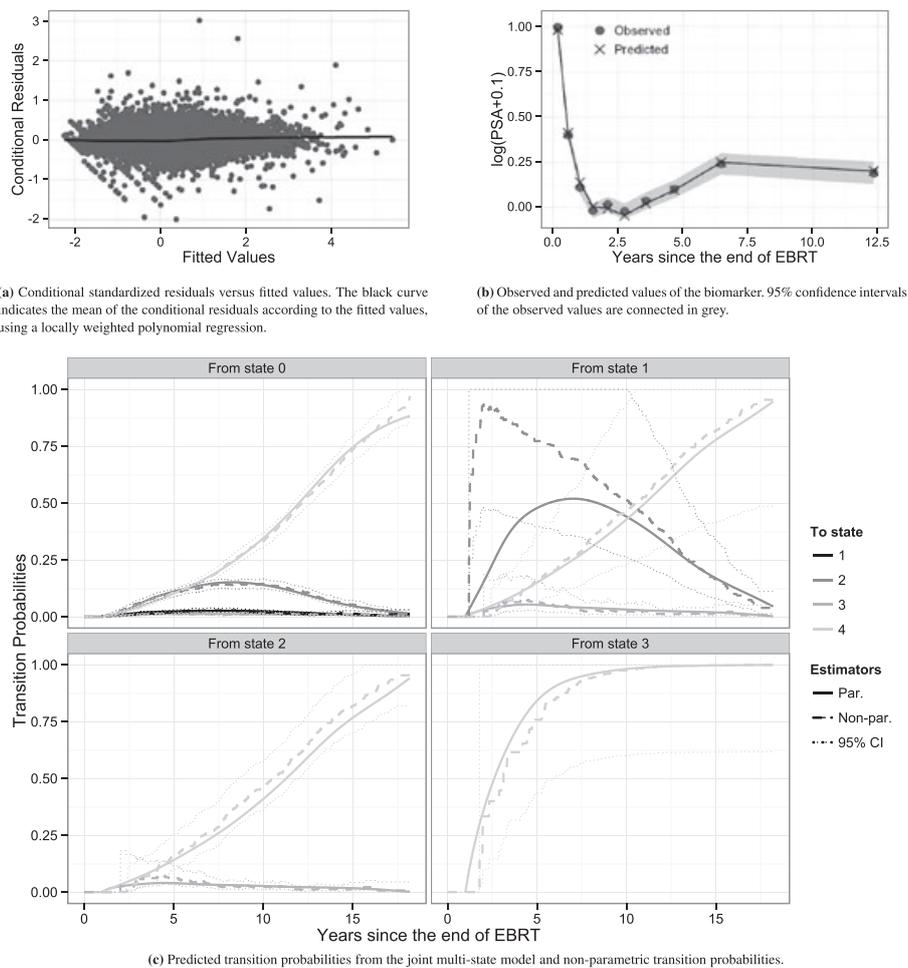


Figure 4. Goodness-of-fit plots for the longitudinal process (a,b) and the multi-state process (c).

5.4. Diagnostics

The parameter estimates of the joint multi-state model were validated by several graphical tools presented in Figure 4. For the longitudinal sub-model, the plotted standardized conditional residuals versus fitted values of the biomarker confirmed the homoscedasticity of the conditional errors. Subject-specific predictions were also compared with observations by plotting the average values by time intervals based on the deciles of the observation times. Ninety-five percent confidence intervals of the observed values were added and confirmed the very good fit of the model to the longitudinal data. For the multi-state sub-model, we focused on $\mathbf{P}(0, t) = \{P_{hk}(0, t)\}$, the matrix of transition probabilities between times 0 and t . We compared our parametric estimator (obtained with the average of the predicted individual transition probabilities from the joint multi-state model) with the Aalen–Johansen estimator (non-parametric estimator of the transition probabilities), both using product integrals. This comparison is fully discussed and detailed in Web Appendix D. These comparisons showed the overall good performances of the joint multi-state model in terms of fit for the transition probabilities, with the exception for transition $1 \rightarrow 2$ for which the immediate pike after EBRT could not be correctly captured by splines.

6. Discussion

The joint model for the longitudinal biomarker PSA and multi-state clinical progression data provides a complete model of prostate cancer progression, which takes into account both classical prognostic factors and PSA dynamics, in order to study factors that influence the transition intensities between clinical health states. The implementation is easy as it relies on `mstate` and `JM` packages. The multi-state data are prepared with `mstate` package, and a slightly modified `jointModel()` function carries out the estimation procedure. The estimation program has been validated by simulations, with very good

performances. Even when the number of subjects experiencing some direct transitions was low, coverage rates remained satisfactory. The simulations underlined, however, some bias in the estimates when the dimension of the random effects increased (≥ 3 random effects) and/or too few quadrature points (3 points in particular) were used. To address this problem, we proposed a two-step procedure that updates the location of the quadrature points and improves the quality of the approximation of the integral over the random effects while keeping a small number of quadrature points. This also reduces substantially the computation time compared with a pseudo-adaptive Gauss–Hermite rule with a much larger number of quadrature points. With this new rule, models with three random effects can be correctly estimated using 9 and 9 quadrature points. Diagnostic graphical tools were also proposed to assess the goodness-of-fit of the model (methodology detailed in Web Supplementary Material D).

The application confirmed that the PSA dynamics strongly impacted the instantaneous risk to experience a clinical recurrence or HT initiation after the end of radiotherapy. The current slope of the biomarker had also a highly significant deleterious effect on the hazard to transit from local recurrence to HT or distant recurrence. Conversely, extrapolating the biomarker's dynamics did not impact anymore the transition intensities from the HT initiation state or the distant recurrence state. This highlights that in the advanced cancers, the PSA – and especially the collected measures prior to the first event – is not of importance anymore. In these situations, other criteria have to be monitored. Note that data posterior to the first clinical recurrence or the HT were not available in our application. When available, it would be of great interest to include them in order to capture the effect of the actual marker dynamics rather than the basal trajectory. However, it would also usually imply a much more complex model for the longitudinal marker as the dynamics might change.

Previous works in prostate cancer had found a strong association between slope of log-PSA and any clinical recurrence (see Sène *et al.* [27]; Taylor *et al.* [28]), by considering all the recurrences in a composite event and the HT as a time-dependent covariate. The limit of these approaches was that in practice considering the type of progression is of major importance as the care greatly depends on the type of risk the patient has. The joint multi-state model formalizes this need. In the same way as it was done with a single event (see Proust-Lima and Taylor [29]; Rizopoulos [30]), individualized dynamic predictions of each type of progression could be derived from this model in order to precisely quantify the risk of each type of progression according to the PSA history. For example, the cumulative probability for subject i to reach state k between times s and t , $s \leq t$, given he was in state h at time s , could be expressed as follows: $\pi_{hk}^i(s, t) = \int_s^t \Pr(E_i(u) = k | E_i(s) = h, Y_i^{(s)}, X_i^{L(s)}, X_i^S) du$, with $Y_i^{(s)}$ the history (i.e. collected measures) of the marker up to time s , $X_i^{L(s)}$ the history of the longitudinal sub-model covariates until time s , and $X_i^S = \{X_{hk,i}^S\}$ the matrix containing the prognostic factors for all the state transitions.

In this article, we made several assumptions. First, we assumed a continuous and Markov multi-state process as it was clinically relevant for the progression of prostate cancer after treatment. However, in other contexts, a semi-Markov process, which considers the time spent in the current state, could be defined as well. In dementia for example, the multi-state process might include three states (healthy, demented and dead) and consider the time spent in the demented state before death (see Commenges *et al.* [31]). The joint multi-state model we proposed and its associated implemented function handle for semi-Markov. Second, through the (semi-)Markov assumption, we assume that the dependency between the transition times for a given subject is entirely explained by the prognostic factors and the marker dynamics. This assumption could be relaxed by including some frailty term in the multi-state model. However, Putter and van Houwelingen [32] pointed out that identifiability of multi-state models with frailties is weak, and the interpretation becomes not obvious. Third, we chose the nature of the dependence function using goodness-of-fit measures in the application, but other strategies could be used. For example, this choice might rely on predictive accuracy measures when focusing on prediction (see Sène *et al.* [33]). Finally, there was no delayed entry in the prostate cancer application. However, the method implicitly handles delayed entry by conditioning the log-likelihood on the state at entry in the study, as it was carried out by Commenges [34].

In summary, we introduce here a first joint model for longitudinal and multi-state clinical progression data. We showed that this model can easily be implemented under R and can be applied in practice through an example, the prostate cancer progression, which is one of many biomedical areas in which such data are collected. This model that captures the complete information about the progression opens to much more precise knowledge of diseases and specific dynamic predictions.

Acknowledgements

The authors thank Paul Sargos and Pierre Richaud from the Institut Bergonié (Bordeaux, France) for their availability and their expertise in clinical interpretations. Computer time for this study was provided by the computing facilities MCIA (Mésocentre de Calcul Intensif Aquitain) of the Université de Bordeaux and of the Université de Pau et des Pays de l'Adour. This work was supported by a joint grant from INSERM and Région Aquitaine, and a grant from the Institut de Recherche en Santé Publique [grant AAP12CanBio16]. The RTOG trial and J. Dignam's efforts were supported by Public Health Service grants U10 CA21661 and U10 CA180822 from the National Cancer Institute, NIH, U.S. Department of Health and Human Services.

References

1. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* 1982; **10**(4):1100–1120.
2. Fisher LD, Lin DY. Time-dependent covariates in the Cox proportional-hazards regression model. *Annual Review of Public Health* 1999; **20**(1):145–157.
3. Faucett CL, Thomas DC. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine* 1996; **15**(15):1663–1685.
4. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. *Biometrics* 1997; **53**(1):330–339.
5. Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* 2004; **14**(3):809–834.
6. Proust-Lima C, Taylor JM, Williams SG, Ankerst DP, Liu N, Kestin LL, Bae K, Sandler HM. Determinants of change in prostate-specific antigen over time and its association with recurrence after external beam radiation therapy for prostate cancer in five large cohorts. *International Journal of Radiation Oncology* Biology* Physics* 2008; **72**(3):782–791.
7. Taylor JM, Yu M, Sandler HM. Individualized predictions of disease progression following radiation therapy for prostate cancer. *Journal of Clinical Oncology* 2005; **23**(4):816–825.
8. Yu M, Taylor JM, Sandler HM. Individual prediction in prostate cancer studies using a joint longitudinal survival–cure model. *Journal of the American Statistical Association* 2008; **103**(481):178–187.
9. Chi YY, Ibrahim JG. Joint models for multivariate longitudinal and multivariate survival data. *Biometrics* 2006; **62**(2):432–445.
10. 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)* 2009; **58**(1):65–81.
11. Kim S, Zeng D, Chambless L, Li Y. Joint models of longitudinal data and recurrent events with informative terminal event. *Statistics in Biosciences* 2012; **4**(2):262–281.
12. Elashoff RM, Li G, Li N. A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics* 2008; **64**(3):762–771.
13. Rizopoulos D. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Chapman and Hall/CRC Biostatistics: Boca Raton, 2012.
14. Dantan E, Joly P, Dartigues JF, Jacqmin-Gadda H. Joint model with latent state for longitudinal and multistate data. *Biostatistics* 2011; **12**(4):723–736.
15. Tom BD, Farewell VT. Intermittent observation of time-dependent explanatory variables: a multistate modelling approach. *Statistics in Medicine* 2011; **30**(30):3520–3531.
16. Andrinopoulou ER, Rizopoulos D, Takkenberg JJ, Lesaffre E. Joint modeling of two longitudinal outcomes and competing risk data. *Statistics in Medicine* 2014; **33**(18):3167–3178.
17. Gould AL, Boye ME, Crowther MJ, Ibrahim JG, Quartey G, Micallef S, Bois FY. Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA bayesian joint modeling working group. *Statistics in Medicine* 2014; **34**(14):2181–2195.
18. Rizopoulos D. JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software* 2010; **35**(9):1–33.
19. De Wreede LC, Fiocco M, Putter H. The mstate package for estimation and prediction in non-and semi-parametric multi-state and competing risks models. *Computer Methods and Programs in Biomedicine* 2010; **99**(3):261–274.
20. Rizopoulos D. Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive gaussian quadrature rule. *Computational Statistics & Data Analysis* 2012; **56**(3):491–501.
21. Beyersmann J, Allignol A, Schumacher M. *Competing Risks and Multistate Models with R*, Science & Business Media. Springer: New York, 2011.
22. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. *Statistics in Medicine* 2013; **32**(23):4118–4134.
23. Brent RP. *Algorithms for Minimization without Derivatives*. Dover Publications: Mineola, New York, 1973.
24. Michalski JM, Winter K, Purdy JA, Parliament M, Wong H, Perez CA, Roach M, Bosch W, Cox JD. Toxicity after three-dimensional radiotherapy for prostate cancer on RTOG 9406 dose Level V. *International Journal of Radiation Oncology* Biology* Physics* 2005; **62**(3):706–713.

25. Pickles T, Kim-Sing C, Morris WJ, Tyldesley S, Paltiel C. Evaluation of the houston biochemical relapse definition in men treated with prolonged neoadjuvant and adjuvant androgen ablation and assessment of follow-up lead-time bias. *International Journal of Radiation Oncology* Biology* Physics* 2003; **57**(1):11–18.
26. Scher HI, Halabi S, Tannock I, Morris M, Sternberg CN, Carducci MA, Eisenberger MA, Higano C, Bubley GJ, Dreicer R, Petrylak D, Kantoff P, Basch E, Kelly WK, Figg WD, Small EJ, Beer TM, Wilding G, Martin A, Hussain M, Prostate Cancer Clinical Trials Working Group. Design and end points of clinical trials for patients with progressive prostate cancer and castrate levels of testosterone: recommendations of the Prostate Cancer Clinical Trials Working Group. *Journal of Clinical Oncology* 2008; **26**(7):1148–1159.
27. Sène M, Bellera CA, Proust-Lima C. Shared random-effect models for the joint analysis of longitudinal and time-to-event data: application to the prediction of prostate cancer recurrence. *Journal de la Société Française de Statistique* 2014; **155**(1):134–155.
28. Taylor JM, Park Y, Ankerst DP, Proust-Lima C, Williams S, Kestin L, Bae K, Pickles T, Sandler H. Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics* 2013; **69**(1):206–213.
29. 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.
30. Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* 2011; **67**(3):819–829.
31. Commenges D, Joly P, Gégout-Petit A, Liqueur B. Choice between semi-parametric estimators of Markov and non-Markov multi-state models from coarsened observations. *Scandinavian Journal of Statistics* 2007; **34**(1):33–52.
32. Putter H, van Houwelingen HC. Frailties in multi-state models: are they identifiable? do we need them? *Statistical Methods in Medical Research* 2015; **24**(6):675–692.
33. Sène M, Taylor JM, Dignam JJ, Jacqmin-Gadda H, Proust-Lima C. Individualized dynamic prediction of prostate cancer recurrence with and without the initiation of a second treatment: development and validation. *Statistical Methods in Medical Research* 2014. [Epub ahead of print].
34. Commenges D. Inference for multi-state models from interval-censored data. *Statistical Methods in Medical Research* 2002; **11**(2):167–182.

Supporting information

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

3.2 Supplementary material

This article was published with a supplementary material available at the publisher's website. This supplementary material is introduced hereinafter. It starts in Section A with examples of R scripts to estimate joint multi-state models. In Section B, it details the multi-step pseudo-adaptive Gauss-Hermite rule which was used in the simulation study and the application of the main manuscript. Results of a supplementary simulation study, which considered 5 states and 10 transitions, are given in Section C. Finally, the supplementary material ends in Section D by detailing the statistical methodology used in the main manuscript to compare parametric and non-parametric transition probabilities.

Supplementary Material for “Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer”

Loïc Ferrer ^{*a}, Virginie Rondeau^a, James J. Dignam^b, Tom Pickles^c, H el ene Jacqmin-Gadda^a and C ecile Proust-Lima^a

^a INSERM U1219, ISPED, Universit e de Bordeaux, Bordeaux, France

^b Department of Public Health Sciences, University of Chicago, and NRG Oncology, U.S.A.

^c Department of Radiation Oncology, University of British Columbia, Canada

* Correspondance author: Lo ic Ferrer, INSERM U1219, ISPED, Universit e de Bordeaux, 146 rue L eo Saignat, 33076 Bordeaux Cedex, France.

Notations are introduced in the main manuscript.

A. Example of R code

`JMstateModel()` function and several detailed examples are available on <https://github.com/LoicFerrer/JMstateModel/>.

We give here the script to estimate the joint multi-state model described in the main simulation study (see Section 4 in the main article), with a non-homogeneous Markov multi-state model which included three states and three transitions. The same covariate, called X in the code below, impacted the longitudinal and multi-state processes. The longitudinal sub-model had a random intercept and two random effects associated to the short term drop and the long term linear trend of the marker. The log baseline intensities were approximated using B-splines and the dependency between the two processes was explained through the true current level and the true current slope of the biomarker.

```
# Load the packages and the function to estimate joint multi-state models:
library(mstate) # Please use the version 0.2.7
library(JM)
source("JMstateModel.R")

# Import two databases which contain longitudinal and multi-state data:
load("data.RData")
```

*loic.ferrer@inserm.fr

```

#####
### Longitudinal sub-part ###
#####

# Fit the longitudinal responses using a linear mixed model:
lmeFit <- lme(fixed = Y ~ (times + I((1 + times)^(-1.2) - 1)) * X,
             data = data_long,
             random = ~ (times + I((1 + times)^(-1.2) - 1)) | id,
             method = "REML",
             control = list(opt = "optim"))

#####
### Multi-state sub-part ###
#####

# Construct the 3*3 matrix of transitions:
tmat <- matrix(NA, 3, 3)
tmat[1, 2:3] <- 1:2
tmat[2, 3] <- 3
dimnames(tmat) <- list(from = c("State_0", "State_1", "State_2"),
                      to = c("State_0", "State_1", "State_2"))

tmat
# The transition '0 -> 1' is called '1', '0 -> 2' is called '2' and
# '1 -> 2' is called '3'.

# Define the covariate in the multi-state sub-part:
covs <- "X"

# The 'msprep()' function divides the multi-state database in order to have
# one line per transition at risk for each subject, with 'Tstart' the
# entry time in the current state, and 'Tstop' the time of transition or
# censorship; 'status' denotes if the transition has been performed:
data_mstate <- msprep(time = c(NA, "time_of_State_1", "time_of_State_2"),
                    status = c(NA, "State_1", "State_2"),
                    data = data_surv,
                    trans = tmat,
                    keep = covs,
                    id = "id")

# 'expand.covs()' permits to define the set of covariates which impacts
# each transition:
data_mstate <- expand.covs(data_mstate, covs,
                        append = TRUE, longnames = FALSE)

# Multi-state model with transition-specific proportional intensities:
coxFit <- coxph(Surv(Tstart, Tstop, status) ~
              X.1 + X.2 + X.3 + strata(trans),
              data = data_mstate,
              method = "breslow",
              x = TRUE,

```

```
model = TRUE)
```

```
#####  
#### Joint multi-state part ####  
#####
```

```
# To define the dependency on the slope of the marker, it is necessary  
# to specify the derivative of the fixed and random parts in the mixed model,  
# and indicate which covariates are kept, :
```

```
dForm <- list(fixed = ~ 1 + I((-1.2) * ((1 + times)^(-2.2))) +  
              X + I((-1.2) * ((1 + times)^(-2.2))):X,  
             indFixed = c(2:3, 5:6),  
             random = ~ 1 + I((-1.2) * ((1 + times)^(-2.2))),  
             indRandom = 2:3)
```

```
# Joint multi-state model with:
```

```
# - true current level and true current slope of the marker as dependence function,  
# - cubic B-splines with 1 internal knot for each log-baseline intensity,  
# - 15 Gauss-Kronrod quadrature points to approximate the integral over time  
#   (by default),  
# - 3 Gauss-Hermite quadrature points in the pseudo-adaptive numerical  
#   integration to approximate the integral over random effects.
```

```
jointFit_1step_GHk3 <-  
  JMstateModel(lmeObject = lmeFit,  
               survObject = coxFit,  
               timeVar = "times",  
               parameterization = "both",  
               method = "spline-PH-aGH",  
               interFact = list(value = ~strata(trans) - 1,  
                                slope = ~strata(trans) - 1,  
                                data = data_mstate),  
               derivForm = dForm,  
               Mstate = TRUE,  
               data.Mstate = data_mstate,  
               ID.Mstate = "id",  
               control = list(GHk = 3, lng.in.kn = 1))
```

```
summary(jointFit_1step_GHk3)
```

```
# Same joint multi-state model with:
```

```
# - 9 Gauss-Hermite quadrature points in the pseudo-adaptive numerical  
#   integration to approximate the integral over random effects.
```

```
jointFit_1step_GHk9 <-  
  JMstateModel(lmeObject = lmeFit,  
               survObject = coxFit,  
               timeVar = "times",  
               parameterization = "both",  
               method = "spline-PH-aGH",  
               interFact = list(value = ~strata(trans) - 1,  
                                slope = ~strata(trans) - 1,  
                                data = data_mstate),  
               derivForm = dForm,  
               Mstate = TRUE,
```

```

        data.Mstate = data_mstate,
        ID.Mstate = "id",
        control = list(GHk = 9, lng.in.kn = 1))
summary(jointFit_1step_GHk9)

# To use the multi-step pseudo-adaptive Gauss-Hermite rule, we have to source
# two functions inspired by JM:
source("modified.log.posterior.b2.R")
source("modified.ranef.jointModel.R")

# Same joint multi-state model with:
# - 9 and 9 Gauss-Hermite quadrature points in the two-step pseudo-adaptive
# numerical integration to approximate the integral over random effects.
# We can choose the posterior mode (true definition) or the posterior mean
# (faster) of the random effects of the fitted joint model (defined in 'init')
# to update the quadrature points. Here the mode is used.
jointFit_2step_GHk9_9 <-
  JMstateModel(lmeObject = lmeFit,
    survObject = coxFit,
    timeVar = "times",
    parameterization = "both",
    method = "spline-PH-aGH",
    interFact = list(value = ~strata(trans) - 1,
      slope = ~strata(trans) - 1,
      data = data_mstate),
    derivForm = dForm,
    Mstate = TRUE,
    data.Mstate = data_mstate,
    ID.Mstate = "id",
    control = list(GHk = 9, lng.in.kn = 1),
    init = jointFit_1step_GHk9,
    init.type.ranef = "mode")
summary(jointFit_2step_GHk9_9)

```

B. Multi-step pseudo-adaptive Gauss-Hermite rule

The main difficulty in the inference is due to the numerical approximation of the integral over the random effects in the likelihood function. To deal with this challenge, Rizopoulos [1] proposed to use the pseudo-adaptive Gauss-Hermite rule. The idea is to first fit the linear mixed sub-model of the joint model, and to use the estimated parameters and the posterior distribution of the random effects to correctly rescale the subject-specific integrands. This part, and notably the following notations, are derived from his book.

We call $A(\cdot)$ a function of the random effects. The pseudo-adaptive Gauss-Hermite rule is defined as:

$$\mathbb{E}\{A(\theta, b_i) | E_i, Y_i; \theta\} \approx 2^{q/2} |\tilde{B}_i|^{-1} \sum_{t_1=1}^K \dots \sum_{t_q=1}^K \pi_t A(\theta, \tilde{r}_t) f(\tilde{r}_t | E_i, Y_i; \theta) \exp(-\|\tilde{r}_t\|^2), \quad (1)$$

where K is the number of quadrature points and $\tilde{r}_t = \tilde{b}_i + \sqrt{2} \tilde{B}_i^{-1} b_t$ uses the prespecified abscissas $b_t = (b_{t_1}, \dots, b_{t_q})^\top$ with corresponding prespecified weights π_t . The empirical Bayes estimates \tilde{b}_i are obtained from the linear mixed sub-model with $\tilde{\theta}_Y$ the corresponding estimated vector of parameters, using:

$$\tilde{b}_i = \arg \max_b \{\log f(Y_i, b; \tilde{\theta}_Y)\}.$$

\tilde{B}_i is the Choleski factor of \tilde{H}_i , where \tilde{H}_i^{-1} is the covariance matrix of \tilde{b}_i :

$$\begin{aligned} \tilde{H}_i &= - \frac{\partial^2 \log f(Y_i, b; \tilde{\theta}_Y)}{\partial b \partial b^\top} \Big|_{b=\tilde{b}_i} \\ &= \frac{Z_i^\top Z_i}{\tilde{\sigma}^2} + \tilde{D}^{-1}, \end{aligned}$$

where $\tilde{\sigma}$ and \tilde{D} are the estimates of σ and D respectively ($\tilde{\sigma}, \tilde{D} \in \tilde{\theta}_Y$).

When the dimension of the random effects increases, it is necessary to use a large number of quadrature points to obtain good estimates. However, it can become difficult to handle big matrices with acceptable computation time. To avoid this problem, we propose a n -step procedure that aims to relocate the quadrature points n times according to the joint models estimates. In the following, for clarity and notations, we only detail the procedure when $n = 2$.

This rule called ‘‘two-step pseudo-adaptive Gauss-Hermite rule’’ is in fact decomposed into three steps: first, the linear mixed sub-model and the survival sub-model are fitted separately to initialize the parameters in the inference algorithm of the joint model as done by default in JM (see Section 3.2 in the main article). \tilde{b}_i and \tilde{H}_i are computed to apply the pseudo-adaptive Gauss-Hermite rule (1). An acceptable number of quadrature points is used and the joint model is estimated. The

corresponding estimated parameter vector $\hat{\theta}$ is obtained, and the location of the quadrature points is updated using the adaptive Gauss-Hermite rule:

$$\mathbb{E}\{A(\theta, b_i) | E_i, Y_i; \theta\} \approx 2^{q/2} |\hat{B}_i|^{-1} \sum_{t_1=1}^{\hat{K}} \dots \sum_{t_q=1}^{\hat{K}} \hat{\pi}_t A(\theta, \hat{r}_t) f(\hat{r}_t | E_i, Y_i, \theta) \exp(-\|\hat{b}_t\|^2), \quad (2)$$

where \hat{K} is the selected number of quadrature points and $\hat{r}_t = \hat{b}_i + \sqrt{2} \hat{B}_i^{-1} \hat{b}_t$ uses prespecified abscissas $\hat{b}_t = (\hat{b}_{t1}, \dots, \hat{b}_{tq})^\top$ with corresponding prespecified weights $\hat{\pi}_t$. The empirical Bayes estimates \hat{b}_i are extracted from the fitted joint model:

$$\hat{b}_i = \arg \max_b \{\log f(E_i, Y_i, b; \hat{\theta})\}.$$

\hat{B}_i is the Choleski factor of \hat{H}_i , where \hat{H}_i^{-1} is the covariance matrix of the random effects:

$$\begin{aligned} \hat{H}_i &= - \frac{\partial^2 \log f(E_i, Y_i, b; \hat{\theta})}{\partial b \partial b^\top} \Big|_{b=\hat{b}_i} \\ &= \frac{Z_i^\top Z_i}{\hat{\sigma}^2} + \hat{D}^{-1} - \sum_{r=0}^{m_i-1} \delta_{i(r+1)} \frac{\partial^2 W_{E_i(T_{ir}), E_i(T_{i(r+1)})}(\hat{b}_i, T_{i(r+1)}; \hat{\theta})^\top \hat{\eta}_{E_i(T_{ir}), E_i(T_{i(r+1)})}}{\partial \hat{b}_i \partial \hat{b}_i^\top} + \\ &\quad \sum_{r=0}^{m_i-1} \sum_{k, k \neq E_i(T_{ir})} \int_{T_{ir}}^{T_{i(r+1)}} \lambda_{E_i(T_{ir}), k}^i(s | \hat{b}_i; \hat{\theta}) \times \\ &\quad \left\{ \frac{[\partial W_{E_i(T_{ir}), k}(\hat{b}_i, s; \hat{\theta})^\top \hat{\eta}_{E_i(T_{ir}), k}]^2}{\partial \hat{b}_i \partial \hat{b}_i^\top} + \frac{\partial^2 W_{E_i(T_{ir}), k}(\hat{b}_i, s; \hat{\theta})^\top \hat{\eta}_{E_i(T_{ir}), k}}{\partial \hat{b}_i \partial \hat{b}_i^\top} \right\} ds, \end{aligned}$$

where $\hat{\sigma}$, \hat{D} and $\hat{\eta}_{hk}$ are the estimates of σ , D and η_{hk} respectively ($\hat{\sigma}, \hat{D}, \hat{\eta}_{hk} \in \hat{\theta}$). Note that the convention $\hat{\eta}_{hk} = 0$ is used for the direct transitions $h \rightarrow k$ which are not observed.

C. Supplementary simulation study

The main simulations, described in Section 4 of the manuscript, considered a multi-state process including three states and three transitions and a longitudinal model including a nonlinear trajectory captured by three random effects. Using samples of 500 subjects, we showed that correct inference was obtained when using a two-step pseudo-adaptive Gauss-Hermite rule with 9 quadrature points at each step. This was not the case with a one-step pseudo-adaptive Gauss-Hermite rule with 3 or 9 quadrature points.

In these supplementary simulations, we investigate whether the inference remains correct when the model includes more states, when some transitions are experienced only by a few subjects, and when the model has much more parameters to be estimated. These supplementary simulations are performed on samples of 1500 subjects with generation parameters chosen from the application data.

Data generation

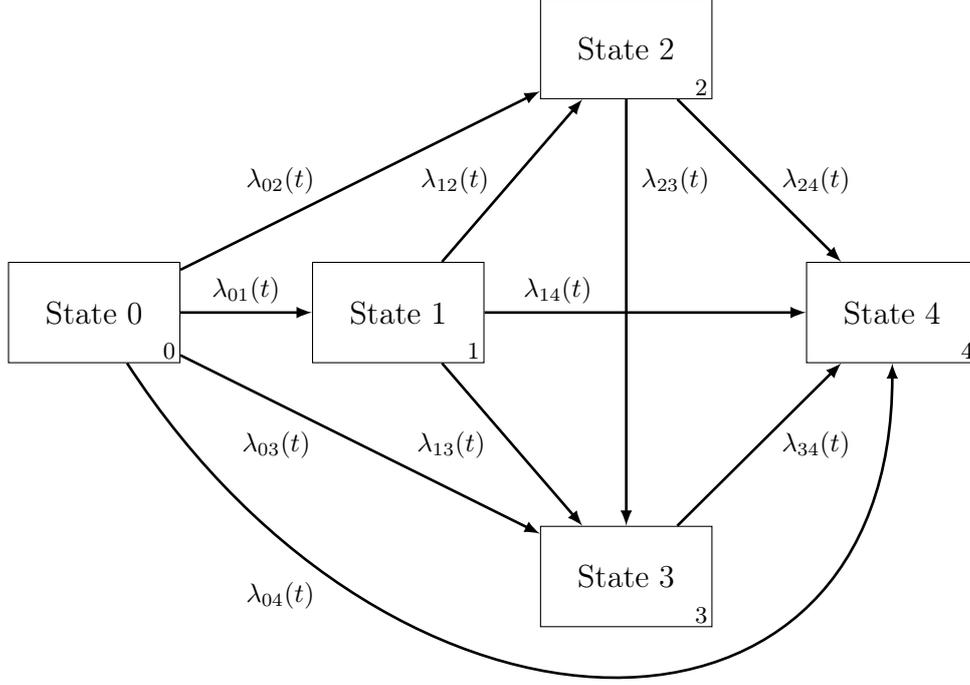
For each subject $i = 1, \dots, 1500$ of the 500 replicates, the longitudinal and multi-state data were generated according to the joint multi-state model defined as:

$$\left\{ \begin{array}{l} Y_{ij} = Y_i^*(t_{ij}) + \epsilon_{ij} \\ = (\beta_0 + \beta_{0,X} X_i + b_{i0}) + (\beta_1 + \beta_{1,X} X_i + b_{i1}) \times t_{ij} + \epsilon_{ij}, \\ \lambda_{hk}^i(t|b_i) = \lambda_{hk,0}(t) \exp \{ \gamma_{hk} X_i + \eta_{hk,\text{level}} Y_i^*(t) + \eta_{hk,\text{slope}} \partial Y_i^*(t) / \partial t \}, \end{array} \right. \quad (3)$$

where the multi-state process that included five states ($h, k \in \{0, 1, 2, 3, 4\}$) and ten transitions is described in Web Figure 1.

As in the application, we considered that some baseline intensities were proportional: $\lambda_{01,0}(t) = \exp(-\zeta_{02})\lambda_{02,0}(t) = \exp(-\zeta_{12})\lambda_{12,0}(t)$, $\lambda_{03,0}(t) = \exp(-\zeta_{13})\lambda_{13,0}(t) = \exp(-\zeta_{23})\lambda_{23,0}(t)$, $\lambda_{04,0}(t) = \exp(-\zeta_{14})\lambda_{14,0}(t) = \exp(-\zeta_{24})\lambda_{24,0}(t) = \exp(-\zeta_{34})\lambda_{34,0}(t)$. Thus, the baseline intensities were only estimated for the transitions $0 \rightarrow 1$, $0 \rightarrow 3$ and $0 \rightarrow 4$.

A linear trajectory of the biomarker was assumed. The times of measurements were $t_{ij} = 0, 0.33, 0.67, \dots, 16.33$. X_i and $b_i = (b_{i0}, b_{i1})^\top$ were generated according to normal distributions with mean 2.04 and variance 0.5, and mean vector $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and variance-covariance matrix $\begin{pmatrix} 0.337 & -0.036 \\ -0.036 & 0.063 \end{pmatrix}$ respectively. The error term ϵ_{ij} was generated from a normal distribution with mean zero and variance 0.230. The log baseline intensities were linear combinations of cubic B-splines with the same knot vector $(0.026, 7.127, 18.201)^\top$.



$$\bar{\psi}_{\text{sim}} = \begin{pmatrix} 618 & 139 & 211 & 44 & 488 \\ 0 & 22 & 85 & 9 & 23 \\ 0 & 0 & 120 & 28 & 148 \\ 0 & 0 & 0 & 13 & 68 \\ 0 & 0 & 0 & 0 & 727 \end{pmatrix}$$

Figure 1: Simulated multi-state process. Arrows indicate the directions of the possible transitions. $\lambda_{hk}(t)$ characterizes the intensity of transition between states h and k at time t . The matrix $\bar{\psi}_{\text{sim}}$ has size $(5, 5)$ and is composed of elements $\bar{\psi}_{\text{sim},(h+1)(k+1)}$, $h, k \in \{0, 1, 2, 3, 4\}$, where $\bar{\psi}_{\text{sim},(h+1)(k+1)}$ is the average number of observed direct transitions $h \rightarrow k$ over the 500 replicates. The diagonal elements $\bar{\psi}_{\text{sim},(h+1)(h+1)}$ denote the average number of patients who were censored in state h . Note that the sum of elements of a row $(h+1)$ of $\bar{\psi}_{\text{sim}}$ corresponds to the average number of patients who experienced the state h .

The vectors of spline coefficients were $(-6.990, -3.428, -5.578, -6.847, -6.067)^\top$ for the transition $0 \rightarrow 1$, $(-7.537, -5.525, -7.191, -4.692, -6.420)^\top$ for the transition $0 \rightarrow 3$, and $(-7.322, -4.345, -3.770, -1.496, -1.261)^\top$ for the transition $0 \rightarrow 4$. Parameters values and knot locations were chosen according to the application data described in the main article.

The procedure to generate the vector of observed times $T_i = (T_{i1}, \dots, T_{im_i})^\top$ was the same as in the main article, with the censoring time generated from an uniform distribution on $[1, 25]$. In this second simulation study, the longitudinal measurements were truncated at T_{im_i} the last observed time of the multi-state process.

Estimated model

The estimation model was the same as the one used to generate the data (3) with $b_i \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D_{11} & D_{12} \\ D_{12} & D_{22} \end{pmatrix}\right)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. The log baseline intensities were approximated by a linear combination of cubic-splines with one internal knot placed at the median of the observed transition times.

Simulation results

The simulation results were obtained through 500 replicates of 1500 individuals. Each joint multi-state model was estimated with 3 quadrature points using the pseudo-adaptive Gauss-Hermite rule (called one-step procedure), and 5–5 quadrature points using the two-step pseudo-adaptive Gauss-Hermite rule (called two-step procedure). The one-step and two-step procedures are detailed in Web Appendix B. The simulation results are presented in Web Table 1.

Overall, these results confirmed the good performances of the implemented function and the necessity to use a certain number of Gauss-Hermite quadrature points. The coverage rates were close to 95% and the relative bias were low, except for some parameters such as $\gamma_{24,X}$ or $\eta_{14,\text{slope}}$, due to the required accuracy. Note that the model performed well even when the number of observed transitions was low, as for the transitions $1 \rightarrow 3$ (9 observed transitions in average) and $1 \rightarrow 4$ (23 observed transitions in average).

Table 1: Simulation results when using a pseudo-adaptive Gauss-Hermite rule with 3 quadrature points, and a two-step pseudo-adaptive Gauss-Hermite rule using 5–5 quadrature points. For each scenario, the statistics depicted are, from left to right: mean, mean standard error, standard deviation, relative bias (in %) and coverage rate (%).

	True value	3 Gauss-Hermite quadrature points (one-step)					5–5 Gauss-Hermite quadrature points (two-step)				
		Mean	StdErr	StdDev	Rel. bias	Cov. rate	Mean	StdErr	StdDev	Rel. bias	Cov. rate
<i>Longitudinal process</i>											
β_0	-0.775	-0.782	0.047	0.047	0.9	95.8	-0.779	0.048	0.050	0.5	95.4
$\beta_{0,X}$	0.531	0.533	0.022	0.023	0.4	94.8	0.532	0.022	0.023	0.3	94.6
β_1	-0.110	-0.107	0.012	0.020	-2.5	77.0	-0.109	0.021	0.020	-0.5	95.6
$\beta_{1,X}$	0.037	0.037	0.006	0.009	0.3	81.8	0.037	0.010	0.010	-0.6	95.4
$\log(\sigma)$	-0.734	-0.734	0.004	0.004	0.0	95.2	-0.734	0.004	0.004	0.0	95.4
<i>Multi-state process</i>											
$\gamma_{01,X}$	0.047	0.034	0.133	0.137	-27.3	93.6	0.028	0.134	0.138	-41.6	93.8
$\gamma_{02,X}$	0.489	0.513	0.109	0.114	5.0	96.2	0.492	0.109	0.109	0.8	95.2
$\gamma_{03,X}$	0.048	0.012	0.234	0.235	-74.3	95.2	0.040	0.236	0.233	-17.5	96.2
$\gamma_{04,X}$	0.058	0.057	0.071	0.066	-2.6	99.2	0.052	0.072	0.068	-10.7	96.4
$\gamma_{12,X}$	0.916	0.928	0.192	0.179	1.3	98.2	0.936	0.193	0.195	2.2	94.4
$\gamma_{13,X}$	0.351	0.256	0.580	0.666	-27.1	95.4	0.288	0.575	0.630	-18.0	95.4
$\gamma_{14,X}$	-0.385	-0.376	0.379	0.449	-2.2	89.6	-0.390	0.381	0.408	1.5	92.2
$\gamma_{23,X}$	-0.210	-0.213	0.297	0.274	1.2	99.0	-0.204	0.297	0.293	-2.9	96.4
$\gamma_{24,X}$	0.007	-0.024	0.130	0.127	-470.3	94.6	-0.006	0.129	0.138	-187.9	93.8
$\gamma_{34,X}$	0.262	0.286	0.195	0.187	8.8	95.0	0.274	0.197	0.204	4.6	94.2
ζ_{02}	-0.959	-1.063	0.396	0.430	10.8	90.8	-1.006	0.398	0.403	4.9	94.2
ζ_{12}	1.496	1.420	0.546	0.503	-5.0	98.4	1.404	0.551	0.549	-6.1	94.8
ζ_{13}	1.384	1.153	1.457	1.493	-16.7	95.4	1.260	1.448	1.488	-8.9	96.0
ζ_{14}	1.469	1.420	0.735	0.824	-3.3	91.6	1.466	0.735	0.780	-0.2	93.0
ζ_{23}	2.563	2.480	0.877	0.932	-3.2	95.6	2.519	0.886	0.858	-1.7	95.8
ζ_{24}	0.511	0.571	0.364	0.330	11.7	97.2	0.520	0.364	0.381	1.8	94.2
ζ_{34}	1.721	1.652	0.476	0.428	-4.0	96.4	1.687	0.483	0.499	-2.0	94.6
$\eta_{01,level}$	0.823	0.814	0.119	0.120	-1.1	94.4	0.834	0.119	0.127	1.4	92.4
$\eta_{02,level}$	0.910	0.907	0.099	0.090	-0.3	96.0	0.917	0.100	0.099	0.8	95.6
$\eta_{03,level}$	0.527	0.560	0.187	0.177	6.3	96.6	0.535	0.189	0.202	1.6	93.2
$\eta_{04,level}$	0.229	0.232	0.051	0.046	1.3	97.2	0.233	0.052	0.050	1.8	94.8
$\eta_{12,level}$	-0.522	-0.529	0.160	0.151	1.2	98.4	-0.522	0.163	0.176	-0.0	93.4
$\eta_{13,level}$	-0.563	-0.429	0.388	0.462	-23.8	90.6	-0.539	0.394	0.450	-4.4	93.0
$\eta_{14,level}$	-0.104	-0.099	0.285	0.280	-4.7	96.4	-0.133	0.288	0.292	27.9	94.2
$\eta_{23,level}$	-0.444	-0.450	0.201	0.197	1.3	97.2	-0.438	0.200	0.201	-1.4	95.6
$\eta_{24,level}$	0.067	0.077	0.087	0.088	15.8	93.0	0.071	0.087	0.092	6.1	93.4
$\eta_{34,level}$	-0.591	-0.590	0.119	0.133	-0.0	94.6	-0.585	0.121	0.128	-1.0	93.6
$\eta_{01,slope}$	0.097	0.099	0.057	0.054	2.3	96.4	0.097	0.057	0.059	-0.4	94.4
$\eta_{02,slope}$	0.252	0.255	0.047	0.050	1.5	93.4	0.250	0.047	0.047	-0.5	93.2
$\eta_{03,slope}$	0.357	0.329	0.110	0.102	-7.7	96.6	0.357	0.111	0.111	0.2	95.4
$\eta_{04,slope}$	-0.572	-0.572	0.047	0.047	0.1	94.4	-0.577	0.048	0.049	0.8	94.4
$\eta_{12,slope}$	0.922	0.927	0.093	0.089	0.5	96.2	0.933	0.095	0.099	1.2	93.6
$\eta_{13,slope}$	1.112	1.064	0.283	0.325	-4.4	91.0	1.110	0.283	0.322	-0.2	93.6
$\eta_{14,slope}$	0.023	0.007	0.299	0.276	-68.5	95.6	0.043	0.299	0.292	87.1	94.2
$\eta_{23,slope}$	0.397	0.392	0.154	0.166	-1.4	93.2	0.391	0.153	0.159	-1.6	94.8
$\eta_{24,slope}$	0.039	0.028	0.094	0.089	-26.2	95.4	0.036	0.094	0.099	-5.9	93.0
$\eta_{34,slope}$	0.596	0.610	0.102	0.118	2.3	93.4	0.600	0.102	0.109	0.5	95.4
<i>Random effects</i>											
D_{11}	0.337	0.335	0.014	0.014	-0.7	93.8	0.336	0.014	0.013	-0.5	95.8
D_{12}	-0.036	-0.035	0.004	0.005	-2.2	91.6	-0.036	0.004	0.004	0.3	95.8
D_{22}	0.063	0.062	0.003	0.003	-1.5	89.0	0.063	0.003	0.003	-0.1	94.4

D. Parametric versus non-parametric transition probabilities

To assess the goodness-of-fit of the model to the data, we compared in Section 5 of the main manuscript the parametric estimator of the transition probabilities to a non-parametric estimator. Note that we detail here the methodology which was used. The following results are based on the book of Andersen *et al.* [2].

Preliminaries

Consider the multi-state process $E = \{E(t), t \geq 0\}$ with values in the finite space $S = \{0, 1, \dots, M\}$, where $E(t)$ denotes the state occupied by an individual at time t . We assume that E is a non-homogeneous Markov process, with left truncation and right censoring. In the following, we will consider that all the introduced multi-state processes guarantee the above properties. The intensity of transition from state $h \in S$ to state $k \in S$ at time t is defined as $\lambda_{hk}(t) = \lim_{dt \rightarrow 0} \frac{\Pr(E(t+dt) = k | E(t) = h)}{dt}$ and we write $\boldsymbol{\lambda}(t) = \{\lambda_{hk}(t)\}$ the $(M+1) \times (M+1)$ matrix of transition intensities. The matrix of cumulative transition intensities is noted $\boldsymbol{\Lambda}(t)$, composed of non-diagonal elements $\Lambda_{hk}(t) = \int_0^t \lambda_{hk}(u) du, \forall h \neq k$, and diagonal elements $\Lambda_{hh}(t) = -\sum_{k \neq h} \Lambda_{hk}(t)$. Let us consider the transition probability $P_{hk}(s, t) = \Pr(E(t) = k | E(s) = h)$, with $s \leq t$, which is the probability that a subject in state h at time s occupies the state k at a later time t . We call $\mathbf{P}(s, t) = \{P_{hk}(s, t)\}$ the matrix of transition probabilities, which satisfies the Chapman-Kolmogorov equation:

$$\mathbf{P}(s, t) = \mathbf{P}(s, u)\mathbf{P}(u, t), \text{ with } 0 \leq s \leq u \leq t. \quad (4)$$

$\mathbf{P}(s, t)$ is the unique solution of the Kolmogorov forward differential equations:

$$\begin{aligned} \mathbf{P}(s, s) &= \mathbf{I}, \\ \frac{\partial}{\partial t} \mathbf{P}(s, t) &= \mathbf{P}(s, t)\boldsymbol{\lambda}(t). \end{aligned} \quad (5)$$

Non-parametric estimator

Let $N_{hk}(t)$ be the number of direct observed transitions from state h to state k up to time t , and $Y_h(t)$ the number of individuals in state h just before time t . The non-parametric estimator of the cumulative intensities, called $\boldsymbol{\Lambda}^*(t)$ has elements $\Lambda_{hk}^*(t)$ estimated through the Nelson-Aalen estimator:

$$\Lambda_{hk}^*(t) = \int_0^t \frac{dN_{hk}(u)}{Y_h(u)}, h \neq k, \quad (6)$$

and $\Lambda_{hh}^*(t) = -\sum_{k \neq h} \Lambda_{hk}^*(t)$. The solution to the Kolmogorov equations (5) permits to express the non-parametric estimate of the transition probabilities $\widehat{\mathbf{P}}^*(s, t)$ using the product-integral:

$$\widehat{\mathbf{P}}^*(s, t) = \prod_{(s, t]} \left(\mathbf{I} + d\widehat{\Lambda}^*(u) \right). \quad (7)$$

where $\widehat{\Lambda}^*(u)$ is the non-parametric estimate of the cumulative transition intensities at time u with $d\widehat{\Lambda}_{hh}^*(u) \geq -1$ for all u , and \mathbf{I} is the $(M + 1) \times (M + 1)$ identity matrix. This estimated matrix of transition probabilities is called Aalen-Johansen estimator.

Let $s < T_1^* < \dots < T_{m^*}^* \leq t$ be the ordered times of observed direct transitions between s and t for all individuals. It can be deduced from (7):

$$\widehat{\mathbf{P}}^*(s, t) = \prod_{l=1}^{m^*} \left(\mathbf{I} + \Delta\widehat{\Lambda}^*(T_l^*) \right). \quad (8)$$

where $\Delta\widehat{\Lambda}^*(T_l^*) = \widehat{\Lambda}^*(T_l^*) - \widehat{\Lambda}^*(T_{l-1}^*)$ and $\Delta\widehat{\Lambda}_{hh}^*(T_l^*) \geq -1$ for all T_l^* .

In our application, we applied:

$$\mathbf{I} + \Delta\widehat{\Lambda}^*(T_l^*) = \begin{pmatrix} 1 - \frac{\Delta N_{0.}(T_l^*)}{Y_0(T_l^*)} & \frac{\Delta N_{01}(T_l^*)}{Y_0(T_l^*)} & \frac{\Delta N_{02}(T_l^*)}{Y_0(T_l^*)} & \frac{\Delta N_{03}(T_l^*)}{Y_0(T_l^*)} & \frac{\Delta N_{04}(T_l^*)}{Y_0(T_l^*)} \\ 0 & 1 - \frac{\Delta N_{1.}(T_l^*)}{Y_1(T_l^*)} & \frac{\Delta N_{12}(T_l^*)}{Y_1(T_l^*)} & \frac{\Delta N_{13}(T_l^*)}{Y_1(T_l^*)} & \frac{\Delta N_{14}(T_l^*)}{Y_1(T_l^*)} \\ 0 & 0 & 1 - \frac{\Delta N_{2.}(T_l^*)}{Y_2(T_l^*)} & \frac{\Delta N_{23}(T_l^*)}{Y_2(T_l^*)} & \frac{\Delta N_{24}(T_l^*)}{Y_2(T_l^*)} \\ 0 & 0 & 0 & 1 - \frac{\Delta N_{3.}(T_l^*)}{Y_3(T_l^*)} & \frac{\Delta N_{34}(T_l^*)}{Y_3(T_l^*)} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

where $N_{h.}(T_l^*) = \sum_{k \neq h} N_{hk}(T_l^*)$ and $\Delta N_{hk}(T_l^*) = N_{hk}(T_l^*) - N_{hk}(T_{l-1}^*)$.

Parametric estimator

For each subject $i \in \{1, \dots, N\}$, let us consider that the observed multi-state process is $\{E_i(t), t \geq 0\}$, where $E_i(t)$ denotes the state occupied by subject i at time t and takes values in the finite space $S = \{0, 1, \dots, M\}$. In our joint multi-state model, we were interested in the transitions intensities:

$$\lambda_{hk}^i(t|b_i) = \lim_{dt \rightarrow 0} \frac{\Pr(E_i(t + dt) = k | E_i(t) = h; b_i)}{dt}, \quad (9)$$

which share the random effects b_i with the longitudinal sub-model. Based on the *Preliminaries* paragraph, let $\Lambda^i(t|b_i) = \{\Lambda_{hk}^i(t|b_i)\}$ be the parametric matrix of cumulative intensities for subject i , and $\mathbf{P}_{hk}^i(s, t|b_i) = \{P_{hk}^i(s, t|b_i)\}$ be the parametric matrix of transition probabilities. From the individual covariates measured at baseline $X_{hk,i}^S$ and the parameters estimated by maximum likelihood $\widehat{\theta}$, we computed for each subject i the individual predictions of the transition intensities $\widehat{\lambda}_{hk}^i(t|\widehat{\theta})$ and

deduced the individual predictions of the cumulative intensities $\widehat{\Lambda}_{hk}^i(t|\hat{\theta})$. To obtain the parametric estimator of the individual transition probabilities $\widehat{\mathbf{P}}^i(s, t|\hat{\theta})$, we used that:

$$\begin{aligned}\widehat{P}_{hh}^i(s, t|\hat{\theta}) &= \exp \left\{ \widehat{\Lambda}_{hh}^i(t|\hat{\theta}) - \widehat{\Lambda}_{hh}^i(s|\hat{\theta}) \right\}, \\ \widehat{P}_{hk}^i(s, t|\hat{\theta}) &= \int_s^t \widehat{P}_{hh}^i(s, u|\hat{\theta}) \widehat{\lambda}_{hk}^i(u|\hat{\theta}) \widehat{P}_{kk}^i(u, t|\hat{\theta}) du, h \neq k.\end{aligned}\quad (10)$$

In practice, when the state space S is large, these integrals become too complex numerically. In the application we therefore preferred to calculate $\widehat{\mathbf{P}}^i(s, t|\hat{\theta})$ through the product-integral:

$$\widehat{\mathbf{P}}^i(s, t|\hat{\theta}) = \prod_{(s,t]} \left(\mathbf{I} + d\widehat{\Lambda}^i(u) \right), \quad (11)$$

with $d\widehat{\Lambda}_{hh}^i(u) \geq -1$ for all u . The matrix of transition probabilities $\widehat{\mathbf{P}}(s, t)$ was then deduced by averaging over the N individual predictions:

$$\widehat{\mathbf{P}}(s, t|\hat{\theta}) = \frac{1}{N} \sum_{i=1}^N \widehat{\mathbf{P}}^i(s, t|\hat{\theta}). \quad (12)$$

Covariance estimation of the non-parametric estimator

The covariance matrix of the Aalen-Johansen estimator of transition probabilities can be estimated by the Greenwood-type estimator:

$$\widehat{\text{cov}}(\widehat{\mathbf{P}}^*(s, t)) = \int_s^t \widehat{\mathbf{P}}^*(u, t)^\top \otimes \widehat{\mathbf{P}}^*(s, u-) \widehat{\text{cov}}(d\widehat{\Lambda}^*(u)) \widehat{\mathbf{P}}^*(u, t) \otimes \widehat{\mathbf{P}}^*(s, u-)^\top, \quad (13)$$

where \otimes denotes the Kronecker product and $^\top$ denotes the vector transpose.

Andersen *et al.* [2] simplified the computations in (13) using the recursion formula:

$$\begin{aligned}\widehat{\text{cov}}(\widehat{\mathbf{P}}^*(s, t)) &= \{(\mathbf{I} + \Delta\widehat{\Lambda}^*(t))^\top \otimes \mathbf{I}\} \widehat{\text{cov}}(\widehat{\mathbf{P}}^*(s, t-)) \{(\mathbf{I} + \Delta\widehat{\Lambda}^*(t)) \otimes \mathbf{I}\} + \\ &\quad \{\mathbf{I} \otimes \widehat{\mathbf{P}}^*(s, t-)\} \widehat{\text{cov}}(\Delta\widehat{\Lambda}^*(t)) \{\mathbf{I} \otimes \widehat{\mathbf{P}}^*(s, t-)\},\end{aligned}\quad (14)$$

where

$$\widehat{\text{cov}}(\Delta\widehat{\Lambda}_{hk}^*(t), \Delta\widehat{\Lambda}_{h'k'}^*(t)) = \begin{cases} \frac{(Y_h(t) - \Delta N_h(t)) \Delta N_h(t)}{Y_h(t)^3}, & \text{for } h = k = h' = k', \\ -\frac{(Y_h(t) - \Delta N_h(t)) \Delta N_{hk'}(t)}{Y_h(t)^3}, & \text{for } h = k = h' \neq k', \\ -\frac{(\delta_{kk'} Y_h(t) - \Delta N_{hk}(t)) \Delta N_{hk'}(t)}{Y_h(t)^3}, & \text{for } h = h', h \neq k, h \neq k', \\ 0, & \text{for } h \neq h', \end{cases}$$

with $\delta_{kk'}$ the Kronecker delta. Note that $\widehat{\text{cov}}(\widehat{\mathbf{P}}^*(s, t))$ and $\widehat{\text{cov}}(\Delta\widehat{\Lambda}^*(t))$ are two $(M+1)^2 \times (M+1)^2$ covariance matrices.

These results may be used to construct the 95% pointwise confidence intervals for the elements

of the Aalen-Johansen estimator:

$$\exp \left\{ \log \hat{P}_{hk}^*(s, t) \pm 1.96 \frac{\sqrt{\widehat{\text{var}}(\hat{P}_{hk}^*(s, t))}}{\hat{P}_{hk}^*(s, t)} \right\}.$$

References

- [1] Rizopoulos D. *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press, 2012.
- [2] Andersen PK, Borgan O, Gill RD, Keiding N. *Statistical Models Based On Counting Processes*. Springer, 1993.

4 Score test for residual dependence between transition times in a joint model for a longitudinal marker and a multi-state process

A limitation of the joint multi-state model we proposed is that it relies on the Markov assumption, which assumes that the future of the multi-state process depends only of its current history (i.e. its current state) and not of its past history (i.e. past state(s)). However, in practice there may still subsist a correlation between individual transition times which is not explained by the model's covariates. For example, some subjects may be more at risk than others to experience many transitions, without this explained by their measured prognostic factors.

To capture this residual intra-subject correlation, an individual frailty term could be included in the multi-state submodel of the joint model. An individual frailty term is a random effect specific to the subject which depicts his propensity to be more likely to transit to another state than others subjects with the same characteristics. However, a joint multi-state model with additional frailty term would be very complex to estimate in practice because its likelihood would involve integrals over time, random effects and frailty term. Moreover, interpretation and identifiability problems could arise [Putter and van Houwelingen, 2011].

We thus propose a score test for the inclusion of a Gaussian frailty term in the multi-state submodel of the joint model, which has the advantage to only require the estimation of the model under the null hypothesis, that is when the frailty term is missing. This score test provides a relevant diagnostic tool in joint multi-state models and, more generally, in multi-state models, because beyond the assumption of frailty, it more generally tests the adjustment to the data.

An analytic expression of the score statistic is given, while the asymptotic variance corrected for the nuisance parameters of the model has to be approximated using finite

difference method. The distribution of the test statistic is then defined in an asymptotic way, that is when the sample size is infinite or the number of observed individual transitions is infinite.

The score test for the inclusion of a Gaussian frailty term in a joint multi-state model is implemented in R, using an easy-to-use function which takes as argument the joint multi-state model estimated when the frailty is omitted.

A simulation study is performed to investigate the performances of such a test, notably by varying the number of subjects and / or the number of observed individual transitions.

This work is then applied on the data from the two cohorts of men treated by radiotherapy for localised prostate cancer. From the use of the score test, we define a reliable model which properly quantifies the effect of each classical prognostic factor and the PSA dynamics on each transition intensity between clinical states.

This paper is currently submitted for publication in a statistical journal.

Score test for residual dependence between transition times in a joint model for a longitudinal marker and a multi-state process

Loïc Ferrer^{1,*}, Anaïs Rouanet², H el ene Jacqmin-Gadda¹, C ecile Proust-Lima¹

¹INSERM, UMR1219, Univ. Bordeaux, F-33076 Bordeaux, France

² MRC Biostatistics Unit, University of Cambridge, United Kingdom

Submitted for publication.

SUMMARY. New joint models have recently been proposed to deal with both a longitudinal process and a multi-state process. These models permit to estimate jointly change over time of a longitudinal marker and transition intensities between multiple states. However, they do not consider any individual frailty in the multi-state model although a frailty could capture a residual correlation between individual transition times or reflect a lack of model fit. Since estimation and interpretation of joint models with an individual frailty are not obvious, we propose instead a score test for the need of a Gaussian frailty term in a joint model for a longitudinal process and a multi-state process. This score test requires only the estimation of the model without frailty. We provide an analytic expression of the score statistic and an approximation of its variance corrected for the nuisance parameters by considering joint models with either shared latent classes or shared random effects. The proposed score test, implemented in R, is validated in a simulation study. A detailed application is carried out on a dataset comprising repeated measurements of the prostate specific antigen and transition times between multiple clinical health states for patients with prostate cancer treated by radiotherapy. The score test is used to validate the assumptions of the model and check its goodness-of-fit.

KEYWORDS. Frailty term; Joint model; Longitudinal process; Multi-state process; Prostate cancer.

4.1 Introduction

Joint modelling was initially introduced to simultaneously model two correlated processes: one longitudinal (e.g. a longitudinal Gaussian marker) and one survival (e.g. time-to-event data), by linking them using a shared latent structure. In the literature, two main types of joint models have been developed: joint models with shared latent classes (JLCM) and joint models with shared random effects (SREM). The former assumes that the population is heterogeneous and that the association between the marker's evolution and the risk of event is depicted through latent classes [Proust-Lima et al., 2014], while

the latter considers that the population is homogeneous and that the intensity of the event process depends on the marker process through a function of the shared individual-specific random effects [Tsiatis and Davidian, 2004]. In both joint modelling approaches, many advances have been proposed recently through varied applications in health studies when focusing on event history data. For instance, Elashoff et al. [2008] and Proust-Lima et al. [2016] extended the standard joint modelling to competing events in SREM and JLCM respectively, Dantan et al. [2011] developed a joint multi-state model with a latent state, Król et al. [2016] proposed a joint model to deal with a longitudinal Gaussian marker, recurrent events and a terminal event. Recently, Rouanet et al. [2016] and Ferrer et al. [2016] introduced new joint models to deal with a longitudinal process and a multi-state process. These models permit to analyse multiple transitions between states, with adjustment on marker dynamics. Indeed, in many clinical applications, the focus is not on a unique type of event, but rather on multiple transitions between clinical health states characterized by a multi-state process. Rouanet et al. [2016] studied simultaneously a longitudinal Gaussian marker and an illness-death model handling interval censoring, by introducing a joint multi-state model with latent classes. The application was performed in elderly subjects to study jointly cognitive decline, interval-censored diagnosis of dementia and death. Ferrer et al. [2016] developed a joint multi-state model with shared random effects and applied it on prostate cancer data by considering the correlation between individual repeated measurements of prostate specific antigen and transition times between clinical health states (local recurrence, initiation of hormonal therapy, metastatic recurrence and death).

In both developments, the multi-state process was assumed to be Markovian which means that, after adjusting on the longitudinal marker trajectory (through shared latent classes or random effects), the future of the multi-state process depends only on the present state, not on the past state(s). This strong assumption is questionable and an alternative way to bypass it would be to consider an individual frailty. The frailty is an individual random effect that captures a residual correlation between transition times, that is a correlation between the multiple transition times of a same subject that is not captured by the covariates included in the model nor by the shared random variables in the joint framework.

However, in practice a joint multi-state model with an additional frailty term would be very difficult to estimate. In the maximum likelihood framework, the likelihood would involve a supplementary integral on the distribution of the frailty. Yet, in shared random effects models, the likelihood already requires several numerically intensive approx-

imations of integrals over the individual random effects from the mixed model, and the estimation of latent class models is also complex due the multi-modality of the likelihood. Putter and van Houwelingen [2015] broached the inclusion of individual frailties in standard multi-state models and specifically discussed model identifiability problems and interpretation of the estimated parameters in the presence of a frailty term.

In this context, the aim of this work is to propose and validate a score test to assess the necessity of a Gaussian frailty term in the multi-state sub-model of a joint multi-state model. The score test has the advantage to only require the estimation of the model under the null hypothesis, i.e. without frailty using joint multi-state models previously proposed [Rouanet et al., 2016; Ferrer et al., 2016]. We specifically defined the score test for the variance of a centered Gaussian frailty term in joint multi-state models considering either shared random effects or latent classes approaches.

The outline of the paper is as follows. Section 4.2 introduces the joint multi-state model with a Gaussian frailty. We provide an analytic expression of the score statistic for testing the need for a frailty, and an approximation of its variance corrected for the nuisance parameters, in Section 4.3. Section 4.4 details a simulation study performed to check the performances of the score test in terms of type-I error rate and statistical power under several scenarii. The score test is then applied on prostate cancer data in Section 4.5 with the objective to validate a model to jointly analyse repeated measures of prostate specific antigen and times of transitions between clinical health states. The paper ends with a discussion in Section 4.6.

4.2 The joint multi-state frailty model

For each subject $i \in \{1, \dots, N\}$, we observe the continuous non-homogeneous process $E_i(t) = \{E_i(t), T_{i0} \leq t \leq C_i\}$ with values $E_i(t)$ in the finite state space $S = \{0, \dots, M\}$. The multi-state process is observed between T_{i0} the left truncation time and C_i the right censoring time. The vector of observed transition time(s) is $T_i = (T_{i1}, \dots, T_{im_i})^\top$ with $T_{ir} < T_{i(r+1)}$ for $r \in \{0, \dots, m_i - 1\}$ and $m_i \geq 1$, and the associated vector of observed transition indicator(s) is $\delta_i = (\delta_{i1}, \dots, \delta_{im_i})^\top$, with $\delta_{i(r+1)}$ equals 1 if a direct transition is observed at $T_{i(r+1)}$ (i.e. $E_i(T_{ir}) \neq E_i(T_{i(r+1)})$) and 0 otherwise. Finally, $T_{im_i} = C_i$ if $\delta_{im_i} = 0$. For each subject i , we also collect the vector of longitudinal Gaussian observations $Y_i = (Y_i(t_{i1}), \dots, Y_i(t_{in_i}))^\top$, with the observation times $\{t_{ij}; j = 1, \dots, n_i\}$ such that $t_{in_i} \leq T_{im_i}$.

4.2.1 Formulation

The joint multi-state model has been previously developed with shared latent classes or shared random effects. For the sake of clarity, we only detail in the main text the score test for a Gaussian frailty term in a joint model with shared random effects. The formulations in the context of a joint model with shared latent classes are given in Supplementary Material, Section 4.9.1.

The joint multi-state frailty model links two sub-models using a function of the shared random effects b_i . The longitudinal evolution of the Gaussian marker is modelled using a linear mixed model while the transition times between states are modelled using proportional intensity models that include functions of the random effects from the mixed model as covariates. An additional Gaussian frailty v_i is added in the multi-state sub-model to capture the residual correlation between individual transition times which is not explained by the covariates:

$$\begin{cases} Y_i(t) &= Y_i^*(t) + \epsilon_i(t) \\ &= X_i^L(t)^\top \beta + Z_i(t)^\top b_i + \epsilon_i(t) \\ \lambda_{hk}^i(t) &= \lambda_{hk,0}(t) \exp(X_{hk,i}^E{}^\top \gamma_{hk} + W_{hk,i}(t|b_i)^\top \eta_{hk} + v_i) \end{cases} \quad (4.1)$$

where $X_i^L(t)$ and $Z_i(t)$ are vectors of possibly time-dependent covariates respectively associated with the vector of fixed effects β and the vector of random effects b_i , where $b_i \sim \mathcal{N}_q(0, D)$. The error terms are independent and identically distributed (iid) such that $\epsilon_i(t) \sim \mathcal{N}(0, \sigma_\epsilon^2)$; b_i and $\epsilon_i(t)$ are independent. The transition intensity $\lambda_{hk}^i(t)$ is defined for any $(h, k) \in S^2$; $\lambda_{hk,0}(t)$ is a parametric baseline intensity; $X_{hk,i}^E$ is the vector of prognostic factors associated with the vector of coefficients γ_{hk} , $W_{hk,i}(t|b_i)$ is a multivariate function of random effects and time that depicts the dependence between the longitudinal marker and each transition intensity between states h and k . In practice, the current marker dynamics are usually chosen, for example $W_{hk,i}(t|b_i) = Y_i^*(t)$ the true current level of the marker, or $W_{hk,i}(t|b_i) = \partial Y_i^*(t)/\partial t$ the true current slope of the marker, or both $W_{hk,i}(t|b_i) = (Y_i^*(t), \partial Y_i^*(t)/\partial t)^\top$. The frailty term v_i , which aims to capture some heterogeneity unexplained by the prognostic factors and the longitudinal marker, is assumed to be Gaussian with mean zero and unknown variance σ_v^2 . Conditionally to the frailty and the marker dynamics, the multi-state process is Markovian.

4.2.2 Estimation

The model (4.1) can be estimated by maximizing the likelihood function $L(\sigma_v^2, \theta)$ where the set of parameters $\{\sigma_v^2, \theta\}$ is to estimate, with θ all the parameters contained in (4.1) except σ_v^2 . Using the independence assumption between the individual longitudinal and multi-state processes conditionally to the individual random effects b_i , and the independence between the individual transition times given the individual frailty v_i , the likelihood can be written as follows:

$$L(\sigma_v^2, \theta) = \prod_{i=1}^N \int_{\mathbb{R}^q} f_Y(Y_i|b_i; \theta) f_b(b_i; \theta) \int_{\mathbb{R}} f_E(E_i|v_i, b_i; \sigma_v^2, \theta) f_v(v_i; \sigma_v^2) dv_i db_i \quad (4.2)$$

where $f(\cdot)$ is a probability density function.

The conditional density of the longitudinal outcome is such that

$$f_Y(Y_i|b_i; \theta) = \frac{1}{(2\pi\sigma_e^2)^{n_i/2}} \exp\left(-\frac{\|Y_i - X_i^L\beta - Z_i b_i\|^2}{2\sigma_e^2}\right)$$

where X_i^L and Z_i are the design matrices with row vectors $X_i^L(t_{ij})$ and $Z_i(t_{ij})$ ($j = 1, \dots, n_i$), respectively, and $\|x\|$ denotes the Euclidean norm of vector x .

The random effects b_i follow a multivariate Gaussian distribution such that

$$f_b(b_i; \theta) = \frac{1}{(2\pi)^{q/2} |D|^{1/2}} \exp\left(-\frac{b_i^\top D^{-1} b_i}{2}\right)$$

with $|D|$ the determinant of matrix D .

The conditional density of the observations of the multi-state process is

$$f_E(E_i|v_i, b_i; \sigma_v^2, \theta) = \prod_{r=0}^{m_i-1} \left[P_{E_i(T_{ir}), E_i(T_{ir})}^i(T_{ir}, T_{i(r+1)}|v_i, b_i) \times \lambda_{E_i(T_{ir}), E_i(T_{i(r+1)})}^i(T_{i(r+1)}|v_i, b_i)^{\delta_{i(r+1)}} \right]$$

with $P_{hh}(s, t) = \exp\left(\int_s^t \lambda_{hh}(u) du\right) = \exp\left(-\sum_{k, k \neq h} \int_s^t \lambda_{hk}(u) du\right)$.

The frailty term v_i follows a Gaussian distribution with expectation 0 and variance σ_v^2 :

$$f_v(v_i; \sigma_v^2) = \frac{1}{(2\pi\sigma_v^2)^{1/2}} \exp\left(-\frac{v_i^2}{2\sigma_v^2}\right)$$

The likelihood in (4.2) involves three sets of integrals (over time, over the individual multivariate random effects, and over the frailty term) which can be numerically approximated using Gaussian quadratures while making sure the number of quadrature points is large enough to ensure correct inference [Ferrer et al., 2016]. The logarithm of the

likelihood can then be optimized using for instance an EM algorithm coupled with a quasi-Newton after a fixed number of iterations as proposed in standard joint models with shared random effects [Rizopoulos, 2012b]. It should be noted however that such estimation would be very complex to achieve in practice.

4.3 Score test for residual dependence between individual transition times

The nullity of the variance of the frailty term can be tested from the model defined in (4.1) using a one-sided score test. Under the null hypothesis, the variance of the Gaussian frailty is null and the frailty equals 0, whereas under the alternative hypothesis, the variance of the frailty is strictly positive and the frailty has an impact on each transition intensity:

$$H_0 : \sigma_v^2 = 0 \quad \text{vs} \quad H_1 : \sigma_v^2 > 0 \quad (4.3)$$

We thus propose a score test to assess this assumption, the score test having the advantage to only require the estimation of the model under the null hypothesis, that is when the frailty term is missing.

4.3.1 Score statistic

The score statistic $U(0, \theta) = U(\sigma_v^2 = 0, \theta)$ is the gradient of the log-likelihood with respect to σ_v^2 , computed under the null hypothesis: $U(0, \theta) = \frac{\partial \log L(\sigma_v^2, \theta)}{\partial \sigma_v^2} \Big|_{\sigma_v^2=0} = \sum_{i=1}^N \frac{\partial \log L_i(\sigma_v^2, \theta)}{\partial \sigma_v^2} \Big|_{\sigma_v^2=0}$. The following analytic expression can be derived in the context of a joint model with shared random effects:

$$U(0, \theta) = \sum_{i=1}^N \frac{1}{2L_i(\sigma_v^2 = 0, \theta)} \times \int_{\mathbb{R}^q} f_Y(Y_i|b_i; \theta) f_E(E_i|b_i; \sigma_v^2 = 0, \theta) f_b(b_i; \theta) \times \left\{ \left[\sum_{r=0}^{m_i-1} \left(\delta_{i(r+1)} + \Lambda_{E_i(T_{ir}), E_i(T_{ir})}^i(T_{ir}, T_{i(r+1)}|b_i; \sigma_v^2 = 0, \theta) \right) \right]^2 + \sum_{r=0}^{m_i-1} \Lambda_{E_i(T_{ir}), E_i(T_{ir})}^i(T_{ir}, T_{i(r+1)}|b_i; \sigma_v^2 = 0, \theta) \right\} db_i \quad (4.4)$$

where $\Lambda_{hk}^i(t) = \int_0^t \lambda_{hk}^i(u) du$. The detailed calculations leading to this expression are given in Section 4.9.1 of the Supplementary Material. The same calculations for a joint latent

class model are given in Section 4.9.1 of the Supplementary Material.

Asymptotically, the score statistic follows a normal distribution: $U(0, \theta) \sim \mathcal{N} \left(0, \text{Var} \{U(0, \theta)\} \right)$.

4.3.2 Variance of the score statistic

The asymptotic variance of the score statistic is approximated from the observed Fisher information matrix of model (4.1) defined as $I(\sigma_v^2, \theta) = \begin{pmatrix} I_{\sigma_v^2 \sigma_v^2} & I_{\sigma_v^2 \theta} \\ I_{\theta \sigma_v^2} & I_{\theta \theta} \end{pmatrix}$.

The variance of the score statistic corrected for the estimation of the nuisance parameters θ , is given by

$$\text{Var} \{U(0, \theta)\} = \left(I_{\sigma_v^2 \sigma_v^2} - I_{\sigma_v^2 \theta} I_{\theta \theta}^{-1} I_{\theta \sigma_v^2} \right) \Big|_{\sigma_v^2=0} \quad (4.5)$$

where $I_{\theta \theta}$ for $\sigma_v^2 = 0$ is estimated from the null model, while $I_{\theta \sigma_v^2}$ and $I_{\sigma_v^2 \sigma_v^2}$ given $\sigma_v^2 = 0$ are computed by the forward finite difference method because no analytic expression was found.

4.3.3 Test statistic

Once the score statistic and the associated variance are computed by replacing θ by $\hat{\theta}_0$, its maximum likelihood estimate under the null hypothesis H_0 , the test statistic can be deduced:

$$T = \frac{\{U(0, \hat{\theta}_0)\}^2}{\text{Var} \{U(0, \hat{\theta}_0)\}} - \inf \left[\frac{\{U(0, \hat{\theta}_0) - \delta\}^2}{\text{Var} \{U(0, \hat{\theta}_0)\}} : \delta > 0 \right] \quad (4.6)$$

and T follows asymptotically a mixture of chi-square distributions: $T \sim \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$.

This expression of the test statistic T was discussed in Verbeke and Molenberghs [2003]. The authors argued that when the null hypothesis of a score test is on the boundary of the parameter space, the test statistic has to consider a correction at the boundary. Thus, the test statistic is null when the score statistic is negative whereas it has the same expression as in the two-sided case (the infimum equals 0) when the score statistic is positive. The score statistic being asymptotically centred, the nullity of the test statistic is asymptotically obtained with probability 1/2, and T follows asymptotically a mixture of chi-square distributions with probabilities 1/2.

4.4 Simulation study

We evaluated the performances of this score test in terms of type-I error rate and statistical power in a simulation study. We considered two scenarii in order to investigate the behaviour of the score test according to different settings of population study with varying numbers of observed transitions:

- in scenario 1, depicted in Figure 4.1, the model considered 3 states ($h, k \in S = \{0, 1, 2\}$) and 3 transitions, as in a unidirectional illness-death model [Andersen and Keiding, 2002];
- in scenario 2, depicted in Figure 4.2, the model considered 5 states ($h, k \in S = \{0, 1, 2, 3, 4\}$) and 10 transitions.

These scenarii notably differed by their average number of observed direct transitions per subject, denoted \bar{M} in the remainder of the paper.

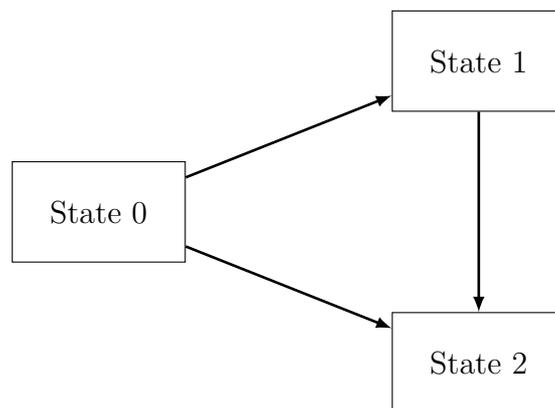


Figure 4.1 – Multi-state representation of scenario 1. Arrows indicate the directions of the possible transitions.

In each scenario, the type-I error rate was investigated under the null hypothesis $\sigma_v^2 = 0$ and the statistical power was investigated under two alternative hypotheses: $\sigma_v^2 = 0.5$ and $\sigma_v^2 = 1$. Each time, 500 replicates of 500, 1000 or 1500 subjects were drawn.

4.4.1 Model for data generation

Data were generated from a joint multi-state frailty model with shared random effects, using the procedures previously described in Crowther and Lambert [2013] and Ferrer et al. [2016] which were extended by including a Gaussian frailty term in the multi-state sub-model. For each subject $i \in 1, \dots, N$, we assumed a linear evolution of the longitudinal marker over time depending on the level of a continuous covariate $X_i \sim \mathcal{N}(2.04, 0.50)$.

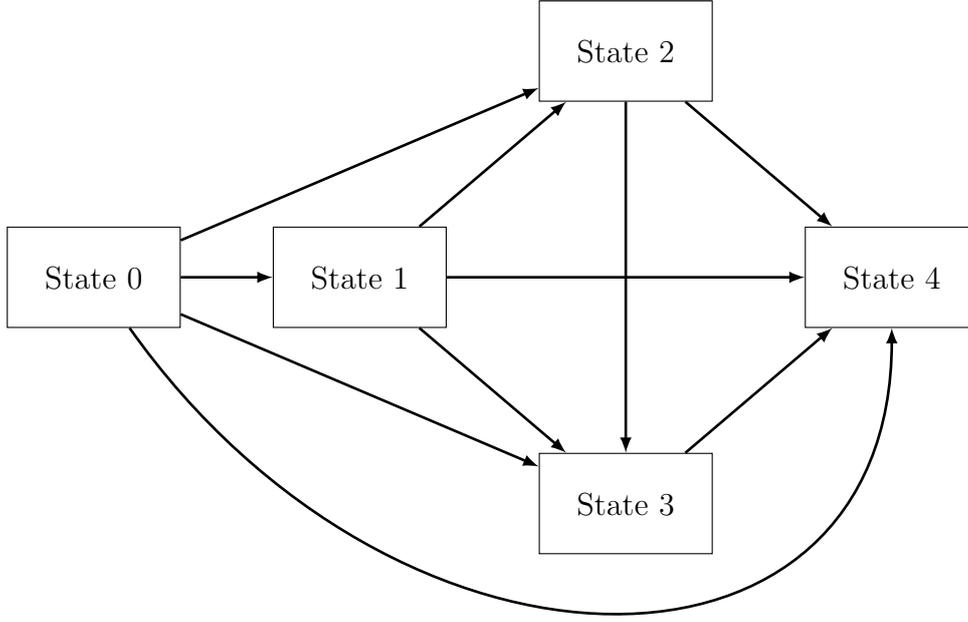


Figure 4.2 – Multi-state representation of scenario 2. Arrows indicate the directions of the possible transitions.

The same covariate X_i was included in the transition intensity along with an effect of both the current level and the current slope of the marker:

$$\left\{ \begin{array}{l} Y_i(t) = Y_i^*(t) + \epsilon_i(t) \\ \quad = (\beta_0 + \beta_{0,X}X_i + b_{i0}) + (\beta_1 + \beta_{1,X}X_i + b_{i1}) \times t + \epsilon_i(t) \\ \lambda_{hk}^i(t) = \lambda_{hk,0}(t) \exp(\gamma_{hk}X_i + \eta_{hk,\text{level}}Y_i^*(t) + \eta_{hk,\text{slope}}\partial Y_i^*(t)/\partial t + v_i) \end{array} \right.$$

In scenario 1, we assumed baseline transition-specific intensities $\lambda_{hk,0}(t)$ generated as linear combinations of cubic B-splines with the same vector of knots $(0.01, 4.14, 7.46, 10.91, 18.20)^\top$ for all transitions. In scenario 2, the baseline transition intensities were proportional from one state to another: $\lambda_{hk,0}(t) = \lambda_0(t) \exp(\zeta_{hk})$ with $\lambda_0(t)$ generated as a linear combination of cubic B-splines with knot vector $(0.03, 7.13, 18.20)^\top$. The longitudinal observations $Y_i = (Y_i(t_{i1}), \dots, Y_i(t_{in_i}))^\top$ were generated at observation times $t_{ij} = 0, 0.33, 0.66, \dots$ and the vector of independent error terms $\epsilon_i = (\epsilon_i(t_{i1}), \dots, \epsilon_i(t_{in_i}))^\top$ was such that $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$. The vector of random effects $b_i = (b_{i0}, b_{i1})^\top$ was drawn from a Gaussian distribution with mean $(0, 0)^\top$ and variance D . The frailty term was generated from a Gaussian distribution with mean zero and variance σ_v^2 , with $\sigma_v^2 \in \{0, 0.5, 1\}$ depending on the studied case. The censoring time was systematically drawn from a uniform distribution on $[1, 22]$. The generating parameters of each scenario are given in Section 4.9.2 of the Supplementary Material.

4.4.2 Model estimation

All the models were estimated while ignoring the frailty, i.e. by using a joint multi-state model with shared random effects [Ferrer et al., 2016]. To achieve accurate point estimates and correct estimation of the corresponding variances, the estimation of shared random effects models requires careful attention to the technique of approximation of the integral over the random effects b_i . In this work, we systematically considered a two-step pseudo adaptive Gauss-Hermite quadrature, proposed in Ferrer et al. [2016], which was shown to provide more accurate estimations than the one-step pseudo adaptive Gauss-Hermite quadrature. In scenario 1 with three states and three transitions, we considered two successive pseudo adaptive Gauss-Hermite quadratures with 5 and 5 points of quadrature. In scenario 2 with five states and ten transitions, we considered successively 9 and 9 points of quadrature as estimations were not accurate enough with a lower number of points.

4.4.3 Type-I error

We evaluated the empirical type-I error rate of the score test for a nominal level of 5% over the 500 replicates of the simulation study. Results are depicted in Table 4.1 for scenarii 1 and 2 and three different population sizes. The type-I error rates were too low when the population was small (500 subjects) or when the number of transitions was small (0.70 transition per subject on average). With an average of 2.84 transitions by subject and at least 1000 subjects in the sample, the type-I error rates were correct.

Table 4.1 – Empirical type-I error of the score test for the inclusion of a frailty in a joint multi-state model under both scenarii. \bar{M} denotes the average number of observed direct transitions per subject.

	Scenario 1	Scenario 2
	$\sigma_v^2 = 0$	$\sigma_v^2 = 0$
	$(\bar{M} = 0.70)$	$(\bar{M} = 2.84)$
$N = 500$	0.008	0.028
$N = 1000$	0.010	0.054
$N = 1500$	0.020	0.060

4.4.4 Power

We computed the empirical statistical power of the score test with a nominal level of 5% over the 500 replicates of the simulation study, considering either $\sigma_v^2 = 0.5$ and $\sigma_v^2 = 1$. Results are depicted in Table 4.2 for scenarii 1 and 2 and the three population sizes. In scenario 2, in which subjects had on average about 2.7 transitions, the statistical power was almost perfect for the three population sizes (500, 1000 or 1500 subjects) and the two considered values for the variance. In contrast, in scenario 1, where subjects had a mean of only approximately 0.75 transitions, statistical power was above 85% only when considering a population of 1500 subjects or a population of 1000 subjects with a variance of the frailty equal to 1.

Table 4.2 – Empirical statistical power of the score test for the inclusion of a frailty in a joint multi-state model under both scenarii. \bar{M} denotes the average number of observed direct transitions per subject.

	Scenario 1		Scenario 2	
	$\sigma_v^2 = 0.5$	$\sigma_v^2 = 1$	$\sigma_v^2 = 0.5$	$\sigma_v^2 = 1$
	($\bar{M} = 0.74$)	($\bar{M} = 0.75$)	($\bar{M} = 2.73$)	($\bar{M} = 2.65$)
$N = 500$	0.278	0.438	0.884	0.990
$N = 1000$	0.568	0.850	0.998	1.000
$N = 1500$	0.846	0.970	1.000	1.000

4.5 Application on prostate cancer data

We analyzed data from 1474 men with a localized prostate cancer treated by external beam radiotherapy (EBRT): 629 patients were from the multi-center RTOG 9406 (Radiation Therapy Oncology Group, USA), and 845 patients were from the cohort of the British Columbia Cancer Agency (BCCA) in Vancouver, with examinations performed from 1994 to 2013, and from 1994 to 2012, respectively. After EBRT, subjects were followed up and the prostate-specific antigen (PSA), which is a well-known biomarker in localized prostate cancer, was repeatedly measured. Several authors have already shown that the PSA dynamics are substantially associated with the risk of post-treatment recurrence [Proust-Lima et al., 2008; Taylor et al., 2013]. Ferrer et al. [2016] extended the analysis on the same data by distinguishing several types of recurrence and death, and by

focusing on the intensities of transitions between the multiple clinical health states. We used the same approach here. The considered states and direct transitions correspond to those depicted in Figure 4.2, with 0, 1, 2, 3 and 4 corresponding to end of EBRT, local recurrence, initiation of a hormonal therapy (additional treatment unplanned at baseline), distant recurrence and death, respectively. The average number of observed direct transitions per subject was 0.92 (with standard deviation 0.88). Further details on the cohorts and data structure are given in Ferrer et al. [2016].

The PSA measurements were collected until the occurrence of the first event. In the joint model, they were log-transformed in order to satisfy the normality assumption of the longitudinal responses. The considered baseline prognostic factors were the pre-radiotherapy PSA level in the log-scale, T-stage which characterizes tumour size (three categories were considered: 2; 3–4 *versus* 1 as reference) and Gleason score which measures the aggressiveness of cancer cells (three categories: 7, 8–10 *versus* 2–6 as reference). A cohort-specific covariate was also considered, coded 1 for RTOG 9406 and -1 for BCCA.

4.5.1 Initial model

We first applied the score test for the inclusion of a Gaussian frailty term in the model proposed in Ferrer et al. [2016]. In this model, three functions of time were considered to fit the biphasic shape of post-therapy PSA trajectory. The baseline intensities of transitions were approximated by cubic B-splines (three internal knots were placed at the quantiles of the observed transition times, and the first external knot was placed at 1 year to take into account the null risk of recurrence before 1 year). The effects of the true current level and the true current slope of PSA were considered on each transition intensity. Finally, covariates were selected in a backward selection procedure both in the longitudinal and multi-state sub-models considered independently. Based on univariate or multivariate Wald tests, covariates with p -value < 0.1 were selected. Based on the model estimates, the null hypothesis for the null variance of a Gaussian frailty term in the joint multi-state model was strongly rejected with p -value equal to 0.007. This result could reflect either a lack of goodness-of-fit or an actual residual correlation between times-to-events, given the random effects. To confirm one of these two options, we tried to improve the goodness-of-fit of the model in a second step.

4.5.2 Improved model

We improved the joint multi-state model by including the age at baseline as covariate and replacing the true PSA level in the multi-state model by a transformation of it so that the considered association function became $W_{hk,i}(t|b_i) = (g(Y_i^*(t)), \partial Y_i^*(t)/\partial t)^\top$ where $Y_i^*(t)$ is the true current level of PSA and $g(Y_i^*(t)) = \text{logit}^{-1}((Y_i^*(t) - 0.71)/0.44)$. This function had been previously found to better capture the effect of current PSA on the risk of recurrence [Proust-Lima et al., 2008], and to satisfy the log-linearity assumption by avoiding very high levels of PSA extrapolated from the longitudinal model in the event history model [Sène et al., 2014]. The covariates selection was done using the same procedure used for the previous model. The quality of the model was significantly improved with an Akaike information criterion (AIC) decreasing from 25574.2 to 25325.3 points. However the score test still rejected the null hypothesis of no frailty with p -value = 0.015.

By examining the individual contributions to the estimated score statistic (Figure 4.3), we identified four subjects that could be considered as outliers (with an arbitrary cut-off at 5 points). Although these particular subjects had a bad prognosis according to T-stage, Gleason and initial PSA, they died very late (15.7, 16.2, 17.2 and 18.2 years after the end of EBRT) without experiencing any type of clinical progression. For the last three subjects, the last time of PSA measurement was 3.1, 1.5 and 1.5 years after the end of EBRT. For the first one, the last PSA measurement occurred 11.0 years after the end of EBRT but his PSA values were extremely high.

4.5.3 Final model

We removed these four subjects and re-ran the strategy of covariate selection. The final selected joint multi-state model was

$$\left\{ \begin{array}{l} Y_i(t) = Y_i^*(t) + \epsilon_i(t) \\ = (\beta_0 + X_i^{L0 \top} \beta_{0,\text{cov}} + b_{i0}) + \\ (\beta_1 + X_i^{L1 \top} \beta_{1,\text{cov}} + b_{i1}) \times f_1(t) + \\ (\beta_2 + X_i^{L2 \top} \beta_{2,\text{cov}} + b_{i2}) \times f_2(t) + \epsilon_i(t) \\ \lambda_{hk}^i(t) = \lambda_{hk,0}(t) \exp \left\{ X_{hk,i}^{E \top} \gamma_{hk} + \begin{pmatrix} g(Y_i^*(t)) \\ \partial Y_i^*(t)/\partial t \end{pmatrix} \begin{pmatrix} \eta_{hk,\text{level}} \\ \eta_{hk,\text{slope}} \end{pmatrix} \right\} \end{array} \right.$$

where the random effects are $b_i = (b_{i0}, b_{i1}, b_{i2})^\top \sim \mathcal{N}_3(0, D)$, D unstructured, and the error terms are iid such that $\epsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$; b_i and $\epsilon_i(t)$ are independent. The three

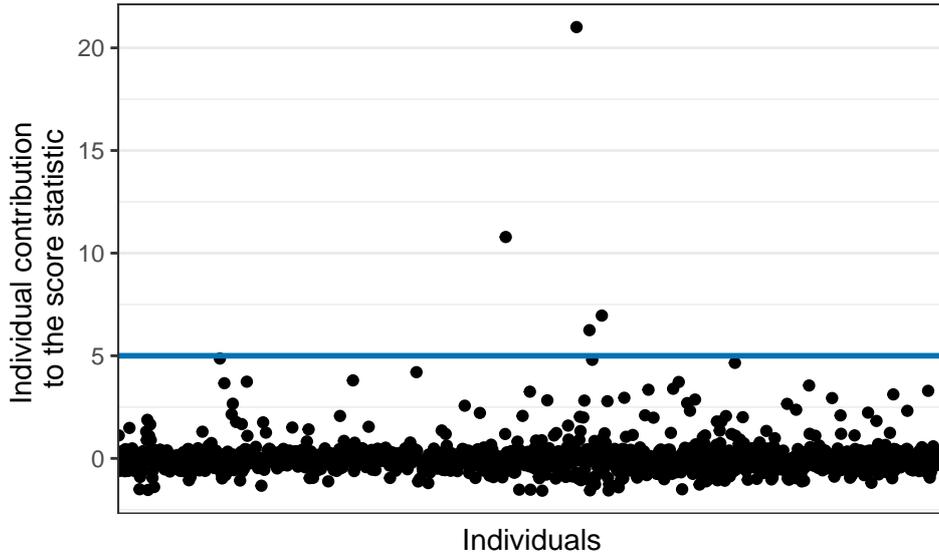


Figure 4.3 – Individual contributions to the score statistic from the improved joint model with $N = 1474$ patients.

functions of time (intercept, $f_1(t) = (1 + t)^{-1.2} - 1$ and $f_2(t) = t$) depicted the biphasic trajectory of the PSA post-EBRT. In the multi-state part, some baseline transition intensities were considered as proportional, as done in the previous applications, with $\lambda_{01,0}(t) = \exp(-\zeta_{02})\lambda_{02,0}(t) = \exp(-\zeta_{12})\lambda_{12,0}(t)$ and $\lambda_{03,0}(t) = \exp(-\zeta_{13})\lambda_{13,0}(t) = \exp(-\zeta_{23})\lambda_{23,0}(t)$, where $\{\zeta_{hk}\}_{hk}$ are unknown proportionality parameters; $\lambda_{04,0}(t)$ was stratified on the cohort. Each considered log baseline intensity was approximated by a linear combination of cubic B-splines with three internal knots placed at the quantiles of the observed times for the considered transition. The selected covariates $X_i^{L0}, X_i^{L1}, X_i^{L2}$ and $\{X_{hk,i}^E\}_{h,k}$ are given in Table 4.4 of the Supplementary Material.

The score test for the inclusion of an additional Gaussian frailty term provided a p -value equals to 0.046, leading us to consider that the results of the model were almost acceptable. Figure 4.4 displays the individual contributions to the score statistic from this final joint model. Contrary to Figure 4.3, Figure 4.4 does not highlight clear outlier.

4.5.4 Interpretation of the final model

The estimates of the association parameters $\{\eta_{hk}\}_{h,k}$ in the final joint multi-state model are given in Table 4.3. For the sake of readability, only the association parameters with the PSA dynamics (current level and current slope) are shown. Other estimates are given in Table 4.4 of the Supplementary Material.

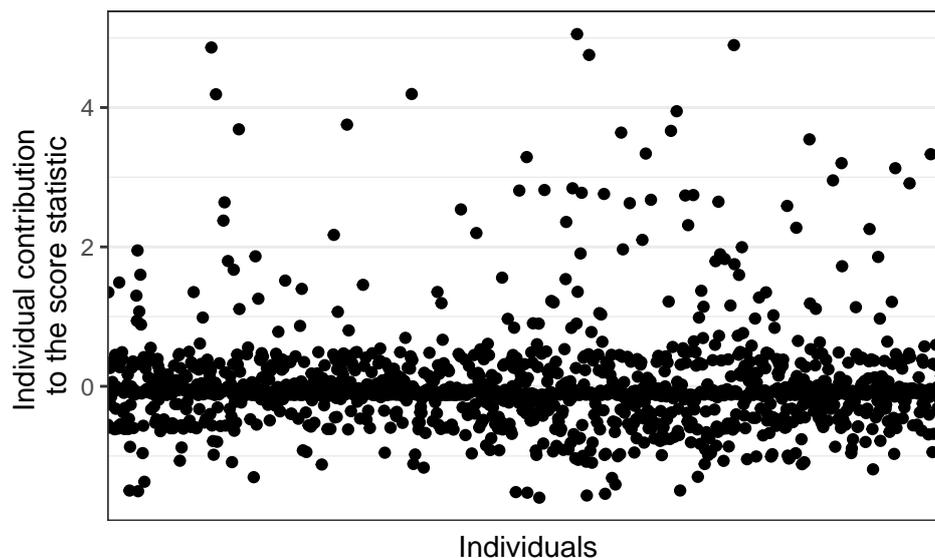


Figure 4.4 – Individual contributions to the score statistic from the final joint model with $N = 1470$ patients.

The results confirmed the strong deleterious impact of PSA increase on the hazard of any type of first recurrence from the end of radiotherapy. However, the PSA level and slope were not significantly associated with the risk of death without recurrence after EBRT. Indeed the corresponding patients did not die from prostate cancer, otherwise they would have experienced a relapse before. This last point was not in agreement with the results from the initial model [Ferrer et al., 2016] which highlighted an unexpected protective effect of the true current PSA level on the direct hazard of death after EBRT. It was clearly due to the outlier subjects identified by the score test. These outlier subjects died very late in the follow-up, certainly due to comorbidities, and with a very short follow-up in the PSA measurements which prevented a correct modelling of their prognosis. For the transition intensity from local recurrence to hormonotherapy states, we found a significant deleterious effect of the true current marker level. Finally, the true current slope of PSA had also a significant deleterious impact on the hazard to transit from local recurrence to distant recurrence states, or from hormonal therapy to distant recurrence states. Overall, these results confirmed that increasing PSA level or slope had a strong deleterious impact on the risk of recurrence post-radiotherapy, but were not necessarily associated with a negative evolution of the disease in advanced stages.

Table 4.3 – Estimates, standard errors and p -values of the association parameters between PSA dynamics and each transition intensity in the joint multi-state model, based on the pooled data after removal of four outliers ($N = 1470$).

	Value	StdErr	p -value
$\eta_{01,\text{level}}$	3.32	0.41	< 0.001
$\eta_{02,\text{level}}$	4.89	0.39	< 0.001
$\eta_{03,\text{level}}$	2.94	0.68	< 0.001
$\eta_{04,\text{level}}$	-0.41	0.23	0.071
$\eta_{12,\text{level}}$	1.90	0.83	0.023
$\eta_{13,\text{level}}$	-2.30	1.32	0.081
$\eta_{14,\text{level}}$	-0.07	0.88	0.939
$\eta_{23,\text{level}}$	-0.29	1.04	0.778
$\eta_{24,\text{level}}$	-0.48	0.62	0.440
$\eta_{34,\text{level}}$	-0.02	0.57	0.974
$\eta_{01,\text{slope}}$	1.33	0.33	< 0.001
$\eta_{02,\text{slope}}$	1.60	0.24	< 0.001
$\eta_{03,\text{slope}}$	1.74	0.54	0.001
$\eta_{04,\text{slope}}$	0.59	0.35	0.088
$\eta_{12,\text{slope}}$	0.46	0.58	0.336
$\eta_{13,\text{slope}}$	3.82	1.07	< 0.001
$\eta_{14,\text{slope}}$	0.70	1.02	0.495
$\eta_{23,\text{slope}}$	0.23	0.51	0.651
$\eta_{24,\text{slope}}$	0.64	0.23	0.005
$\eta_{34,\text{slope}}$	-0.56	0.42	0.186

4.6 Discussion

In many biometrical applications, the focus should not be on the occurrence of the first event of clinical interest, but rather on transitions between clinical health states which can be characterized by a multi-state process. Similarly, in the presence of longitudinal markers, a joint model including a multi-state sub-model should generally be preferred to standard joint models for a unique right censored time to event or several competing causes.

However, when handling multi-state data, one could wonder whether the Markov assumption holds or whether a frailty term should be included in order to capture the residual correlation between individual transition times. In this contribution, we proposed a score test to assess the need to consider a Gaussian frailty term in a joint multi-state model. As pointed out by Putter and van Houwelingen [2015] in the context of multi-state models (without a longitudinal marker), we do not know in practice if such a frailty explains the correlation between individual transition times conditional on covariates or if it is a sort of error term that reflects a lack of fit in the model, possibly due to the violation of the proportional hazards assumption. Thus the score test we propose does not only check the Markov assumption of multi-state models without frailties, but also provides an overall goodness-of-fit tool, as illustrated in the application with the detection of four outliers.

We provided an analytic expression of the score statistic, but we had to numerically derive the variance of the score statistic. The empirical variance, as developed in Jacqmin-Gadda *and others* [2010] and recommended by Freedman [2007], could not be used because the observed transition times were not independent and identically distributed. We thus approximated the asymptotic variance corrected for the nuisance parameters of the model using a forward finite difference method.

The score test was validated using an extensive simulation study. It highlighted the need to have large samples (i.e. a substantial number of subjects and a substantial average number of transitions per subject) to obtain good performances of the score test, in particular a correct type-I error. These results confirmed the previous works carried out on the score test for a frailty term in the context of recurrent events [Sinha, 2013] or on the likelihood ratio test for variance components in mixed models [Crainiceanu and Ruppert, 2004]. However, one could argue that for a diagnostic test, a low type-I error rate (around 2% for instance) is not necessarily unacceptable as the main objective is to detect a lack of fit. The low type-I error rates found in cases where the number of

subjects and/or the number of transitions by individual are small are due to a point mass at zero more important than expected [Crainiceanu and Ruppert, 2004]. To correct for this, some authors proposed permutation techniques [Sinha, 2013]. However, they were not applicable in our context since transition times are not independently distributed.

Although the score test development was detailed in the main manuscript for a joint model with shared random effects, the methodology also applies to a joint model with shared latent classes. Calculations are given in Section 4.9.1 of the Supplementary Material: the structure of the score $U(0, \theta)$ remains the same, but the integral over the random effects is replaced by a weighted sum over the classes. Indeed, the joint model with latent classes can be seen as a joint model with shared random effects that are discrete. Furthermore the methodology also applies to a standard multi-state model that does not consider any longitudinal marker: in that case, the integral over the random effects in equation (4.4) disappears.

We showed that the score test we proposed allows not only to capture a residual correlation between individual transition times, but also reflects a lack of fit of the model. In practice, this score test should be used in addition to other analyses of goodness-of-fit, for example using simple residuals diagnostic plots [Dobson and Henderson, 2003; Rizopoulos et al., 2010] or by comparing observations and model's predictions [Ferrer et al., 2016].

4.7 Software

The joint multi-state model with shared random effects and the associated score test for the inclusion of a frailty are implemented in R. Extensions to deal with a transformation of the current marker's level as prognostic factor are also included. The R functions with examples of application can be found at <https://github.com/LoicFerrer/JMstateModel/> for practical use.

Acknowledgements

We warmly thank James Dignam (Department of Public Health Sciences, University of Chicago and NRG Oncology, Chicago, U.S.A.) and Tom Pickles (Department of Radiation Oncology, University of British Columbia, Vancouver, Canada) for sharing their data on prostate cancer progression. This work was supported by a joint grant from INSERM and Région Aquitaine. Computer time for this study was provided by the computing facilities

MCIA (Mésocentre de Calcul Intensif Aquitain) of the Université de Bordeaux and of the Université de Pau et des Pays de l'Adour.

4.8 References

- ANDERSEN, PER KRAGH AND KEIDING, NIELS. Multi-state models for event history analysis. *Statistical Methods in Medical Research* 11(2), 91–115, 2002. 53, 112
- CRAINICEANU, CIPRIAN M AND RUPPERT, DAVID. Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66(1), 165–185, 2004. 121, 122
- CROWTHER, MICHAEL J AND LAMBERT, PAUL C. Simulating biologically plausible complex survival data. *Statistics in Medicine* 32(23), 4118–4134, 2013. 112, 145
- DANTAN, E, JOLY, P, DARTIGUES, J-F AND JACQMIN-GADDA, H. Joint model with latent state for longitudinal and multistate data. *Biostatistics* 12(4), 723–736, 2011. 35, 67, 106
- ELASHOFF, ROBERT M, LI, GANG AND LI, NING. A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics* 64(3), 762–771, 2008. 35, 67, 106, 201
- FERRER, LOÏC, RONDEAU, VIRGINIE, DIGNAM, JAMES, PICKLES, TOM, JACQMIN-GADDA, HÉLÈNE AND PROUST-LIMA, CÉCILE. Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer. *Statistics in Medicine* 35(22), 3933–3948, 2016. 70, 106, 107, 109, 112, 114, 115, 116, 119, 122, 130, 139, 145, 163
- FREEDMAN, DAVID A. How can the score test be inconsistent? *The American Statistician* 61(4), 291–295, 2007. 121
- JACQMIN-GADDA, HÉLÈNE, PROUST-LIMA, CÉCILE, TAYLOR, JEREMY MG AND COM-MENGES, DANIEL. Score test for conditional independence between longitudinal outcome and time to event given the classes in the joint latent class model. *Biometrics* 66(1), 11–19, 2010. 121
- KRÓL, AGNIESZKA, FERRER, LOÏC, PIGNON, JEAN-PIERRE, PROUST-LIMA, CÉCILE, DUCREUX, MICHEL, BOUCHÉ, OLIVIER, MICHIELS, STEFAN AND RONDEAU, VIRGINIE. Joint model for left-censored longitudinal data, recurrent events and terminal event: Predictive abilities of tumor burden for cancer evolution with application to the ffd 2000–05 trial. *Biometrics* 72(3), 907–916, 2016. 35, 67, 106, 138, 157, 204
- PROUST-LIMA, CÉCILE, DARTIGUES, JEAN-FRANÇOIS AND JACQMIN-GADDA, HÉLÈNE. Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: a latent process and latent class approach. *Statistics in Medicine* 35(3), 382–398, 2016. 36, 106, 192
- PROUST-LIMA, CÉCILE, SÉNE, MBÉRY, TAYLOR, JEREMY MG AND JACQMIN-GADDA, HÉLÈNE. Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research* 23(1), 74–90, 2014. 62, 105

SCORE TEST FOR A FRAILTY TERM IN THE JOINT MULTI-STATE MODEL

- PROUST-LIMA, CÉCILE, TAYLOR, JEREMY MG, WILLIAMS, SCOTT G, ANKERST, DONNA P, LIU, NING, KESTIN, LARRY L, BAE, KYOUNGWA AND SANDLER, HOWARD M. Determinants of change in prostate-specific antigen over time and its association with recurrence after external beam radiation therapy for prostate cancer in five large cohorts. *International Journal of Radiation Oncology Biology Physics* 72(3), 782–791, 2008. 32, 115, 117, 145
- PUTTER, HEIN AND VAN HOUWELINGEN, HANS C. Frailties in multi-state models: Are they identifiable? do we need them? *Statistical Methods in Medical Research* 24(6), 675–692, 2015. 107, 121
- RIZOPOULOS, DIMITRIS. *Joint Models for Longitudinal and Time-to-Event Data: with Applications in R*. Chapman & Hall/CRC, Boca Raton, 2012. 63, 65, 110, 139, 201
- ROUANET, ANAÏS, JOLY, PIERRE, DARTIGUES, JEAN-FRANÇOIS, PROUST-LIMA, CÉCILE AND JACQMIN-GADDA, HÉLÈNE. Joint latent class model for longitudinal data and interval-censored semi-competing events: Application to dementia. *Biometrics* 72(4), 1123–1135, 2016. 106, 107, 203
- SÈNE, MBÉRY, BELLERA, CARINE A AND PROUST-LIMA, CÉCILE. Shared random-effect models for the joint analysis of longitudinal and time-to-event data: application to the prediction of prostate cancer recurrence. *Journal de la Société Française de Statistique* 155(1), 134–155, 2014. 117, 203
- SINHA, SANJOY K. The use of score tests for frailty variance components in recurrent event data. *Journal of Biometrics & Biostatistics*, 2013. 121, 122
- TAYLOR, JEREMY MG, PARK, YONGSEOK, ANKERST, DONNA P, PROUST-LIMA, CECILE, WILLIAMS, SCOTT, KESTIN, LARRY, BAE, KYOUNGWA, PICKLES, TOM AND SANDLER, HOWARD. Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics* 69(1), 206–213, 2013. 115, 138, 145, 157
- TSIATIS, ANASTASIOS A AND DAVIDIAN, MARIE. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* 14(3), 809–834, 2004. 35, 61, 62, 106, 136, 192
- VERBEKE, GEERT AND MOLENBERGHS, GEERT. The use of score tests for inference on variance components. *Biometrics* 59(2), 254–262, 2003. 111

4.9 Supplementary material

4.9.1 Formulation of the score statistic for a Gaussian frailty term in a joint model for a longitudinal process and a multi-state process

This section details the formulations of the score statistic for a Gaussian frailty term in a joint multi-state model with either shared random effects or shared latent classes.

Formulation of the score statistic for a Gaussian frailty term in a joint multi-state model with shared random effects

In this part, we use the notations given in Section 4.2 of the main manuscript.

The joint model can be written as

$$\begin{cases} Y_i(t) &= Y_i^*(t) + \epsilon_i(t) \\ &= X_i^L(t)^\top \beta + Z_i(t)^\top b_i + \epsilon_i(t) \\ \lambda_{hk}^i(t) &= \lambda_{hk,0}(t) \exp(X_{hk,i}^E{}^\top \gamma_{hk} + W_{hk,i}(t|b_i)^\top \eta_{hk} + \sigma_v v_i^*) \end{cases} \quad (4.7)$$

where v_i^* follows a standard Gaussian distribution.

The associated likelihood function is then

$$L(\sigma_v^2, \theta) = \prod_{i=1}^N \int_{\mathbb{R}^q} f_Y(Y_i|b_i; \theta) f_b(b_i; \theta) \int_{\mathbb{R}} f_E(E_i|v_i^*, b_i; \sigma_v^2, \theta) f_{v^*}(v_i^*) dv_i^* db_i \quad (4.8)$$

where $f_Y(Y_i|b_i; \theta)$ and $f_b(b_i; \theta)$ are defined in the same way as in the manuscript, and

$$f_E(E_i|v_i^*, b_i; \sigma_v^2, \theta) = \prod_{r=0}^{m_i-1} \left[P_{E_i(T_{ir}), E_i(T_{ir})}^i(T_{ir}, T_{i(r+1)}|v_i^*, b_i) \times \lambda_{E_i(T_{ir}), E_i(T_{i(r+1)})}^i(T_{i(r+1)}|v_i^*, b_i)^{\delta_{i(r+1)}} \right]$$

The density of the standardised random effect is

$$f_{v^*}(v_i^*) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{v_i^{*2}}{2}\right)$$

The score test is then applied to test the absence of the frailty:

$$H_0 : \sigma_v^2 = 0 \quad \text{vs} \quad H_1 : \sigma_v^2 > 0 \quad (4.9)$$

The score statistic is defined as the gradient of the log-likelihood (4.8) with respect to σ_v^2 , computed under the null hypothesis: $U(\sigma_v^2 = 0, \theta) = U(0, \theta) = \left. \frac{\partial \log L(\sigma_v^2, \theta)}{\partial \sigma_v^2} \right|_{\sigma_v^2=0} = \sum_{i=1}^N \left. \frac{\partial \log L_i(\sigma_v^2, \theta)}{\partial \sigma_v^2} \right|_{\sigma_v^2=0}$. One can rewrite the score statistic:

$$\begin{aligned} U(0, \theta) &= \sum_{i=1}^N U_i(0, \theta) \\ &= \sum_{i=1}^N \left. \frac{\partial \log L_i(\sigma_v^2, \theta)}{\partial \sigma_v^2} \right|_{\sigma_v^2=0} \\ &= \sum_{i=1}^N \frac{1}{L_i(\sigma_v^2, \theta)} \times \left. \frac{\partial L_i(\sigma_v^2, \theta)}{\partial \sigma_v^2} \right|_{\sigma_v^2=0} \end{aligned} \quad (4.10)$$

where $U_i(0, \theta)$ is the contribution of subject i to the score statistic.

The first derivative of the likelihood with respect to σ_v^2 can be expressed as

$$\frac{\partial L_i(\sigma_v^2, \theta)}{\partial \sigma_v^2} = \int_{\mathbb{R}^q} f_Y(Y_i | b_i; \theta) f_b(b_i; \theta) \int_{\mathbb{R}} \frac{\partial f_E(E_i | v_i^*, b_i; \sigma_v^2, \theta)}{\partial \sigma_v^2} f_{v^*}(v_i^*) dv_i^* db_i$$

with

$$\begin{aligned} \frac{\partial f_E(E_i | v_i^*, b_i; \sigma_v^2, \theta)}{\partial \sigma_v^2} &= f_E(E_i | v_i^*, b_i; \sigma_v^2, \theta) \frac{\partial \log f_E(E_i | v_i^*, b_i; \sigma_v^2, \theta)}{\partial \sigma_v^2} \\ &= \frac{v_i^*}{2\sigma_v} \times \psi(E_i | v_i^*, b_i; \sigma_v^2, \theta) \end{aligned}$$

where $\psi(E_i | v_i^*, b_i; \sigma_v^2, \theta) = f_E(E_i | v_i^*, b_i; \sigma_v^2, \theta) \times \sum_{r=0}^{m_i-1} (\delta_{i(r+1)} + \Lambda_{E_i(T_{ir}), E_i(T_{ir})}(T_{ir}, T_{i(r+1)} | v_i^*, b_i; \sigma_v^2, \theta))$.

One deduces

$$\frac{\partial L_i(\sigma_v^2, \theta)}{\partial \sigma_v^2} = \frac{\int_{\mathbb{R}^q} f_Y(Y_i | b_i; \theta) f_b(b_i; \theta) \int_{\mathbb{R}} \psi(E_i | v_i^*, b_i; \sigma_v^2, \theta) v_i^* f_{v^*}(v_i^*) dv_i^* db_i}{2\sigma_v L_i(\sigma_v^2, \theta)}$$

Expressed under the null hypothesis $\sigma_v^2 = 0$, this expression unfortunately becomes unusable:

$$\left. \frac{\partial L_i(\sigma_v^2, \theta)}{\partial \sigma_v^2} \right|_{\sigma_v^2=0} = \frac{0}{0}$$

One bypasses the problem using the L'Hôpital's rule:

$$\left. \frac{\partial L_i(\sigma_v^2, \theta)}{\partial \sigma_v^2} \right|_{\sigma_v^2=0} = \frac{\int_{\mathbb{R}^q} f_Y(Y_i|b_i; \theta) f_b(b_i; \theta) \int_{\mathbb{R}} \frac{\partial \psi(E_i|v_i^*, b_i; \sigma_v^2, \theta)}{\partial \sigma_v^2} v_i^* f_{v^*}(v_i^*) dv_i^* db_i}{2 \frac{\partial \sigma_v L_i(\sigma_v^2, \theta)}{\partial \sigma_v^2}} \Bigg|_{\sigma_v^2=0}$$

$$\text{with } \frac{\partial [\sigma_v L_i(\sigma_v^2, \theta)]}{\partial \sigma_v^2} = \frac{L_i(\sigma_v^2, \theta)}{2\sigma_v} + \sigma_v \frac{\partial L_i(\sigma_v^2, \theta)}{\partial \sigma_v^2}$$

$$\text{and } \frac{\partial \psi(E_i|v_i^*, b_i; \sigma_v^2, \theta)}{\partial \sigma_v^2} = \frac{v_i^*}{2\sigma_v} \phi(E_i|v_i^*, b_i; \sigma_v^2, \theta),$$

$$\text{where } \phi(E_i|v_i^*, b_i; \sigma_v^2, \theta) = f_E(E_i|v_i^*, b_i; \sigma_v^2, \theta) \times \left\{ \left[\sum_{r=0}^{m_i-1} (\delta_{i(r+1)} + \Lambda_{E_i(T_{ir}), E_i(T_{ir})}(T_{ir}, T_{i(r+1)}|v_i^*, b_i; \sigma_v^2, \theta)) \right]^2 + \sum_{r=0}^{m_i-1} \Lambda_{E_i(T_{ir}), E_i(T_{ir})}(T_{ir}, T_{i(r+1)}|v_i^*, b_i; \sigma_v^2, \theta) \right\}.$$

Then the gradient of the likelihood w.r.t. σ_v^2 under the null hypothesis can be computed as

$$\left. \frac{\partial L_i(\sigma_v^2, \theta)}{\partial \sigma_v^2} \right|_{\sigma_v^2=0} = \frac{\int_{\mathbb{R}^q} f_Y(Y_i|b_i; \theta) f_b(b_i; \theta) \int_{\mathbb{R}} \phi(E_i|v_i^*, b_i; \sigma_v^2, \theta) v_i^{*2} f_{v^*}(v_i^*) dv_i^* db_i}{2 L_i(\sigma_v^2, \theta) + 4 \sigma_v^2 \frac{\partial L_i(\sigma_v^2, \theta)}{\partial \sigma_v^2}} \Bigg|_{\sigma_v^2=0}$$

Since $\int_{\mathbb{R}} v_i^{*2} f_{v^*}(v_i^*) dv_i^* = \mathbb{E}(v_i^{*2}) = 1$ and $\left[\sigma_v^2 \frac{\partial L_i(\sigma_v^2, \theta)}{\partial \sigma_v^2} \right] \Big|_{\sigma_v^2=0} = 0$, we deduce

$$\left. \frac{\partial L_i(\sigma_v^2, \theta)}{\partial \sigma_v^2} \right|_{\sigma_v^2=0} = \frac{\int_{\mathbb{R}^q} f_Y(Y_i|b_i; \theta) f_b(b_i; \theta) \phi(E_i|v_i^*, b_i; \sigma_v^2 = 0, \theta) db_i}{2 L_i(\sigma_v^2 = 0, \theta)}$$

Finally the following formulation of the score statistic is obtained:

$$\begin{aligned} U(0, \theta) &= \sum_{i=1}^N \frac{1}{2L_i(\sigma_v^2 = 0, \theta)} \times \\ &\quad \int_{\mathbb{R}^q} f_Y(Y_i|b_i; \theta) f_E(E_i|b_i; \sigma_v^2 = 0, \theta) f_b(b_i; \theta) \times \\ &\quad \left\{ \left[\sum_{r=0}^{m_i-1} (\delta_{i(r+1)} + \Lambda_{E_i(T_{ir}), E_i(T_{ir})}^i(T_{ir}, T_{i(r+1)}|b_i; \sigma_v^2 = 0, \theta)) \right]^2 + \right. \\ &\quad \left. \sum_{r=0}^{m_i-1} (\Lambda_{E_i(T_{ir}), E_i(T_{ir})}^i(T_{ir}, T_{i(r+1)}|b_i; \sigma_v^2 = 0, \theta)) \right\} db_i \end{aligned} \quad (4.11)$$

Formulation of the score statistic for a Gaussian frailty term in a joint multi-state model with shared latent classes

The joint latent class multi-state model relies on the assumption that the population is heterogeneous and can be divided into G homogeneous sub-groups, with specific marker evolutions and transition intensities. First, the latent variable c_i denotes the latent class membership and its probability is defined by a multinomial logistic model, adjusted on time-independent covariates X_i^P . Given the latent class g ($g = 1, \dots, G$), the longitudinal marker is described by a mixed model with class-specific parameters β_g and the transition intensity from any state h to state l involves class-specific baseline intensities $\lambda_{hkg,0}(\cdot)$ and parameters γ_{hkg} . In order to test a residual correlation between times-to-events, conditional on the latent classes, we consider an individual Gaussian frailty term $v_i^* \sim \mathcal{N}(0, 1)$ impacting all the transition intensities. The model is thus fully defined for any latent class g by:

$$\left\{ \begin{array}{l} P(c_i = g | X_i^P) = \frac{e^{\zeta_{0g} + X_i^{P\top} \zeta_{1g}}}{\sum_m^G e^{\zeta_{0m} + X_i^{P\top} \zeta_m}} \\ Y_i(t) = Y_i^*(t) + \epsilon_i(t) \\ \quad = X_i^L(t)^\top \beta_g + Z_i(t)^\top b_i + \epsilon_i(t) \\ \lambda_{hkg}^i(t) = \lambda_{hkg,0}(t) \exp(X_{hk,i}^{E\top} \gamma_{hkg} + \sigma_v v_i^*) \end{array} \right. \quad (4.12)$$

with $b_i \sim \mathcal{N}(0, \sigma_g^2 D)$, $\epsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_e^2)$. Note that to ensure identifiability, we fix $\zeta_{0G} = \zeta_{1G} = 0$ and $\sigma_G^2 = 1$, considering the G^{th} class as the reference one.

The associated likelihood is written as follows:

$$\begin{aligned} L(\sigma_v^2, \theta) &= \prod_{i=1}^N L_i(\sigma_v^2, \theta) \\ &= \prod_{i=1}^N \sum_{g=1}^G P(c_i = g) f_Y(Y_i | c_i = g; \theta) \int_{\mathbb{R}} f_E(E_i | v_i^*, c_i = g; \sigma_v^2, \theta) f_{v^*}(v_i^*) dv_i^* \end{aligned} \quad (4.13)$$

where $f_Y(Y_i | c_i = g; \theta)$ is the density of a multivariate Gaussian distribution with mean $X_i^L \beta_g$ and variance $\sigma_g^2 Z_i D Z_i^\top + \sigma_e^2 I_{n_i}$, and

$$f_E(E_i | v_i^*, c_i = g; \sigma_v^2, \theta) = \prod_{r=0}^{m_i-1} \left[P_{E_i(T_{ir}), E_i(T_{ir})}^i(T_{ir}, T_{i(r+1)} | v_i^*, c_i = g) \times \lambda_{E_i(T_{ir}), E_i(T_{i(r+1)})}^i(T_{i(r+1)} | v_i^*, c_i = g)^{\delta_{i(r+1)}} \right]$$

Following the exact same calculation steps as described in the previous section for a joint model with shared random effects, we obtain the following formulation of the score statistic in the joint latent class framework:

$$\begin{aligned}
 U(0, \theta) = & \sum_{i=1}^N \frac{1}{2L_i(\sigma_v^2 = 0, \theta)} \times \\
 & \sum_{g=1}^G P(c_i = g) f_Y(Y_i | c_i = g; \theta) f_E(E_i | c_i = g; \sigma_v^2 = 0, \theta) \times \\
 & \left\{ \left[\sum_{r=0}^{m_i-1} \left(\delta_{i(r+1)} + \Lambda_{E_i(T_{ir}), E_i(T_{ir})}^i (T_{ir}, T_{i(r+1)} | c_i = g; \sigma_v^2 = 0, \theta) \right) \right]^2 + \right. \\
 & \left. \sum_{r=0}^{m_i-1} \left(\Lambda_{E_i(T_{ir}), E_i(T_{ir})}^i (T_{ir}, T_{i(r+1)} | c_i = g; \sigma_v^2 = 0, \theta) \right) \right\} \quad (4.14)
 \end{aligned}$$

4.9.2 Simulation data generation

The simulation study carried out in Section 3 of the main manuscript considered two scenarii with different settings of population study. We refer here to the notations in the main manuscript. Data were generated in both scenarii using a joint multi-state frailty model:

$$\left\{ \begin{array}{l}
 Y_i(t) = Y_i^*(t) + \epsilon_i(t) \\
 = (\beta_0 + \beta_{0,X} X_i + b_{i0}) + (\beta_1 + \beta_{1,X} X_i + b_{i1}) \times t + \epsilon_i(t) \\
 \lambda_{hk}^i(t) = \lambda_{hk,0}(t) \exp(\gamma_{hk} X_i + \eta_{hk, \text{level}} Y_i^*(t) + \eta_{hk, \text{slope}} \partial Y_i^*(t) / \partial t + v_i)
 \end{array} \right.$$

In scenario 1, the fixed-effects of the longitudinal sub-model were $\beta_0 = -0.79, \beta_{0,X} = 0.54, \beta_1 = -0.10$ and $\beta_{1,X} = 0.03$. The random effects were such that $b_i = (b_{i0}, b_{i1})^\top \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.35 & -0.04 \\ -0.04 & 0.06 \end{pmatrix} \right)$ and the variance of the centered normal error term was $\sigma_e^2 = 0.23$. For the multi-state sub-model, each log baseline intensity was generated using cubic B-splines with the vector of spline coefficients $(-9.20, -3.50, -5.00, -3.90, -3.50, -2.50, -2.00)^\top$ for transition $0 \rightarrow 1$, $(-11.67, -5.82, -6.41, -4.71, -3.37, -2.01, -1.70)^\top$ for transition $0 \rightarrow 2$ and $(-3.44, -4.18, -4.30, -3.99, -2.63, -2.44, -1.32)^\top$ for transition $1 \rightarrow 2$. The effects of the baseline prognostic factors were $\gamma_{01} = 0.27, \gamma_{02} = 0.75, \gamma_{12} = 0.78$ and the effects of the PSA dynamics were $\eta_{01, \text{level}} = 1.02, \eta_{02, \text{level}} = 0.36, \eta_{12, \text{level}} = -0.03, \eta_{01, \text{slope}} = 1.04, \eta_{02, \text{slope}} = -1.23, \eta_{12, \text{slope}} = 0.99$.

In scenario 2, the longitudinal sub-model considered the fixed-effects $\beta_0 = -0.78, \beta_{0,X} = 0.53, \beta_1 = -0.11$ and $\beta_{1,X} = 0.04$ and the variance of the error term $\sigma_e^2 = 0.23$. The covariance matrix of the random effects was $D = \begin{pmatrix} 0.34 & -0.04 \\ -0.04 & 0.06 \end{pmatrix}$. In the multi-state sub-

model, the baseline transition intensities were proportional from one state to the other: $\lambda_{hk,0}(t) = \lambda_0(t) \exp(\zeta_{hk})$ with $\zeta_{02} = -1.50, \zeta_{03} = -1.00, \zeta_{04} = -2.25, \zeta_{12} = 0.60, \zeta_{13} = 0.10, \zeta_{14} = 0.25, \zeta_{23} = 0.50, \zeta_{34} = 0.50$. ζ_{01} and ζ_{24} were null and not specified in the model estimation step. The reference baseline intensity $\lambda_0(t)$ was generated as a linear combination of cubic B-splines with vector of coefficients $(-1.80, -1.30, -1.30, -1.30, -1.00)^\top$. The effects of the baseline prognostic factors were $\gamma_{01} = 0.05, \gamma_{02} = 0.49, \gamma_{03} = 0.05, \gamma_{04} = 0.06, \gamma_{12} = 0.92, \gamma_{13} = 0.35, \gamma_{14} = -0.38, \gamma_{23} = -0.21, \gamma_{24} = 0.01, \gamma_{34} = 0.26$. Finally, the effects of the marker dynamics were $\eta_{01,\text{level}} = 0.82, \eta_{02,\text{level}} = 0.91, \eta_{03,\text{level}} = 0.53, \eta_{04,\text{level}} = 0.23, \eta_{12,\text{level}} = -0.52, \eta_{13,\text{level}} = -0.56, \eta_{14,\text{level}} = -0.10, \eta_{23,\text{level}} = -0.44, \eta_{24,\text{level}} = 0.07, \eta_{34,\text{level}} = -0.59, \eta_{01,\text{slope}} = 0.10, \eta_{02,\text{slope}} = 0.25, \eta_{03,\text{slope}} = 0.36, \eta_{04,\text{slope}} = -0.57, \eta_{12,\text{slope}} = 0.92, \eta_{13,\text{slope}} = 1.11, \eta_{14,\text{slope}} = 0.02, \eta_{23,\text{slope}} = 0.40, \eta_{24,\text{slope}} = 0.04, \eta_{34,\text{slope}} = 0.60$.

4.9.3 Complementary results for the application

We introduce here the complementary results for the application developed in the main manuscript. Table 4.4 depicts the parameter estimates (except the association parameter estimates which are depicted in Table 4.3 of the main manuscript), in the final joint multi-state model applied on prostate cancer data.

Overall the longitudinal and multi-state parameters were in agreement with the previous results obtained and discussed in Ferrer et al. [2016]. For the multi-state sub-part, they notably confirmed that having a poor prognostic at baseline was not always associated with higher hazards to experience clinical events, after adjustment for the true current PSA dynamics.

The adjustment on age at baseline, which was not done in Ferrer et al. [2016], logically indicated that older subjects had higher risk of directly dying after either EBRT or local recurrence or hormonal therapy. The same result was found for the transition intensity between the local recurrence and distant recurrence states. These interpretations were made given the other prognostic factors in the model.

Table 4.4 – Parameter estimates, standard errors and p -values in the joint multi-state model, based on the pooled data after removal of four outliers ($N = 1470$).

Estimates of the association parameters $\{\eta_{hk}\}_{h,k}$ are given in Table 4.3 of the main manuscript.

	Longitudinal Process				Multi-state Process		
	Value	StdErr	p -value		Value	StdErr	p -value
β_0	-0.25	0.06	< 0.001	$\gamma_{(01,03,12),iPSA}$	-0.06	0.10	0.505
$\beta_{0,iPSA}$	0.80	0.03	< 0.001	$\gamma_{02,iPSA}$	0.01	0.10	0.933
$\beta_{0,cohort}$	-0.01	0.02	0.580	$\gamma_{04,iPSA}$	0.22	0.08	0.004
β_1	0.73	0.14	< 0.001	$\gamma_{(13,14),iPSA}$	-0.50	0.24	0.034
$\beta_{1,iPSA}$	0.87	0.06	< 0.001	$\gamma_{(23,24,34),iPSA}$	-0.20	0.09	0.020
$\beta_{1,tstage2}$	0.38	0.08	< 0.001	$\gamma_{(04,13,14,24),age}$	0.07	0.01	< 0.001
$\beta_{1,tstage3-4}$	0.44	0.13	0.001	$\gamma_{23,age}$	-0.01	0.02	0.379
$\beta_{1,cohort}$	-0.04	0.04	0.414	$\gamma_{01,tstage2}$	1.01	0.26	< 0.001
β_2	-0.18	0.04	< 0.001	$\gamma_{01,tstage3-4}$	0.95	0.32	0.003
$\beta_{2,iPSA}$	0.18	0.02	< 0.001	$\gamma_{02,tstage2}$	-0.05	0.16	0.774
$\beta_{2,tstage2}$	0.14	0.02	< 0.001	$\gamma_{02,tstage3-4}$	0.24	0.20	0.234
$\beta_{2,tstage3-4}$	0.25	0.04	< 0.001	$\gamma_{03,tstage2}$	0.84	0.40	0.034
$\beta_{2,gleason7}$	0.07	0.02	< 0.001	$\gamma_{23,tstage3-4}$	0.33	0.55	0.549
$\beta_{2,gleason8-10}$	0.21	0.04	< 0.001	$\gamma_{23,tstage2}$	1.04	0.51	0.041
$\beta_{2,cohort}$	-0.05	0.01	< 0.001	$\gamma_{23,tstage3-4}$	0.93	0.68	0.171
$\log(\sigma)$	-1.30	0.01		$\gamma_{(01,02),gleason7}$	0.10	0.13	0.417
				$\gamma_{(01,02),gleason8-10}$	0.08	0.19	0.652
D_{11}	0.37	0.03		$\gamma_{03,gleason7}$	0.76	0.31	0.015
D_{12}	0.35	0.03		$\gamma_{03,gleason8-10}$	-0.13	0.64	0.837
D_{13}	0.00	0.01		$\gamma_{13,gleason7}$	1.09	0.81	0.175
D_{22}	0.14	0.02		$\gamma_{13,gleason8-10}$	1.20	1.08	0.265
D_{23}	0.26	0.04		$\gamma_{23,gleason7}$	1.21	0.39	0.002
D_{33}	1.74	0.05		$\gamma_{23,gleason8-10}$	0.36	0.58	0.540
				$\gamma_{(01,34),cohort}$	-0.27	0.09	0.002
				$\gamma_{(13,23),cohort}$	0.95	0.18	< 0.001
				$\gamma_{14,cohort}$	-0.78	0.35	0.024
				$\gamma_{24,cohort}$	-0.47	0.10	< 0.001
				$\zeta_{(12,23)}$	4.47	0.70	< 0.001

D_{ij} denotes the ij -element of the covariance matrix for the random effects. $\gamma_{(hk,h'k'),x}$ denotes the effect of covariate X on the intensities of transitions $h \rightarrow k$ and $h' \rightarrow k'$, i.e. $\gamma_{(hk,h'k'),x} = \gamma_{hk,x} = \gamma_{h'k',x}$. Similarly, $\zeta_{(12,13)}$ is indicated since $\zeta_{12} = \zeta_{13}$.

5 Individual dynamic predictions using landmarking and joint modelling: validation of estimators and robustness assessment

Joint modelling is the most natural approach when we are confronted to a longitudinal biomarker and time-to-event data with competing risks, which are both correlated. Standard joint models with shared random effects specify a mixed sub-model for the repeated measurements of the longitudinal marker, a cause-specific proportional hazard model for the occurrence time of the competing event, and links these two sub-models using a function of the shared random effects such as the current dynamics of the marker. These two sub-models are then estimated simultaneously for correctly taking into account the correlation between the longitudinal process and the survival process. In the prediction setting, as supported in Suresh et al. [2017] the joint models also satisfies the consistency condition introduced by Jewell and Nielsen [1993] which stipulates that the hazard function and the marker dynamics must be linked at all time points to give consistent dynamic predictions.

However, inference may be complex in such models: model estimation and prediction steps may be difficult to achieve to provide accurate predictions. Moreover the joint model with shared random effects requires restrictive assumptions, such as proportional hazards according to each covariate effect, or parametric specifications of both the whole longitudinal marker trajectory and the association structure between the marker dynamics and each cause-specific hazard of event.

To reduce the possible bias due to these assumptions and avoid hard computational procedures, the landmarking approach has been introduced [van Houwelingen, 2014]. The main idea is to consider at each time point, called landmark time, only the subjects at risk and their covariates history collected until this time. Classical landmark models are cause-

specific proportional hazard (CS PH) models applied from the landmark time point, and, to reduce the possible bias due to the proportional hazard assumption, an administrative censoring is also applied at the end of the prediction window. Some authors [Fine and Gray, 1999; Scheike et al., 2008; Andersen and Pohar Perme, 2010] also focused on a direct estimation of the quantity of interest, which is a probability while CS PH models are built on hazard functions and may be not adapted to the predictive setting. These landmark models do not provide consistent dynamic predictions as joint models, and are adjusted on the marker dynamics predicted at the landmark time point.

In this work we are interested in the prediction of individual cumulative incidences of event, computed from a landmark time point and based on individual covariates history including biomarker data. Using the landmarking or joint modelling approaches, several proposals have been published in the literature to compute such quantity [Maziarz et al., 2017; Sène et al., 2016; Rizopoulos, 2011]. However none formally defined the quantity of interest, validated its estimators, and compared them according to several scenarii of misspecification. Moreover no concept of uncertainty around the predictions was introduced in the landmark approach [Proust-Lima and Taylor, 2009; Sweeting et al., 2016; Rizopoulos *and others*, 2017].

We fill the gap through an extensive simulation study which properly defines the quantity of interest, and validates the estimators from the joint modelling and landmarking approaches. The predictive abilities of these estimators are then compared in terms of accuracy of prediction, discriminatory power, and robustness to the model hypotheses though several cases of well- and mis-specification such as the violation of the PH assumption, the misspecification of the association function between the longitudinal and survival processes, and the misspecification of the longitudinal trajectory of the marker. The models and estimators are implemented in R.

This paper is currently submitted for publication in a statistical journal.

Individual dynamic predictions using landmarking and joint modelling: validation of estimators and robustness assessment

Loïc Ferrer^{1,*}, Hein Putter², Cécile Proust-Lima¹

¹INSERM, UMR1219, Univ. Bordeaux, F-33076 Bordeaux, France

²Leiden University Medical Center, Leiden, the Netherlands

Submitted for publication.

SUMMARY. After the diagnosis of a disease, one major objective is to predict cumulative probabilities of events such as clinical relapse or death from the individual information collected up to a prediction time, including usually biomarker repeated measurements. Several competing estimators have been proposed to calculate these individual dynamic predictions, mainly from two approaches: joint modelling and landmarking. These approaches differ by the information used, the model assumptions and the complexity of the computational procedures. It is essential to properly validate the estimators derived from joint models and landmark models, quantify their variability and compare them in order to provide key elements for the development and use of individual dynamic predictions in clinical follow-up of patients. Motivated by the prediction of two competing causes of progression of prostate cancer from the history of prostate-specific antigen, we conducted an in-depth simulation study to validate and compare the dynamic predictions derived from these two methods. Specifically, we formally defined the quantity to estimate and its estimators, proposed techniques to assess the uncertainty around predictions and validated them. We also compared the individual dynamic predictions derived from joint models and landmark models in terms of prediction error, discriminatory power, efficiency and robustness to model assumptions. We show that these prediction tools should be handled with care, in particular by properly specifying models and estimators.

KEYWORDS. Competing risks; Dynamic Prediction; Landmarking; Joint modelling; Prediction accuracy; Robustness.

5.1 Introduction

After diagnosis and subsequent treatment of cancer, patients are typically monitored via repeated measurements of biomarkers. For example, in patients with prostate cancer treated by radiotherapy, the Prostate Specific Antigen (PSA) is measured routinely. Precisely predicting the individualized probabilities of events such as clinical relapse for these patients from their individual information collected until the prediction time has become

a central issue [Proust-Lima and Taylor, 2009; Goldstein et al., 2017]. Personalized treatment strategies can indeed be proposed according to the updated individual probabilities [Sène et al., 2016], or the planning of the next biomarker measurement can be optimized [Rizopoulos et al., 2015].

Two main approaches have been proposed to compute individual dynamic predictions: joint modelling and landmarking. These differ in the used information, the model assumptions and the complexity of computational procedures.

The *joint modelling (JM) approach* simultaneously models the repeated measurements of the biomarker (e.g., using a linear mixed model in standard JM) and the time-to-event data (e.g., using a proportional hazards model in standard JM) by linking them using a function of shared random effects [Tsiatis and Davidian, 2004]. This approach has the advantage of taking into account the endogenous nature of biomarkers [Kalbfleisch and Prentice, 2011], of only requiring one model estimation for any prediction time, and of modelling the progression of the disease as a whole, which makes it very popular. But it is often based on simplifying assumptions (e.g., proportional hazards, number of random effects) and may be complex to estimate, so that it should be handled carefully and can remain difficult to apply in practice.

The *landmarking approach* consists of adjusting standard survival models considering only the subsample of subjects still at risk at the prediction time and the longitudinal information collected up to the prediction time [van Houwelingen, 2007]. These models induce significantly less numerical problems and reduce the possible estimation bias related to the proportional hazards assumption. However, as they do not fully explore the collected information during the follow-up and the correlation between the marker and the time of event, they can produce sub-efficient estimators [Huang *and others*, 2016] and are only an approximation of the (correct) joint estimator. Indeed, as supported in Suresh et al. [2017] they do not satisfy the consistency condition introduced by Jewell and Nielsen [1993] which stipulates that the hazard function and the marker dynamics must be linked at all time points to give consistent dynamic predictions. In the presence of longitudinal biomarkers, the landmark approach can result in several models. Most of the time a survival model (cause-specific proportional hazards) is adjusted on the last observed value of the biomarker. In addition to truncation at the prediction time, censoring is administered at the end of the prediction window to reduce possible bias related to the proportionality of hazards. But this approach does not take into account fluctuations of the biomarker, its observation at discrete times and measurement errors. To circumvent this problem, the last observed value of the biomarker may be replaced by its

predicted value at the prediction time (obtained from a linear mixed model) [Sweeting et al., 2016; Rizopoulos *and others*, 2017]. This two-step model considers the same method of truncation and administrative censoring. It takes into account all the collected information of the biomarker until the prediction time for the subjects at risk. But the event probabilities must be deduced by approximation and the model is not completely freed of the proportional hazards assumption. In the context of competing risks, rather than using a cause-specific proportional hazards model, the conditional probabilities of event can be directly estimated by a dynamic pseudo-observations approach [Nicolaie et al., 2013b]. This approach directly models the conditional probabilities of event and is freed from the proportionality hazards assumption. By considering the predicted value of the biomarker at the prediction time as a covariate, it also takes into account the trajectory of the biomarker. But this approach requires the specification of a link function and can still provide less efficient estimators than the joint model. Note that, in these landmark models as well as in joint models, any function of the biomarker trajectory parameters can be used instead of the biomarker predicted value at the prediction time.

In addition, although of central interest in many recent works, estimators of dynamic predictions and of their uncertainty were never formally validated while there exist several competing proposals in the joint modelling framework [Maziarz et al., 2017; Sène et al., 2016; Rizopoulos, 2011]. In the landmarking approach with longitudinal biomarkers, some estimators were also proposed but not validated, and no concept of uncertainty was introduced [Proust-Lima and Taylor, 2009; Sweeting et al., 2016; Rizopoulos *and others*, 2017].

Motivated by the prediction of competing progressions of prostate cancer from the PSA history, we first proposed estimators of individual dynamic predictions and of their uncertainty with 95% confidence intervals for the joint and the landmark approaches, and we validated them in a simulation study. We then compared the predictive accuracy of the models under several scenarios to explore their robustness to misspecification.

The rest of the paper is organized as follows. Section 5.2 introduces the prediction models and the derived estimators of dynamic predictions and of their uncertainty. Section 5.3 briefly describes the motivating data. The simulation studies are carried out in Section 5.4 for validating the proposed estimators and comparing them in terms of prediction accuracy. The paper ends with a discussion in Section 5.5.

5.2 Prediction models

Let us consider the competing risks setting where the subjects are at risk to experience K competing events. For each subject i ($i = 1, \dots, N$), we denote T_i the earliest time-to-event and $\delta_i = k$ the cause of event, with $k \in 1, \dots, K$. In the presence of censoring, we observe the event time $T_i^\dagger = \min(T_i, C_i)$ with C_i the censoring time, and the indicator of event becomes $\Delta_i = \delta_i \cdot \mathbb{1}\{T_i \leq C_i\}$ with $\mathbb{1}$ the indicator function. We also observe \mathcal{X}_i the (possibly exogenous time-dependent) covariates collected until the event time, and Y_i an endogenous longitudinal marker measured repeatedly with $Y_i(t_{ij})$ the observed measure at time t_{ij} for $j = 1, \dots, n_i$ and $t_{in_i} \leq T_i^\dagger$. In the following, $\mathcal{X}_i(s)$ denotes the history of \mathcal{X}_i until time s , $\mathcal{Y}_i(s) = \{Y_i(t_{ij}) : 0 \leq t_{ij} \leq s, j = 1, \dots, n_i(s)\}$ denotes the history of the marker until s , and the model formulations assume a longitudinal marker with Gaussian distribution.

5.2.1 Definition of individual dynamic prediction

In this paper we are interested in the individual cumulative probability of the event of cause k between times s and $s + w$ for a new subject \star , with s the landmark time (or prediction time) and w the horizon. This probability, also called landmark specific cumulative incidence of cause k is defined as

$$\pi_\star^k(s, w) = \Pr(s < T_\star \leq s + w, \delta_\star = k | T_\star > s, \mathcal{Y}_\star(s), \mathcal{X}_\star(s)). \quad (5.1)$$

We focus on several parametric models that express this quantity of interest as a function of a vector of parameters θ :

$$\pi_\star^k(s, w; \theta) = \Pr(s < T_\star \leq s + w, \delta_\star = k | T_\star > s, \mathcal{Y}_\star(s), \mathcal{X}_\star(s); \theta). \quad (5.2)$$

In practice, θ is unknown and is replaced by $\hat{\theta}_{\mathcal{I}}$, its estimate from the considered observed data in the learning sample \mathcal{I} . In the remainder of the manuscript, this subscript is omitted for the sake of simplicity, and the estimated quantity of interest is denoted $\hat{\pi}_\star^k(s, w; \hat{\theta})$.

Two kinds of variability of $\hat{\pi}_\star^k(s, w; \hat{\theta})$ can be defined to quantify the uncertainty of this estimator. The first one consists in considering the variance of $\pi_\star^k(s, w; \theta)$ conditional to $T_\star > s$, $\mathcal{Y}_\star(s)$ and $\mathcal{X}_\star(s)$, with respect to the estimated parameters $\hat{\theta}$ [Proust-Lima and Taylor, 2009; Taylor et al., 2013; Król et al., 2016; Maziarz et al., 2017]. The second one consists in considering the variance as only conditional to $T_\star > s$ and $\mathcal{X}_\star(s)$ by taking into account both the variability of the estimated parameters $\hat{\theta}$ and the variability of $\mathcal{Y}_\star(s)$ due

to measurement errors in the marker's observations [Yu et al., 2008; Rizopoulos, 2011]. In this contribution, we propose techniques to quantify both sources of uncertainty although simulations focus only on the former.

5.2.2 Joint model

Model formulation

The joint model considers the full collected information $\mathcal{I} = \{(T_i^\dagger, \Delta_i, \mathcal{Y}_i(T_i^\dagger), \mathcal{X}_i(T_i^\dagger)); i = 1, \dots, N\}$. It is decomposed into two sub-models linked by a function of a shared latent structure. The most popular joint model [Rizopoulos, 2012b] links a linear mixed model for the repeated measurements of the marker and a cause-specific proportional hazards model for the specific hazard of each cause of event k using a function of shared random effects:

$$\begin{cases} Y_i(t) &= m_i(t) + \epsilon_i(t) \\ &= X_i^L(t)^\top \beta + Z_i(t)^\top b_i + \epsilon_i(t), \\ \lambda_i^k(t) &= \lambda_{k,0}(t) \exp \left\{ X_{k,i}^E{}^\top \gamma_k + W_{k,i}(t|b_i; \beta)^\top \eta_k \right\}, \end{cases}$$

where $t > 0$ and $\lambda_i^k(t)$ denotes the hazard function of cause k at time t , with $k = 1, \dots, K$. In the longitudinal sub-part, $X_i^L(t)$ and $Z_i(t)$ denote vectors of covariates (possibly time-dependent) associated respectively with the vector of fixed effects β and the vector of random effects $b_i, b_i \sim \mathcal{N}_q(0, D)$. The error term is $\epsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$; the random effects and error terms are independent. In the survival sub-part, $\lambda_{k,0}(t)$ denotes the parametric baseline hazard of cause k at time t . The vector of covariates $X_{k,i}^E$ is associated with the vector of coefficients γ_k . Note that for simplicity, we do not consider any exogenous time-dependent prognostic variable although this is not a requirement. The (possibly multivariate) function $W_{k,i}(t|b_i; \beta)$ denotes the function of dependence between the longitudinal process and the hazard of event of cause k , such as for example the unbiased current level of the marker $m_i(t)$, the unbiased current slope $\partial m_i(t)/\partial t$, or both $(m_i(t), \partial m_i(t)/\partial t)^\top$.

The joint model assumes proportional hazards (PH) between levels of covariates.

It can be estimated in the maximum likelihood framework by using the independence between the longitudinal process $\mathcal{Y}_i(T_i^\dagger)$ and the survival process (T_i^\dagger, Δ_i) conditionally on the random effects b_i . The likelihood involves integrals over the random effects and time that have to be numerically solved, usually using Gaussian quadratures. Note that the number of quadrature points has to be chosen carefully to provide correct inference [Ferrer et al., 2016].

Cumulative incidence estimator

Once the model is estimated, the vector of parameters $\widehat{\theta}$ and its variance matrix $\widehat{V}(\widehat{\theta})$ are obtained and we are able to compute for each new subject \star the predicted conditional cumulative incidence of cause k for all the possible horizons w and landmark times s :

$$\widehat{\pi}_\star^k(s, w; \widehat{\theta}) = \int_{\mathbb{R}^q} \Pr(s < T_\star \leq s + w, \delta_\star = k | T_\star > s, \mathcal{X}_\star(s), b_\star; \widehat{\theta}) f(b_\star | T_\star > s, \mathcal{Y}_\star(s), \mathcal{X}_\star(s); \widehat{\theta}) db_\star. \quad (5.3)$$

Another estimator of (5.2) (faster but less accurate) can be obtained by approximating the integral over the random effect distribution by the integrand computed at the modal point. This Laplace approximation of (5.3) will be called conditional estimator. The complete formulas of the marginal and conditional estimators are detailed in Section 5.7.1 of the Supplementary Material.

To validate our estimator in the simulation study, we mainly considered its variance as conditional to $T_\star > s$ and $\mathcal{Y}_\star(s)$. The corresponding 95% confidence interval of (5.3) can be obtained using parametric bootstrap techniques. The procedure is realized as follows:

Consider a large L ; for each $l = 1, \dots, L$,

1. generate $\widetilde{\theta}^{(l)} \sim \mathcal{N}(\widehat{\theta}, \widehat{V}(\widehat{\theta}))$;
2. compute $\widetilde{\pi}_\star^{k,(l)}(s, w; \widetilde{\theta}^{(l)}) = \int_{\mathbb{R}^q} \Pr(s < T_\star \leq s + w, \delta_\star = k | T_\star > s, \mathcal{X}_\star(s), b_\star; \widetilde{\theta}^{(l)}) f(b_\star | T_\star > s, \mathcal{Y}_\star(s), \mathcal{X}_\star(s); \widetilde{\theta}^{(l)}) db_\star$.

Compute the 95% confidence interval from the 2.5th and 97.5th percentiles of $\{\widetilde{\pi}_\star^{k,(l)}(s, w; \widetilde{\theta}^{(l)}); l = 1, \dots, L\}$.

The same procedure can be used with the conditional estimator of the probability by replacing the expression in step 2.

The additional variability due to measurement errors in the observations $\mathcal{Y}_\star(s)$ can be easily taken into account by adding a step to this algorithm which draws marker measurements from their estimated distribution:

- 1bis. generate $\mathcal{Y}_\star^{(l)}(s) = \{Y_\star^{(l)}(t_{\star j}) : 0 \leq t_{\star j} \leq s, j = 1, \dots, n_\star(s)\}$
with $Y_\star^{(l)}(t_{\star j}) \sim \mathcal{N}(X_\star^L(t_{\star j})^\top \widehat{\beta} + Z_\star(t_{\star j})^\top \widehat{b}_\star, \widehat{\sigma}^2)$; $\widehat{b}_\star = \mathbb{E}(b_\star | \mathcal{Y}_\star(s), \mathcal{X}_\star(s); \widehat{\theta})$, $\widehat{\beta}$ and $\widehat{\sigma}$ subsets of $\widehat{\theta}$.

The l -th bootstrapped estimator is then obtained by replacing $\mathcal{Y}_\star(s)$ by $\mathcal{Y}_\star^{(l)}(s)$ in step 2.

5.2.3 Landmark cause-specific proportional hazards model

An alternative to joint models is landmark models which only consider subjects at risk at a landmark time s and the longitudinal information $\{\mathcal{Y}(s), \mathcal{X}(s)\}$ collected until s . When considering PH landmark models, administrative censoring is applied at the end of the prediction window $s + w$ in order to reduce the possible bias entailed by a violation of the PH assumption. The considered information becomes $\mathcal{I} = \{(\xi_i(s, w), \Psi_i(s, w), \mathcal{Y}_i(s), \mathcal{X}_i(s)); i = 1, \dots, N^\dagger(s)\}$, with $\xi_i(s, w) = \min(T_i^\dagger, s + w)$, $\Psi_i(s, w) = \Delta_i \cdot \mathbb{1}\{s < T_i \leq s + w\}$ and $N^\dagger(s) = \sum_{i=1}^N \mathbb{1}\{T_i^\dagger > s\}$.

Model formulation

The landmark cause-specific (CS) proportional hazards (PH) model is defined by

$$\lambda_i^k(t) = \lambda_{k,0}(t) \exp \left\{ X_{k,i}^E \top \gamma_k + W_{k,i}(s) \top \eta_k \right\},$$

where $t > s$, $\lambda_{k,0}(\cdot)$ is an unspecified cause-specific baseline hazard function and $W_{k,i}(s)$ is a multivariate function that depicts the dynamics of the marker extrapolated at time s . The model is estimated by maximizing the Cox partial likelihood [Cox, 1972] for each considered pair of landmark and horizon times. Note that for the sake of clarity, we did not use a subscript s, w for the model parameters although they are different for each (s, w) .

To take into account the information of the marker before the landmark time s , one can consider the last observed value only, i.e. $W_{k,i}(s) = Y_i(t_{in_i(s)})$. However, this technique, called "naive landmark model" assumes that the marker is measured without error and considers neither the whole trajectory of the marker until s nor the subject-specific gap between $t_{in_i(s)}$ and s .

A better alternative is to deduce the value of $W_{k,i}(s)$ at time s from a linear mixed model estimated on the marker measurements collected until s in subjects at risk at s . This technique, called the "two-stage landmark model" considers $W_{k,i}(s) = \widehat{W}_{k,i}(s|\widehat{b}_i; \widehat{\beta})$, where $\widehat{\beta}$ is the vector of estimated fixed effects and $\widehat{b}_i = \mathbb{E}(b_i|\mathcal{Y}_i(s), \mathcal{X}_i(s); \widehat{\theta}) = \widehat{D}Z_i^\top \widehat{V}_i^{-1}(Y_i - X_i^L \widehat{\beta})$ is the vector of empirical Bayes estimates of the individual random effects, with $\widehat{V}_i = Z_i \widehat{D} Z_i^\top + \widehat{\sigma}^2 I_{n_i(s)}$. Here X_i^L and Z_i are the matrices of covariates with respectively the row vectors $X_i^L(t_{ij})^\top$ and $Z_i(t_{ij})^\top$, and the column vector Y_i is with elements Y_{ij} , for $j = 1, \dots, n_i(s)$. I is the identity matrix.

Cumulative incidence estimator

With the two-stage approach, the predicted conditional cumulative incidence of cause k for subject \star is

$$\hat{\pi}_\star^k(s, w; \hat{\theta}) = \Pr(s < T_\star \leq s + w, \delta_\star = k | T_\star > s, \mathcal{X}_\star(s), \hat{b}_\star; \hat{\theta}), \quad (5.4)$$

where $\hat{\theta}$ is the vector of estimated parameters (with associated estimated variance $\widehat{V}(\hat{\theta})$), and $\hat{b}_\star = \mathbb{E}(b_\star | \mathcal{Y}_\star(s), \mathcal{X}_\star(s); \hat{\theta})$.

To estimate valid 95% confidence intervals, it is necessary to take into account the variability due to the parameter and baseline hazard estimates. The same parametric bootstrap technique as described for the joint model can be used for the parameter estimates but it can't be applied for the baseline hazard estimates. The unspecified cumulative baseline hazard $\Lambda_{k,0}(t) = \int_s^t \lambda_{k,0}(u) du$ is estimated using the Breslow's estimator [Breslow, 1972], $\hat{\Lambda}_{k,0}(t) = \int_s^t \hat{\Pi}_k^{(0)}(\hat{\theta}, u)^{-1} d\bar{J}_k(u)$ where $\hat{\Pi}_k^{(0)}(\hat{\theta}, u) = \frac{1}{N^\dagger(s)} \sum_{i=1}^{N^\dagger(s)} \mathbb{1}\{\xi_i(s, w) \geq u\} \exp\{X_{k,i}^E \top \hat{\gamma}_k + \widehat{W}_{k,i}(s|\hat{b}_i; \hat{\beta}) \top \hat{\eta}_k\}$ and $\bar{J}_k(u) = \frac{1}{N^\dagger(s)} \sum_{i=1}^{N^\dagger(s)} \mathbb{1}\{\xi_i(s, w) \leq u, \Psi_i(s, w) = k\}$. We propose a procedure that combines parametric bootstrap to take into account the variability associated to $\hat{\theta}$ and perturbation-resampling methods, inspired by Sinnott and Cai [2016], to take into account the variability associated to $\hat{\Lambda}_{k,0}(\cdot)$.

This technique, which also avoids hard computational cost, is validated in the simulation study in Section 5.4.1. The full procedure is realized as follows:

For each bootstrap sample $l = 1, \dots, L$, where L is large enough;

1. generate $\tilde{\theta}^{(l)} \sim \mathcal{N}(\hat{\theta}, \widehat{V}(\hat{\theta}))$ and deduce $\tilde{b}_\star^{(l)} = \mathbb{E}(b_\star | \mathcal{Y}_\star(s), \mathcal{X}_\star(s); \tilde{\theta}^{(l)})$;
2. for each subject $i \in 1, \dots, N^\dagger(s)$ of the learning sample, generate $\nu_i^{(l)} \sim 4 \cdot \text{Beta}(1/2, 3/2)$;
3. compute $\tilde{\Lambda}_{k,0}^{(l)}(t) = \int_s^t \hat{\Pi}_k^{(0),(l)}(\tilde{\theta}^{(l)}, u)^{-1} d\bar{J}_k^{(l)}(u)$ with $\hat{\Pi}_k^{(0),(l)}(\tilde{\theta}^{(l)}, u) = \frac{1}{N^\dagger(s)} \sum_{i=1}^{N^\dagger(s)} \nu_i^{(l)} \mathbb{1}\{\xi_i(s, w) \geq u\} \exp\{X_{k,i}^E \top \tilde{\gamma}_k^{(l)} + \widehat{W}_{k,i}(s|\tilde{b}_i^{(l)}; \tilde{\beta}^{(l)}) \top \tilde{\eta}_k^{(l)}\}$ and $\bar{J}_k^{(l)}(u) = \frac{1}{N^\dagger(s)} \sum_{i=1}^{N^\dagger(s)} \nu_i^{(l)} \mathbb{1}\{\xi_i(s, w) \leq u, \Psi_i(s, w) = k\}$;
4. for each subject \star of the validation sample, deduce $\tilde{\pi}_\star^k(s, w; \tilde{\theta}^{(l)}) = f(\{\tilde{\Lambda}_{k,0}^{(l)}(u), s < u \leq s + w\}; k = 1, \dots, K, \tilde{\theta}^{(l)}, \tilde{b}_\star^{(l)})$, where $f(\cdot)$ is a function specified in Section 5.7.1 of the Supplementary Material.

Compute the 95% confidence interval from the 2.5th and 97.5th percentiles of $\{\tilde{\pi}_\star^{k,(l)}(s, w; \tilde{\theta}^{(l)}); l = 1, \dots, L\}$.

Similarly as in joint models, additional uncertainty on the individual marker measurements can be considered by adding a step 1.bis to perturb the individual marker's observations.

Using a naive approach, the predicted conditional cumulative incidence of cause k for subject \star is

$$\hat{\pi}_\star^k(s, w; \hat{\theta}) = \Pr(s < T_\star \leq s + w, \delta_\star = k | T_\star > s, \{X_{k,\star}^E; k = 1, \dots, K\}, Y_\star(t_{\star n_\star(s)}); \hat{\theta}), \quad (5.5)$$

and the same technique combining parametric bootstrap and perturbation-resampling can be used to obtain 95% confidence intervals. Note however that using the naive approach, the variance of $\hat{\pi}_\star^k(s, w; \hat{\theta})$ necessarily neglects the variability due to the measurement errors in the marker measurements.

5.2.4 Landmark model based on pseudo-observations

Cause-specific hazard models rely on the PH assumption and require the computation of integrals over time in the individual cumulative incidences. To avoid these issues, some authors have focused on the direct modelling of the individual cumulative incidences with for example the Fine-Gray model [Fine and Gray, 1999], the binomial regression models [Scheike et al., 2008] or the pseudo-value approach [Andersen and Pohar Perme, 2010]. The latter is developed here. The pseudo-observation approach does not require the PH assumption, hence the considered information is $\mathcal{I} = \{(T_i^\dagger, \Delta_i, \mathcal{Y}_i(s), \mathcal{X}_i(s)); i = 1, \dots, N^\dagger(s)\}$.

Model formulation

For subjects at risk at time s , we are interested in the expectation of $\mu_i^k(s, w) = \mathbb{1}(T_i \leq s + w, \delta_i = k)$. In presence of censoring, this quantity is not always observable. Thus the idea is to define the dynamic jackknife pseudo-observation [Nicolai et al., 2013b] of the non-parametric estimator of $\pi^k(s, w)$: $\hat{\mu}_i^k(s, w) = N^\dagger(s) \hat{F}^k(s, w) - (N^\dagger(s) - 1) \hat{F}_{(-i)}^k(s, w)$, where $N^\dagger(s)$ is the number of subjects at risk at s and $\hat{F}^k(s, w)$ is the Aalen-Johansen estimate of $\pi^k(s, w)$ [Andersen et al., 1993].

To include the dynamic information on the marker until s , the same two-stage approach as defined in Section 5.2.3 can be used to deduce $\widehat{W}_{k,i}(s|\hat{b}_i; \hat{\beta})$ in those still at risk in s . The pseudo-observation and the prognostic factors are then linked through a

generalized linear model with link function g :

$$g \left[\mathbb{E} \{ \widehat{\mu}_i^k(s, w) | T_i^\dagger > s \} \right] = \gamma_{0,k} + X_{k,i}^{E \top} \gamma_{1,k} + \widehat{W}_{k,i}(s | \widehat{b}_i; \widehat{\beta})^\top \eta_k.$$

The model is thus estimated using generalized estimating equations (GEE) [Andersen and Pohar Perme, 2010].

Cumulative incidence estimator

The predicted conditional cumulative incidence can directly be expressed as

$$\widehat{\pi}_\star^k(s, w; \widehat{\theta}) = \Pr(s < T_\star \leq s + w, \delta_\star = k | T_\star > s, \mathcal{X}_\star(s), \widehat{b}_\star; \widehat{\theta}), \quad (5.6)$$

with $\widehat{b}_\star = \mathbb{E}(b_\star | \mathcal{Y}_\star(s), \mathcal{X}_\star(s); \widehat{\theta})$, where $\widehat{\theta}$ is the vector of estimated parameters (with associated estimated variance matrix $\widehat{V}(\widehat{\theta})$). For example with the cloglog link function g ($g(x) = \text{cloglog}(x) = \log\{-\log(1-x)\}$), it can be expressed as: $\widehat{\pi}_\star^k(s, w; \widehat{\theta}) = 1 - \exp \left[- \exp \{ \widehat{\gamma}_{0,k} + X_{k,i}^{E \top} \widehat{\gamma}_{1,k} + \widehat{W}_{k,i}(s | \widehat{b}_i; \widehat{\beta})^\top \widehat{\eta}_k \} \right]$.

The 95% confidence intervals of (5.6) may be calculated using parametric bootstrap:

Consider a large L ; for each $l = 1, \dots, L$,

1. generate $\widetilde{\theta}^{(l)} \sim \mathcal{N}(\widehat{\theta}, \widehat{V}(\widehat{\theta}))$ and deduce $\widetilde{b}_\star^{(l)} = \mathbb{E}(b_\star | \mathcal{Y}_\star(s), \mathcal{X}_\star(s); \widetilde{\theta}^{(l)})$;
2. compute $\widetilde{\pi}_\star^{k,(l)}(s, w; \widetilde{\theta}^{(l)}) = \widehat{\pi}_\star^{k,(l)}(s, w | \widetilde{b}_\star^{(l)}; \widetilde{\theta}^{(l)})$.

Compute the 95% confidence interval from the 2.5th and 97.5th percentiles of $\{ \widetilde{\pi}_\star^{k,(l)}(s, w; \widetilde{\theta}^{(l)}); l = 1, \dots, L \}$.

Uncertainty from $\mathcal{Y}_\star(s)$ in the variance of $\widehat{\pi}_\star^k(s, w; \widehat{\theta})$ can again be accounted for by perturbing the observations $\mathcal{Y}_\star(s)$ in an additional 1.bis step, as described in Section 5.2.3.

5.2.5 Implementation

The estimation of the prediction models and the computation of the derived estimators were performed in R using standard packages and extensions coded by the authors, with the `JM` package for the joint model, the `survival` package for the landmark cause-specific proportional hazards models and the `pseudo` and `geepack` packages for the landmark model based on pseudo-values. Examples of codes used for the manuscript writing can be found in Section 5.7.5 of the Supplementary Material, and detailed examples can be found at <https://github.com/LoicFerrer/> for practical use.

5.3 Motivating data

The paper relies on simulation studies inspired by the data analyzed in Ferrer et al. [2016]. In this study, patients ($N = 1474$) had a clinically localized prostate cancer and were treated by external beam radiotherapy. After the end of the radiotherapy, repeated measurements of the Prostate Specific Antigen (PSA) were collected until the occurrence of a clinical event defined as the recurrence of the disease (local/distant recurrence, initiation of hormonal therapy or death due to the prostate cancer) or death due to an other cause. Post-treatment PSA trajectory was mostly biphasic with a short term drop followed by a stable or slight increase [Proust-Lima et al., 2008]. Several authors [Proust-Lima and Taylor, 2009; Taylor et al., 2013] showed that including these post-treatment PSA dynamics in dynamic prediction tools of disease recurrence highly reduced the prediction error.

5.4 Simulation studies

Two simulation studies were performed, one for the validation of the estimators (Section 5.4.1), and a second for their comparison and the assessment of their robustness to misspecification (Section 5.4.2). Both simulation studies relied on the same following design.

$R = 500$ learning samples of $N = 1000$ subjects as well as a validation sample of $N^{\text{new}}(0) = 500$ subjects were generated from a joint model with parameter values θ_0 [Crowther and Lambert, 2013; Ferrer et al., 2016]. The models detailed in section 5.2 were estimated on each learning sample r ($r = 1, \dots, R$) and the derived estimators of cumulative incidence were computed for a given horizon w on the $N^{\text{new}}(s)$ subjects ($\star = 1, \dots, N^{\text{new}}(s)$) of the validation sample who did not experience any event before landmark time s . The simulation design was built to validate the estimator defined in (5.2) and its associated variance conditional to the full set of individual observations for prediction, so that we did not resample the observed markers in the prediction sample.

For each replicate r , we then compared the true generated cumulative incidence $\pi_{\star}^k(s, w; \theta_0) = \int_{\mathbb{R}^q} \pi_{\star}^k(s, w | b_{\star}; \theta_0) f(b_{\star} | T_{\star} > s, \mathcal{Y}_{\star}(s), \mathcal{X}_{\star}(s); \theta_0) db_{\star}$ with the estimators $\hat{\pi}_{\star, r}^k(s, w; \hat{\theta})$.

5.4.1 Simulation study I: Validation of the estimators $\hat{\pi}_*^k(s, w; \hat{\theta})$

To validate the proposed estimators, we checked the distributions over the individuals of the estimated relative bias and the estimated coverage rates for $\pi_*^k(s, w; \theta_0)$. We also investigated the efficiency of the estimators with the mean relative change in the confidence interval widths.

Model specification

For each subject i (learning or validation sample), data were generated according to the joint model:

$$\begin{cases} Y_i(t) &= m_i(t) + \epsilon_i(t) \\ &= (\beta_0 + \beta_{0,X}X_i + b_{i0}) + (\beta_1 + \beta_{1,X}X_i + b_{i1})t + \epsilon_i(t), \\ \lambda_i^k(t) &= \lambda_{k,0}(t) \exp \left\{ \gamma_k X_i + \eta_{1,k} m_i(t) + \eta_{2,k} \frac{\delta m_i(t)}{\delta t} \right\}, \end{cases}$$

where $\log(\lambda_{k,0}(t))$ is a combination of cubic B-splines with one internal knot, k is the cause of event (Recurrence ; Death); X_i is a continuous variable. The coefficients and the distribution of the covariates used for the generation data correspond to those obtained on the motivating data with X_i the PSA level before treatment initiation. They are given in Section 5.7.3 of the Supplementary Material.

Results

Due to the duration of the procedures, the simulations were run for two landmark times $s = 1, 5$, one horizon time $w = 3$ and 200 subjects randomly selected from the validation sample. $R = 499$ and $R = 486$ replicates were considered for $s = 1$ and $s = 5$ respectively, due to convergence problems in the landmark model estimation.

Figures 5.1a and 5.1b depict respectively the distribution over the subjects of the relative bias of the estimator and the coverage rates of its 95% confidence interval both for the joint and two-stage landmark CS PH models for landmark times $s = 1$ and $s = 5$ and one horizon time $w = 3$. The box plots highlight the correct estimation of $\pi_*^k(s, w; \theta_0)$, except for the conditional expression from the joint model in the earlier landmark times ($s = 1$). This confirms that considering the modes of the distributions a posteriori of the random effects (defined in Section 5.7.1 of the Supplementary Material) in the conditional estimator is valid only when there is enough longitudinal information. The coverage rates which are very close to 0.95 validate the proposed 95% confidence interval computations

for both approaches. Finally the comparison of the widths of the 95% confidence intervals according to the joint and two-stage landmark CS PH models (Figure 5.1c) confirms that the joint model estimator is much more efficient than the landmark CS PH estimator. This result was expected because the included information in the landmark models is lower than the one in the joint model.

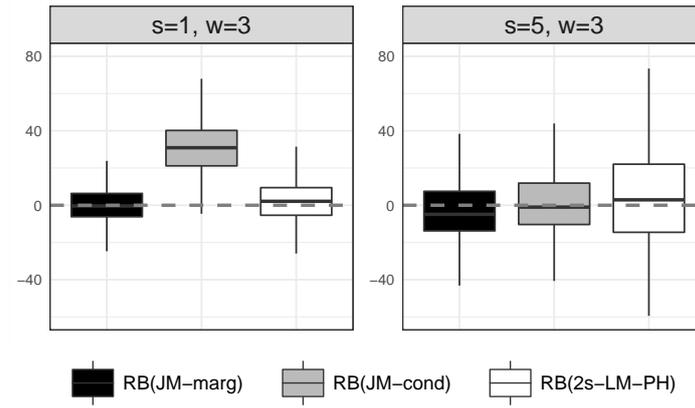
5.4.2 Simulation study II: Robustness to models hypotheses

The second simulation study aimed to compare the performances of the different approaches to provide individual dynamic predictions. We relied for that on predictive accuracy and explored their robustness to the models hypotheses.

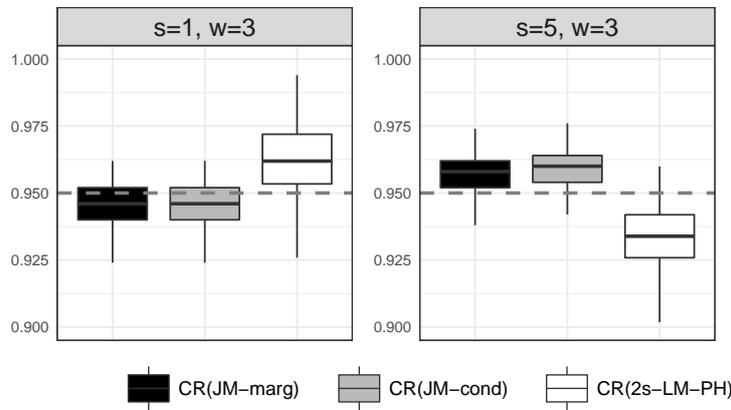
Predictive accuracy was assessed through both the Mean Squared Error of Prediction (MSEP) popularized through the Brier Score (BS) [Gerds and Schumacher, 2006] and the Area Under the ROC curve (AUC) [Heagerty *and others*, 2000]. The former assesses both calibration and discrimination abilities of the methods while the latter only focuses on discrimination ability and as such, neglects an important aspect of predictive accuracy [Blanche *and others*, 2015]. As in a simulation setting, we directly used the true individual prediction rather than the event indicator for the BS and did not have to deal with censoring. This lead to the computation of a standard MSEP on the validation sample :
$$\text{MSEP}_r^k(s, w) = \frac{1}{N^{\text{new}}(s)} \times \sum_{\star=1}^{N^{\text{new}}(s)} \left(\pi_{\star}^k(s, w; \theta_0) - \hat{\pi}_{\star, r}^k(s, w; \hat{\theta}) \right)^2.$$
 For the AUC, we applied the definition adapted to the competing risks setting [Blanche *and others*, 2015]:
$$\text{AUC}_r^k(s, w) = \Pr \left(\hat{\pi}_{i, r}^k(s, w; \hat{\theta}) > \hat{\pi}_{j, r}^k(s, w; \hat{\theta}) \mid \Delta_i^k(s, w) = 1, T_i > s, \Delta_j^k(s, w) = 0, T_j > s \right),$$
 where $\Delta_i^k(s, w) = \mathbb{1}\{s < T_i \leq s + w, \delta_i = k\}$ with $\delta_i = k$ the cause of event ; i and j are here two subjects for prediction (see Section 5.7.2 of the Supplementary Material for details on the AUC formulation). Note that both AUC and MSEP estimators were intrinsically model free as we did not have to deal with censoring.

We considered four scenarios: correct specification of the joint model, misspecification of the dependence function, violation of the proportional hazards assumption, and misspecification of the longitudinal trajectory of the marker. The distribution of the covariates and the coefficients used for the generation data in the four cases can be found in Section 5.7.3 of the Supplementary Material. Under each scenario, prediction models were compared two by two using boxplots of the differences in the predictive accuracy measures over the R replicates.

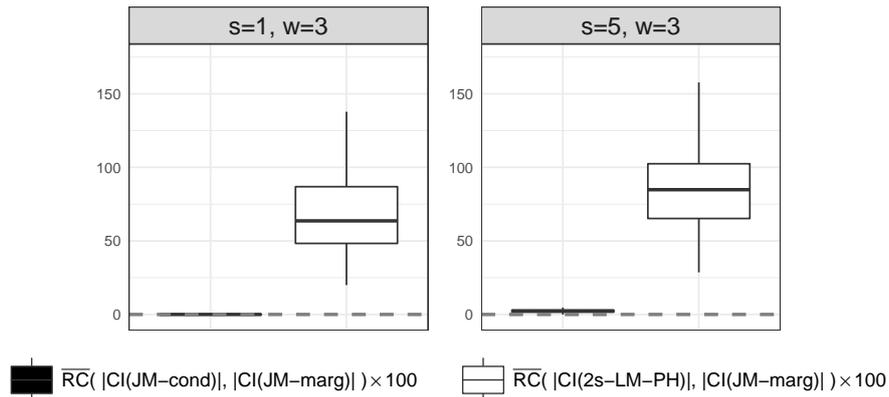
In the main manuscript, we present the results on MSEP and refer to the results on AUC which are detailed in Supplementary Material. For a given replicate r and a given landmark time s , only models that converged were considered.



(a) Evaluation of the estimators: distribution over the individuals $\star = 1, \dots, N(s)$ of the relative bias (RB, in %). The dashed line represents the 0.



(b) Evaluation of the confidence intervals: distribution over the individuals $\star = 1, \dots, 200$ of the coverage rates (CR) for the 95% confidence intervals of $\pi_{\star}^{\text{Rec.}}(s, w; \theta)$. The dashed line represents the 0.95.



(c) Evaluation of the estimator efficiency: distribution over the individuals $\star = 1, \dots, 200$ of the mean relative changes ($\overline{\text{RC}}$, in %) of the 95% confidence interval (CI) widths. The dashed line represents the 0.

Figure 5.1 – Evaluation of the estimators in terms of relative bias (a), coverage rates (b) and mean relative changes of the confidence intervals widths (c). Considered are the marginal estimator and the conditional estimator from the joint model (denoted JM-marg and JM-cond, respectively) and the estimator based on the two-stage cause-specific landmark model (denoted 2s-LM-PH).

Case 1: Correct specification of the joint model

For the well-specified case, data generation and specification of the joint and landmark models in the estimation and prediction steps were the same as in Section 5.4.1.

Figure 5.2 shows differences of MSEP for 8 pairs of landmark and horizon times ($s = 1, 3, 5, 8$ and $w = 1.5, 3$). As expected, the joint model performed better than the landmark models for all the pairs (s, w) . Once again, the conditional estimator of the predicted probability in the joint model was much worse than its marginal alternative in the earliest landmark times, but gave similar performances from $s = 5$.

When considering discriminatory power only with AUC (Figure 5.9 in Supplementary Material), the joint model still performed better than landmark models, especially the naive landmark model. However, no difference was highlighted between the conditional and marginal estimators suggesting a problem of calibration rather than discrimination for the conditional estimator in the earliest times. It can be noted that the convergence problems in the model estimations usually arose from insufficient considered information in landmarking.

Case 2: Misspecification of the dependence function

To investigate a misspecification of the dependence structure, we used the same generated data as in case 1 but the prediction models neglected the slope of the marker in the estimation and prediction steps. Yet the latter had a strong impact on the risk of recurrence.

The distributions over the replicates of the differences of MSEP for all the selected pairs of landmark and horizon times are depicted in Figure 5.3. Relative to the marginal estimator from the joint model, the other estimators behave similarly as in case 1. Mainly, the joint model remained better than the landmark models. However, neglecting the slope in the dependence structure induced a large increase in the MSEP of the marginal estimator from the joint model which is provided in each graph in Figure 5.3 and Figure 5.2; for instance, for $(s, w) = (1, 3)$, the MSEP increases from 0.323 to 1.114 when neglecting the slope in the dependence structure. This underlines the great importance of correctly specifying the dependence function in these models. The examination of AUC differences (Figure 5.11 in Supplementary Material) led to the same conclusions, except that the AUC under case 1 was not systematically better than the one under case 2.

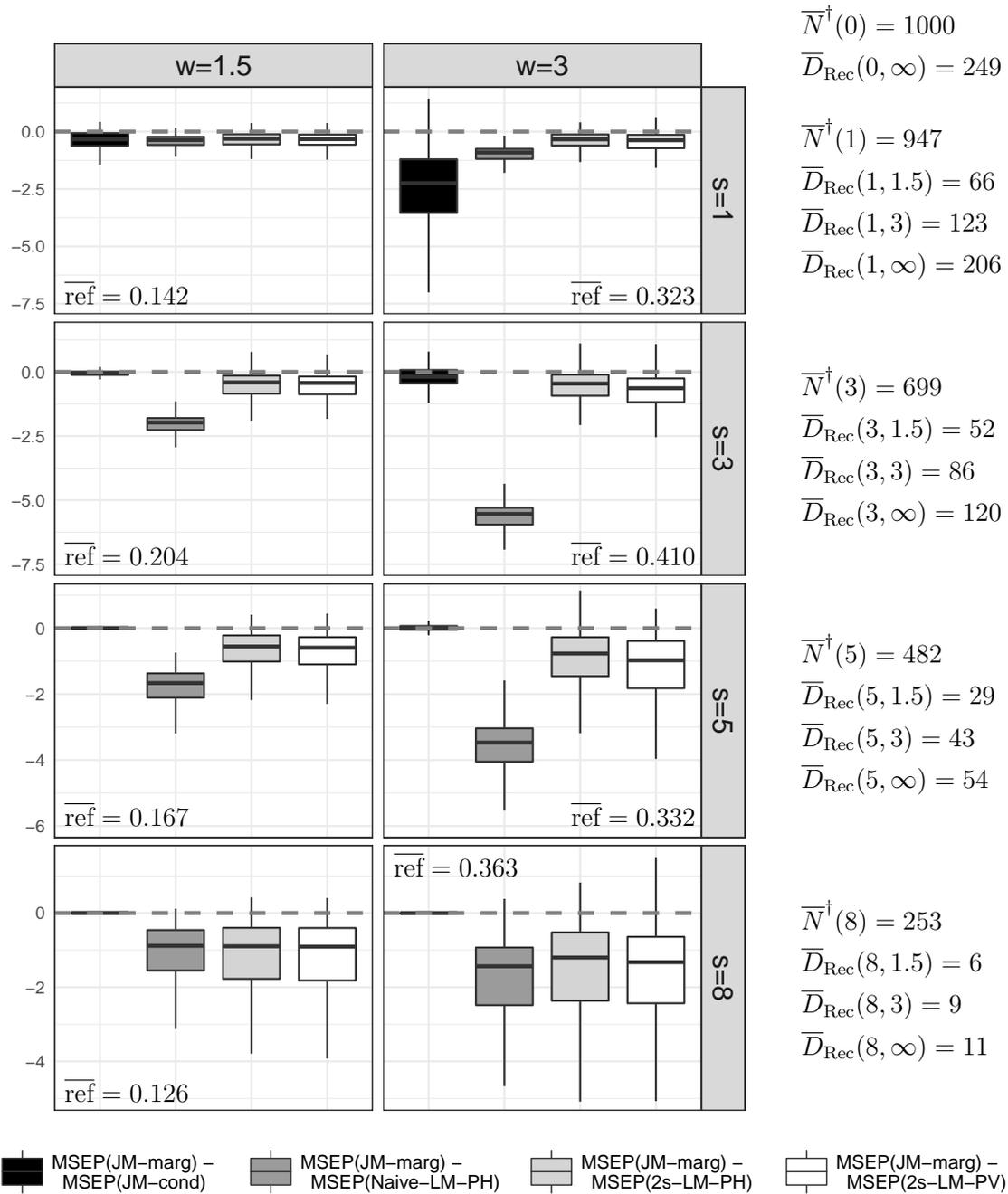


Figure 5.2 – Boxplots of the differences ($\times 1000$) of Mean Square Error of Prediction (MSEP) between the marginal estimator from the joint model (denoted JM-marg) and alternatives in the case of correct specification of the joint model (case 1). Considered are the conditional estimator from the joint model (JM-cond), the estimators from cause-specific landmark models using a two-stage or naive approach (2s-LM-PH and Naive-LM-PH, respectively) and the two-stage pseudo value model (2s-LM-PV). The distributions are depicted over $R = 499, 494, 486, 389$ replicates for 4 landmark times $s = 1, 3, 5, 8$ respectively, with 2 considered horizons $w = 1.5$ and $w = 3$. $\bar{\text{ref}}$ denotes the mean MSEP ($\times 1000$) using the marginal estimator from the joint model for each (s, w) . $\bar{N}^\dagger(s)$ is the mean number of subjects at risk at s and $\bar{D}_{\text{Rec}}(s, w)$ is the mean number of recurrences occurred between s and $s + w$.

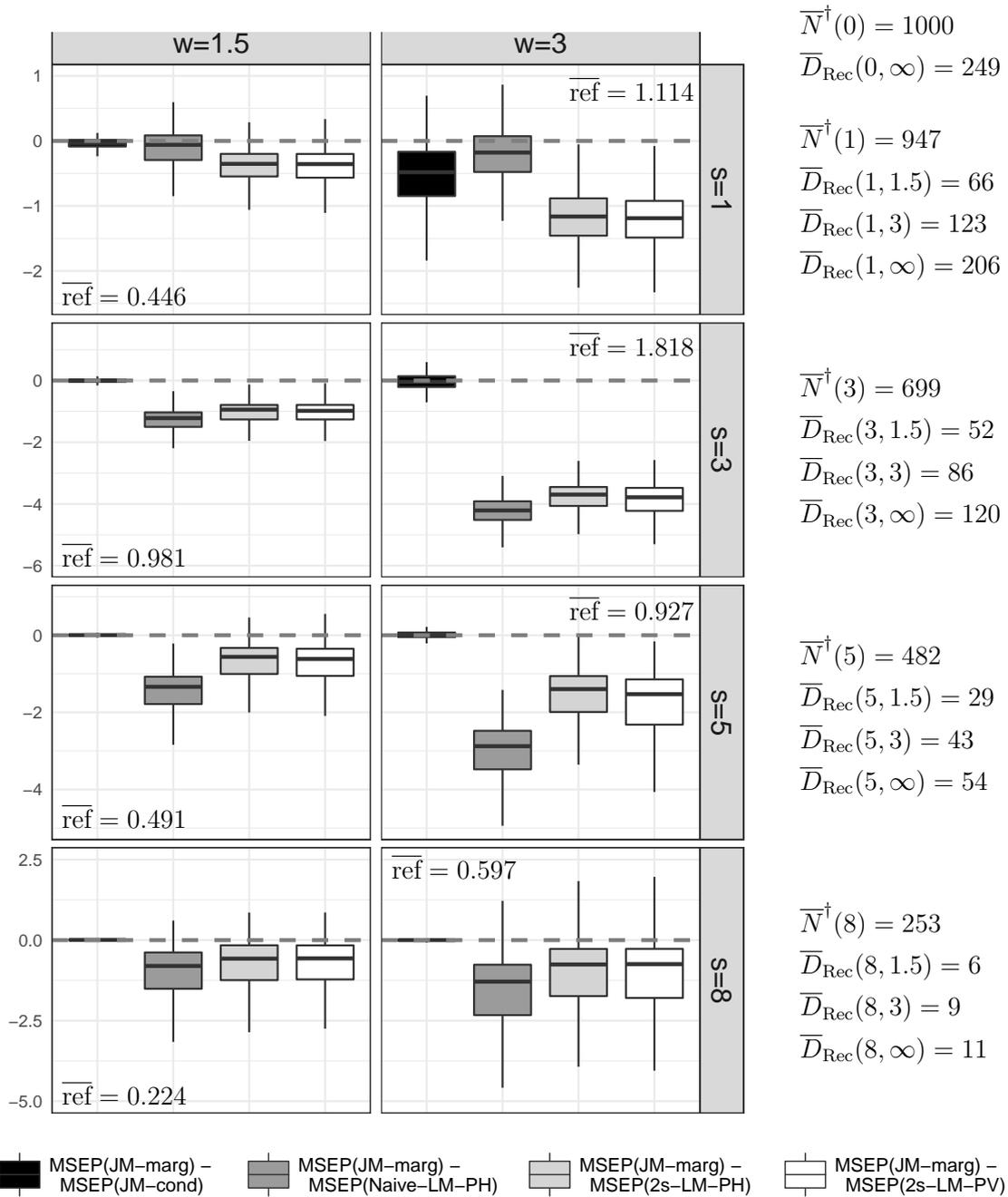


Figure 5.3 – Boxplots of the differences ($\times 1000$) of Mean Square Error of Prediction (MSEP) between the marginal estimator from the joint model (denoted JM-marg) and alternatives in the case of misspecification of the dependence function (case 2). Considered are the conditional estimator from the joint model (JM-cond), the estimators from cause-specific landmark models using a two-stage or naive approach (2s-LM-PH and Naive-LM-PH, respectively) and the two-stage pseudo value model (2s-LM-PV). The distributions are depicted over $R = 499, 498, 497, 428$ replicates for 4 landmark times $s = 1, 3, 5, 8$ respectively, with 2 considered horizons $w = 1.5$ and $w = 3$. $\bar{\text{ref}}$ denotes the mean MSEP ($\times 1000$) using the marginal estimator from the joint model for each (s, w) . $\bar{N}^\dagger(s)$ is the mean number of subjects at risk at s and $\bar{D}_{\text{Rec}}(s, w)$ is the mean number of recurrences occurred between s and $s + w$.

Case 3: Violation of the proportional hazards assumption

The robustness of the models to a violation of the proportional hazard assumption was checked by considering an interaction with $\log(1 + t)$ for the parameters associated with the marker dynamics in the generation model:

$$\left\{ \begin{array}{l} Y_i(t) = m_i(t) + \epsilon_i(t) \\ \quad = (\beta_0 + \beta_{0,X}X_i + b_{i0}) + (\beta_1 + \beta_{1,X}X_i + b_{i1})t + \epsilon_i(t), \\ \lambda_i^k(t) = \lambda_{k,0}(t) \exp \left\{ \gamma_k X_i + \eta_{1,k} \log(1 + t)m_i(t) + \eta_{2,k} \log(1 + t) \frac{\delta m_i(t)}{\delta t} \right\}. \end{array} \right.$$

For all the prediction models, the estimation and prediction steps did not consider this interaction with $\log(1 + t)$.

Boxplots of the differences of MSEF over the replicates are depicted in Figure 5.4. Even under this strong violation of the PH assumption, the performances of the two-stage landmark and joint models remained comparable. Furthermore one can note that the pseudo-value approach was not better than the models based on proportional hazards. Again, the same conclusions can be drawn from the differences in AUC (Figure 5.12 in Supplementary Material).

To illustrate the behavior of each model to this misspecification, Figure 5.5a depicts the time-varying coefficient $\eta_{1,k} * \log(1 + t)$ used in the data generation and the time-invariant parameters estimated in the joint and two-stage landmark CS PH models for one random replicate. The landmark model permitted to obtain estimated parameters closer to the generated one (except for $s = 8$ because only 8 and 13 subjects experienced the event between 8 and $8 + w$ for $w = 1.5$ and $w = 3$, respectively) but these estimates had also large variances because of the considered information which might explain the non superiority of landmark approaches to this misspecification.

Case 4: Misspecification of the longitudinal trajectory of the marker

The last case explored the performances of the prediction models when the longitudinal trend of the marker was misspecified. Data were generated using a joint model with a biphasic shape of the marker:

$$\left\{ \begin{array}{l} Y_i(t) = m_i(t) + \epsilon_i(t) \\ \quad = (\beta_0 + \beta_{0,X}X_i + b_{i0}) + (\beta_1 + \beta_{1,X}X_i + b_{i1})((1 + t)^{-1.2} - 1) + \\ \quad \quad (\beta_2 + \beta_{2,X}X_i + b_{i2})t + \epsilon_i(t), \\ \lambda_i^k(t) = \lambda_{k,0}(t) \exp \left\{ \gamma_k X_i + \eta_{1,k} m_i(t) + \eta_{2,k} \frac{\delta m_i(t)}{\delta t} \right\}. \end{array} \right.$$

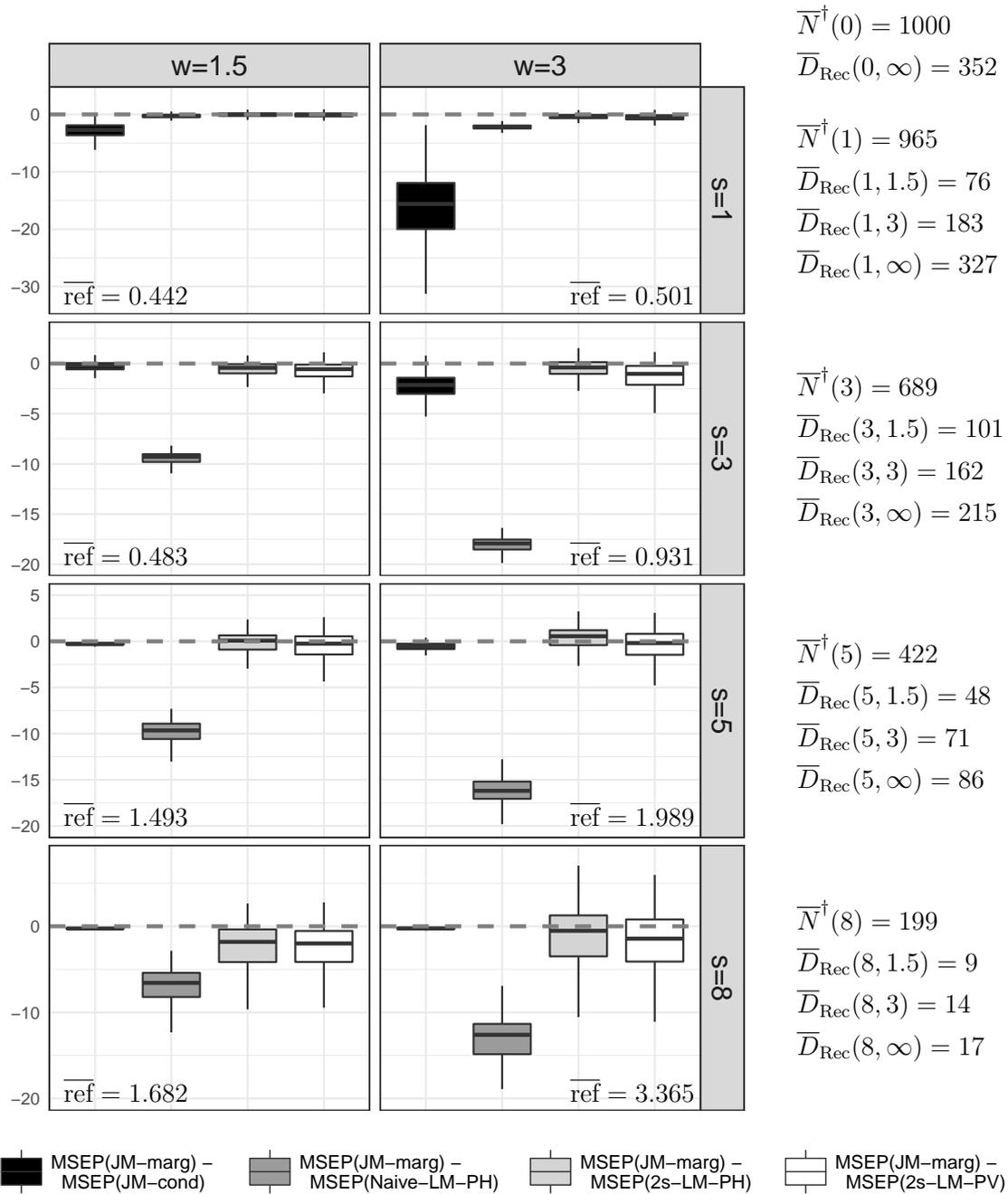
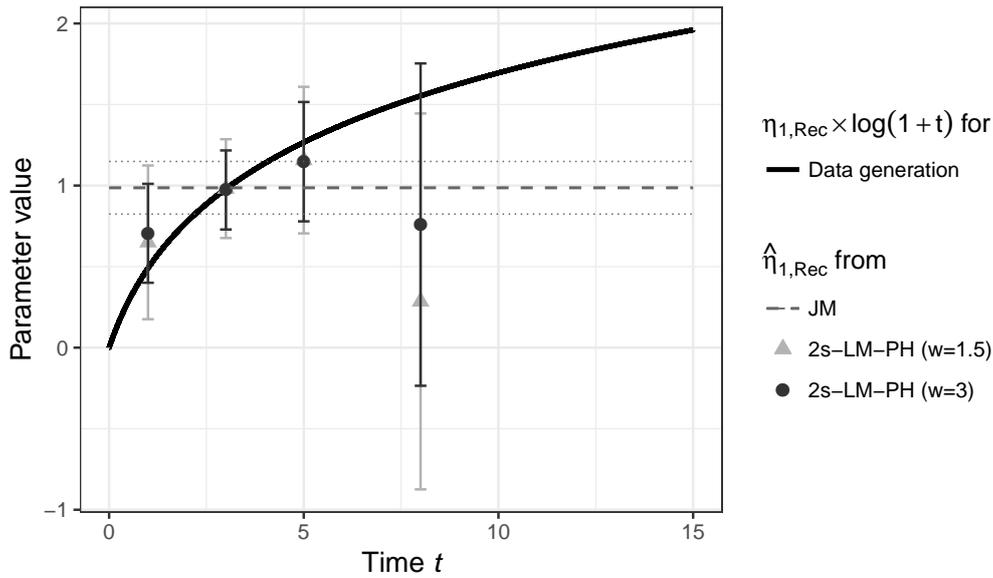
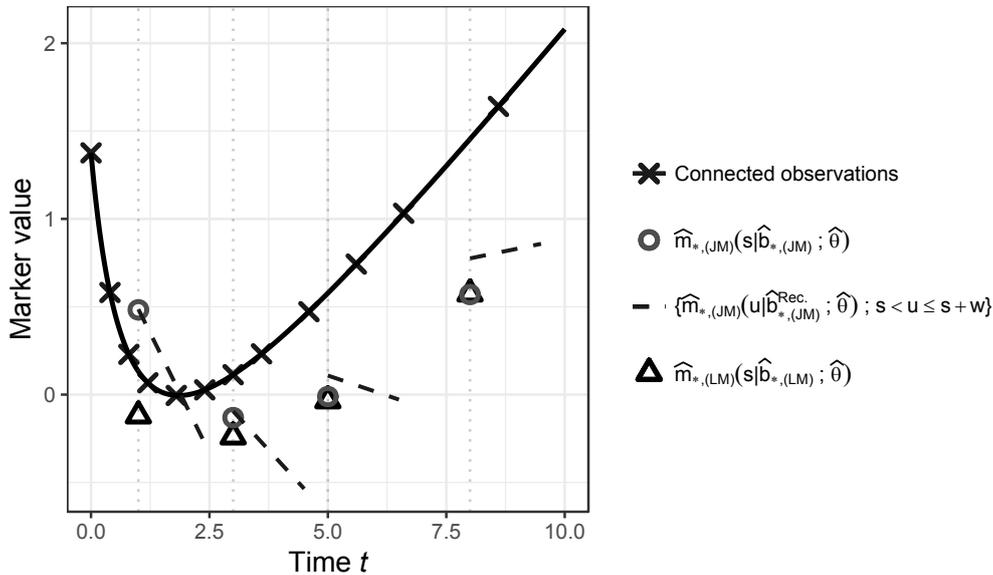


Figure 5.4 – Boxplots of the differences ($\times 1000$) of Mean Square Error of Prediction (MSEP) between the marginal estimator from the joint model (denoted JM-marg) and alternatives in the case of substantial violation of the PH assumption (case 3). Considered are the conditional estimator from the joint model (JM-cond), the estimators from cause-specific landmark models using a two-stage or naive approach (2s-LM-PH and Naive-LM-PH, respectively) and the two-stage pseudo value model (2s-LM-PV). The distributions are depicted over $R = 485, 326, 294, 188$ replicates for 4 landmark times $s = 1, 3, 5, 8$ respectively, with 2 considered horizons $w = 1.5$ and $w = 3$. $\bar{\text{ref}}$ denotes the mean MSEP ($\times 1000$) using the marginal estimator from the joint model for each (s, w) . $\bar{N}^\dagger(s)$ is the mean number of subjects at risk at s and $\bar{D}_{\text{Rec}}(s, w)$ is the mean number of recurrences occurred between s and $s + w$.



(a) Example of log relative risk of recurrence associated with the current marker level in the generation data with a time-varying effect ($\eta_{1,Rec} \times \log(1 + t)$). Are also depicted the estimated effects ($\hat{\eta}_{1,Rec}$) in the joint model (JM) and the two-stage cause specific landmark models (2s-LM-PH) with horizons $w = 1.5$ and $w = 3$, and their associated estimated 95% confidence intervals.



(b) Example of nonlinear marker trajectory considered in the generation data for a mean subject \star with $b_\star = 0$ and $X_\star = 2.04$. Are also represented the current values of the marker actually predicted in the estimation models (denoted $\hat{m}_*(JM)$ for the joint model and $\hat{m}_*(LM)$ for the two-stage landmark models) which are based for the joint model on $\hat{b}_{\star,(JM)} = \arg \max_b f(b|T_\star > s, \mathcal{Y}_\star(s), \mathcal{X}_\star(s); \hat{\theta})$ in the denominator of the conditional estimator definition and $\hat{b}_{\star,(JM)}^{Rec} = \arg \max_b f(b|T_\star > s, \delta_\star = Rec., \mathcal{Y}_\star(s), \mathcal{X}_\star(s); \hat{\theta})$ in the numerator (see formula in Section 5.7.1 of the Supplementary Material), and for the two-stage landmark models on $\hat{b}_{\star,(LM)} = \mathbb{E}(b_\star | \mathcal{Y}_\star(s), \mathcal{X}_\star(s); \hat{\theta})$.

Figure 5.5 – Illustrative examples of the model behaviors on a randomly selected replicate for misspecified cases 3 and 4.

For the estimation of the predicted probabilities of event using joint and two-stage landmark models, we considered a linear trajectory over time for the marker. As shown in Figure 5.5b, the degree of misspecification of the longitudinal marker trend was severe but it was made on purpose to clearly show the impact of such misspecification.

Figure 5.6 displays the boxplots of differences in MSE_P for the 8 pairs (s, w) . The landmark models performed much better than the joint models for landmark times $s = 1, 3, 5$; at landmark time $s = 8$, performances of joint and landmark models became roughly similar. Such result was expected. The joint model incorrectly assumed a linear trajectory for the marker on the whole follow-up while the landmark model, by considering only the longitudinal information collected until s , assumed a linear trajectory only until s which was more realistic at earliest landmark times even if still far from being well specified. The same conclusions can be drawn from the examination of the AUC differences (Figure 5.10 in Supplementary Material), except that even at the latest landmark time (8 years), the landmark models remained much more discriminatory than the joint model.

Figure 5.5b illustrates the differences in the predicted current values of the marker used in the estimated models to predict the risk of recurrence for a hypothetical subject \star with $b_\star = 0$ and $X_\star = 2.04$ in a randomly selected replication. The difference between the predicted levels of the marker are very different at the earliest landmark times between the joint and landmark models while they get closer at $s = 8$. Moreover the joint model actually uses the predicted current level of the marker rather than the predicted marker value in s which does not necessarily follows the generated path.

To explore whether such differences were due to the severe misspecification of our example, we considered a second longitudinal marker trend in Section 5.7.2 of the Supplementary Material. This supplementary case considered a small degree of misspecification of the longitudinal marker by considering some slight fluctuations with splines in the generation model compared to the wellspecified case 1. Although slightly misspecified, the superiority of joint model over landmark approaches previously found in case 1 almost completely disappeared. This confirmed the high sensitivity to any kind of misspecification of the marker trajectory in the joint model.

5.5 Discussion

With the development of personalized medicine, it is important to provide valid and powerful tools to clinicians for the computation of individual probabilities of specific events such as landmark conditional cumulative incidences. These predictions are expected to

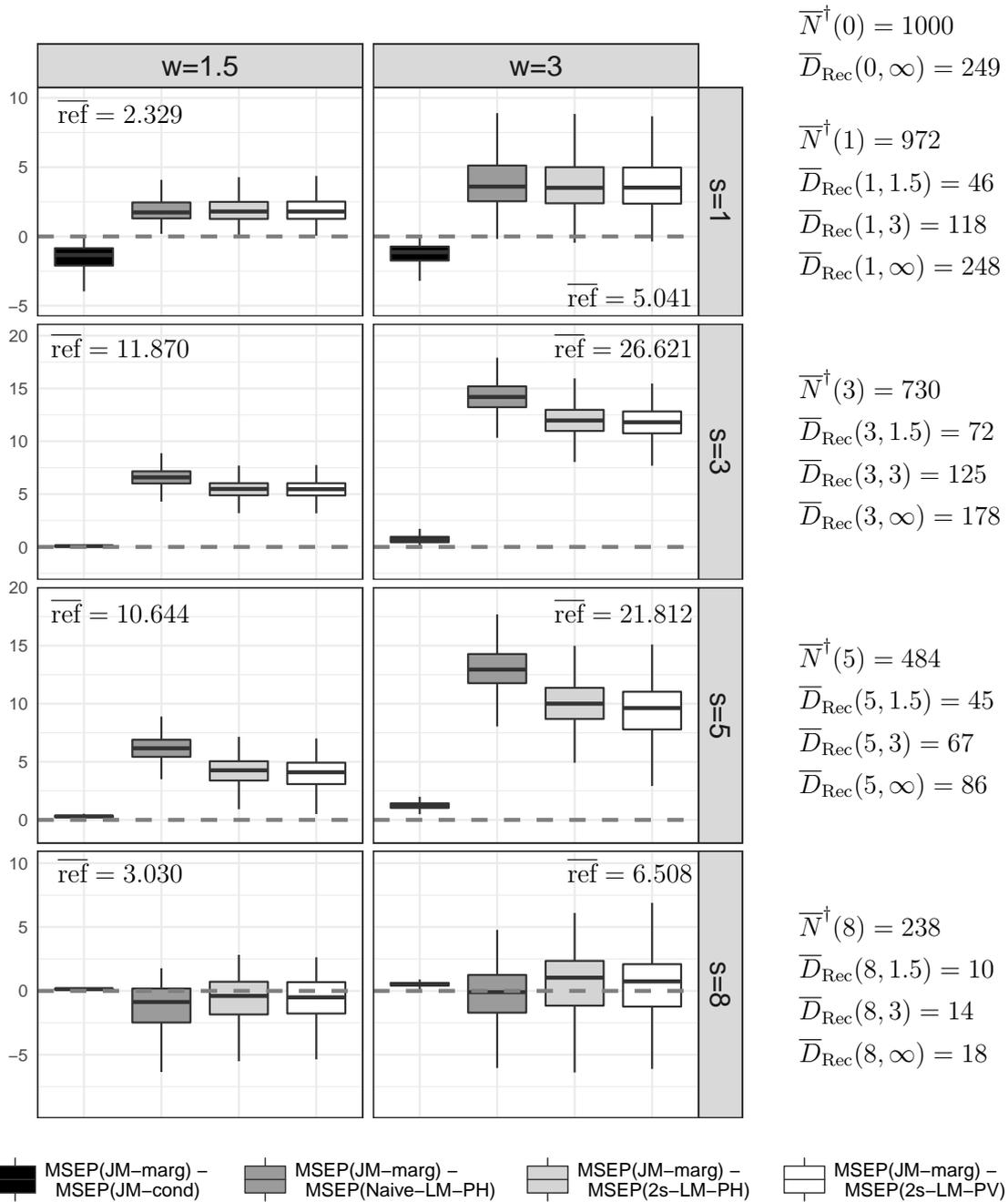


Figure 5.6 – Boxplots of the differences ($\times 1000$) of Mean Square Error of Prediction (MSEP) between the marginal estimator from the joint model (denoted JM-marg) and alternatives in the case of substantial misspecification of the longitudinal marker trajectory (case 4). Considered are the conditional estimator from the joint model (JM-cond), the estimators from cause-specific landmark models using a two-stage or naive approach (2s-LM-PH and Naive-LM-PH, respectively) and the two-stage pseudo value model (2s-LM-PV). The distributions are depicted over $R = 500, 496, 476, 357$ replicates for 4 landmark times $s = 1, 3, 5, 8$ respectively, with 2 considered horizons $w = 1.5$ and $w = 3$. $\overline{\text{ref}}$ denotes the mean MSEP ($\times 1000$) using the marginal estimator from the joint model for each (s, w) . $\overline{N}^\dagger(s)$ is the mean number of subjects at risk at s and $\overline{D}_{\text{Rec}}(s, w)$ is the mean number of recurrences occurred between s and $s + w$.

be used in clinical practice, notably to adapt individual strategies of treatment or to plan the patient-specific optimal screening time in clinical trials.

Several authors [Rizopoulos, 2011; Maziarz et al., 2017] already proposed estimators of the individual landmark conditional cumulative incidence $\pi_{\star}^k(s, w)$, but surprisingly none was formally validated using simulation studies. Our first objective was thus to formally define the quantity of interest $\pi_{\star}^k(s, w)$ and provide estimators (along with 95% confidence interval) both for the landmarking and the joint modelling approaches and properly validate them by comparing generated and estimated expressions of $\pi_{\star}^k(s, w)$. The computation of the generated quantity of interest was not obvious because it involved an integral over the latent structure shared by the longitudinal process and the survival process, the data being generated from a joint model. Note that in some papers, such quantity of interest is not correctly defined [Barrett and Su, 2017; Rizopoulos *and others*, 2017; Sweeting et al., 2016]. The marginal estimator from the joint model obtained very good performance in general whereas it was showed that the conditional estimator from the joint model necessitates sufficient individual longitudinal information collected until the prediction time.

Quantification of the uncertainty around individual predictions is essential for the decision making in clinical practice. In the landmark approach no solution was ever proposed, and a vagueness prevailed in the joint modelling literature with some definitions conditional on the observations [Proust-Lima and Taylor, 2009; Taylor et al., 2013; Król et al., 2016; Maziarz et al., 2017], others taking into account the measurement error of observations either along with the population parameter uncertainty [Yu et al., 2008; Rizopoulos, 2011] or without [Desmée *and others*, 2017], and finally many skipping the uncertainty issue as in landmarking [Proust-Lima and Taylor, 2009; Sweeting et al., 2016; Rizopoulos *and others*, 2017]. We thus introduced two definitions of uncertainty (conditioned or not on the observations) and proposed corresponding Monte Carlo methods to compute them in a unified manner for joint and landmark models. Compared to the joint model, the estimator based on the two-stage landmark cause-specific proportional hazards model confirmed its expected poor efficiency, with wide confidence intervals when only a few subjects experienced the event in the prediction window. Note that, although we chose to validate the estimators based on the confidence interval conditioned on the observations (using the corresponding adequate simulation setting), it would appear more natural to also account for the error of measurement of the biomarker history when providing confidence bands for individual follow-up.

Our second objective was to properly compare the landmark models and the joint model through several cases of well- and mis-specification. Indeed, a series of papers showed comparisons of prediction models in dynamic predictions [Goldstein et al., 2017; Huang *and others*, 2016; Sweeting et al., 2016] but none of them evaluated their robustness to misspecifications although most proposed methods were parametric. To our knowledge, only one contribution explored the problem of misspecification of prediction models (only longitudinal trajectory and functional dependency) very recently in a small simulation study and concluded to the superiority of joint models approach over landmark models [Rizopoulos *and others*, 2017].

In our extensive simulation study, we found that in the case of correctly specified model, the joint model performed better than the landmark models, as expected. In the case of misspecification of the dependence structure between the longitudinal process and the survival process, the difference of performances between approaches did not change as also concluded by Rizopoulos *and others* [2017]. But more importantly, the performance of misspecified models was much worse in terms of prediction error. Regarding the PH assumption, the two landmark models we proposed better dealt with this assumption than the joint model: dynamic pseudo-values did not require the PH assumption at all, and our cause-specific hazards landmark models limited the PH assumption to the prediction window with an administrative censoring at the end of it. Yet, the impact of PH assumption violation on the estimators derived from the joint model remained limited, suggesting that the violation of the PH assumption should be extreme to entail a tangible impact on the estimated cumulative incidences in the joint model. Finally, we showed that the correct specification of the marker trajectory was essential to provide good predictions with joint models (and with landmark models to a much lesser extent). We demonstrated the major loss of performance of the joint model in a severe case of misspecification to illustrate the limit but we also found in Supplementary Material that even a slight misspecification of the trajectory (usually considered as acceptable) impacted the prediction error of the models mainly, and eliminated for example the gain of using the joint model over the landmark model at shorter landmark times or when the horizon time increased. The previous simulation study also suggested this lack of robustness [Rizopoulos *and others*, 2017] although not emphasized.

As usual in prediction model development, comparisons were made in terms of both Mean Square Error of Prediction which measures a trade-off between calibration and discrimination, and Area Under the ROC Curve which only targets discriminatory power.

With the perspective in mind of providing quantified individual predictions, we mostly relied on the prediction error to assess both calibration and discrimination even though most conclusions were also drawn from AUC examinations, yet sometimes to a lesser extent probably explained by the lesser sensitivity of AUC [Pencina *and others*, 2008] and the possibly preserved discriminatory power in the presence of worse calibration.

To conclude with recommendations, we emphasize the need to carefully define the quantity of interest, its estimator and the type of associated uncertainty. The several cases of misspecification warned us on the necessity to precisely specify the dependence structure between the longitudinal marker dynamics and the risk of event. Finally, the specification of the longitudinal marker trend should be studied with extreme care, especially when using joint modelling. Researchers should be warned that the use of sophisticated methods such as the joint models may allow obtaining accurate and efficient estimators only when they are correctly specified. Otherwise, estimators might be off the mark. Landmark models seem less sensitive to the misspecification of the longitudinal marker trajectory but are as sensitive as joint models regarding the dependence structure. In addition, they provide considerably less efficient estimators and may induce convergence problems, notably when the landmark time increases and thus the considered information is too poor.

Acknowledgements

The authors thank warmly Boris P. Hejblum for his valuable advice and assistance. This work was supported by a joint grant from INSERM and Région Aquitaine, a grant from the Institut OpenHealth and a grant from the Réseau Franco-Néerlandais. Computer time for this study was provided by the computing facilities MCIA (Mésocentre de Calcul Intensif Aquitain) of the Université de Bordeaux and of the Université de Pau et des Pays de l'Adour.

5.6 References

- ANDERSEN, PER KRAGH, BORGAN, ORNULF, GILL, RICHARD D AND KEIDING, NEILS. *Statistical Models Based On Counting Processes*. Springer-Verlag, New York, 1993. 47, 53, 56, 69, 143
- ANDERSEN, PER KRAGH AND POHAR PERME, MAJA. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research* 19(1), 71–99, 2010. 52, 134, 143, 144

- BARRETT, JESSICA AND SU, LI. Dynamic predictions using flexible joint models of longitudinal and time-to-event data. *Statistics in Medicine* 36(9), 1447–1460, 2017. 157, 200
- BLANCHE, PAUL, PROUST-LIMA, CÉCILE, LOUBÈRE, LUCIE, BERR, CLAUDINE, DARTIGUES, JEAN-FRANÇOIS AND JACQMIN-GADDA, HÉLÈNE. Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics* 71(1), 102–113, 2015. 147, 165
- BRESLOW, NORMAN E. Discussion of Professor Cox's paper. *Journal of the Royal Statistical Society, Series B* 34, 216–217, 1972. 51, 142
- COX, DAVID R. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* 34, 187–220, 1972. 34, 51, 141
- CROWTHER, MICHAEL J AND LAMBERT, PAUL C. Simulating biologically plausible complex survival data. *Statistics in Medicine* 32(23), 4118–4134, 2013. 112, 145
- DESMÉE, SOLÈNE, MENTRÉ, FRANCE, VEYRAT-FOLLET, CHRISTINE, SÉBASTIEN, BERNARD AND GUEDJ, JÉRÉMIE. Nonlinear joint models for individual dynamic prediction of risk of death using hamiltonian monte carlo: application to metastatic prostate cancer. *BMC Medical Research Methodology* 17(1), 105, 2017. 157
- FERRER, LOÏC, RONDEAU, VIRGINIE, DIGNAM, JAMES, PICKLES, TOM, JACQMIN-GADDA, HÉLÈNE AND PROUST-LIMA, CÉCILE. Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer. *Statistics in Medicine* 35(22), 3933–3948, 2016. 70, 106, 107, 109, 112, 114, 115, 116, 119, 122, 130, 139, 145, 163
- FINE, JASON P AND GRAY, ROBERT J. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* 94(446), 496–509, 1999. 143
- GERDS, THOMAS A AND SCHUMACHER, MARTIN. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal* 48(6), 1029–1040, 2006. 147
- GOLDSTEIN, BENJAMIN A, POMANN, GINA MARIA, WINKELMAYER, WOLFGANG C AND PENCINA, MICHAEL J. A comparison of risk prediction methods using repeated observations: an application to electronic health records for hemodialysis. *Statistics in Medicine* 36, 2750—2763, 2017. 35, 136, 158
- HEAGERTY, PATRICK J, LUMLEY, THOMAS AND PEPE, MARGARET S. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics* 56(2), 337–344, 2000. 147
- HUANG, XUELIN, YAN, FANGRONG, NING, JING, FENG, ZIDING, CHOI, SANGBUM AND CORTES, JORGE. A two-stage approach for dynamic prediction of time-to-event distributions. *Statistics in Medicine* 35, 2167–2182, 2016. 136, 158
- JEWELL, NICHOLAS P AND NIELSEN, JENS P. A framework for consistent prediction rules based on markers. *Biometrika* 80, 153–164, 1993. 133, 136
- KALBFLEISCH, JOHN D AND PRENTICE, ROSS L. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, 2011. 34, 60, 136

INDIVIDUAL DYNAMIC PREDICTIONS

- KRÓL, AGNIESZKA, FERRER, LOÏC, PIGNON, JEAN-PIERRE, PROUST-LIMA, CÉCILE, DUCREUX, MICHEL, BOUCHÉ, OLIVIER, MICHIELS, STEFAN AND RONDEAU, VIRGINIE. Joint model for left-censored longitudinal data, recurrent events and terminal event: Predictive abilities of tumor burden for cancer evolution with application to the ffd 2000–05 trial. *Biometrics* 72(3), 907–916, 2016. 35, 67, 106, 138, 157, 204
- MAZIARZ, MARLENA, HEAGERTY, PATRICK, CAI, TIANXI AND ZHENG, YINGYE. On longitudinal prediction with time-to-event outcome: Comparison of modeling options. *Biometrics* 73(1), 83–93, 2017. 35, 36, 134, 137, 138, 157, 206
- NICOLAIE, MA, VAN HOUWELINGEN, JC, DE WITTE, TM AND PUTTER, H. Dynamic pseudo-observations: A robust approach to dynamic prediction in competing risks. *Biometrics* 69(4), 1043–1052, 2013. 36, 137, 143, 206
- PENCINA, MICHAEL J, D’AGOSTINO, RALPH B AND VASAN, RAMACHANDRAN S. Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. *Statistics in Medicine* 27(2), 157–172, 2008. 159
- PROUST-LIMA, CÉCILE AND TAYLOR, JEREMY MG. 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, 2009. 134, 136, 137, 138, 145, 157
- PROUST-LIMA, CÉCILE, TAYLOR, JEREMY MG, WILLIAMS, SCOTT G, ANKERST, DONNA P, LIU, NING, KESTIN, LARRY L, BAE, KYOUNGHWAN AND SANDLER, HOWARD M. Determinants of change in prostate-specific antigen over time and its association with recurrence after external beam radiation therapy for prostate cancer in five large cohorts. *International Journal of Radiation Oncology Biology Physics* 72(3), 782–791, 2008. 32, 115, 117, 145
- RIZOPOULOS, DIMITRIS. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* 67(3), 819–829, 2011. 36, 134, 137, 139, 157, 206
- RIZOPOULOS, DIMITRIS. *Joint Models for Longitudinal and Time-to-Event Data: with Applications in R*. Chapman & Hall/CRC, Boca Raton, 2012. 63, 65, 110, 139, 201
- RIZOPOULOS, DIMITRIS, MOLENBERGHS, GEERT AND LESAFFRE, EMMANUEL. Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal* (in press), 2017. 134, 137, 157, 158
- RIZOPOULOS, DIMITRIS, TAYLOR, JEREMY MG, VAN ROSMALEN, JOOST, STEYERBERG, EWOUT W AND TAKKENBERG, JOHANNA JM. Personalized screening intervals for biomarkers using joint models for longitudinal and survival data. *Biostatistics* 17, 149–164, 2015. 32, 136, 208
- SCHEIKE, THOMAS H, ZHANG, MEI-JIE AND GERDS, THOMAS A. Predicting cumulative incidence probability by direct binomial regression. *Biometrika* 95(1), 205–220, 2008. 52, 134, 143
- SÈNE, MBÉRY, TAYLOR, JEREMY MG, DIGNAM, JAMES J, JACQMIN-GADDA, HÉLÈNE AND PROUST-LIMA, CÉCILE. Individualized dynamic prediction of prostate cancer recurrence with and without the initiation of a second treatment: Development and validation. *Statistical Methods in Medical Research* 25(6), 2972–2991, 2016. 32, 36, 134, 136, 137, 208
- SINNOTT, JENNIFER A AND CAI, TIANXI. Inference for survival prediction under the regularized cox model. *Biostatistics* 17(4), 692–707, 2016. 142

- SURESH, KRITHIKA, TAYLOR, JEREMY MG, SPRATT, DANIEL E, DAIGNAULT, STEPHANIE AND TSODIKOV, ALEXANDER. Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model. *Biometrical Journal* (in press), 2017. 133, 136
- SWEETING, MICHAEL J, BARRETT, JESSICA K, THOMPSON, SIMON G AND WOOD, ANGELA M. The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the aric study. *Statistics in Medicine* (in press), 2016. 134, 137, 157, 158
- TAYLOR, JEREMY MG, PARK, YONGSEOK, ANKERST, DONNA P, PROUST-LIMA, CECILE, WILLIAMS, SCOTT, KESTIN, LARRY, BAE, KYOUNGWHA, PICKLES, TOM AND SANDLER, HOWARD. Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics* 69(1), 206–213, 2013. 115, 138, 145, 157
- TSIATIS, ANASTASIOS A AND DAVIDIAN, MARIE. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* 14(3), 809–834, 2004. 35, 61, 62, 106, 136, 192
- VAN HOUWELINGEN, HANS C. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics* 34(1), 70–85, 2007. 36, 59, 136, 206
- YU, MENGANG, TAYLOR, JEREMY M G AND SANDLER, HOWARD M. Individual prediction in prostate cancer studies using a joint longitudinal survival–cure model. *Journal of the American Statistical Association* 103(481), 178–187, 2008. 63, 139, 157

5.7 Supplementary material

5.7.1 Complete formulas of the estimators

This section details the formulations of the estimators defined in the main manuscript from the landmark cause-specific proportional hazards models (Section 5.2.3) and the joint model (Section 5.2.2). Refer to the main manuscript for the notations. Please note that all these estimators involve the computation of integrals over time that can be avoided using product-integrals. Detailed information on this operation can be found in the supplementary material of Ferrer et al. [2016].

Joint model: marginal estimator

$$\begin{aligned} \hat{\pi}_*^k(s, w; \hat{\theta}) &= \frac{\int_{\mathbb{R}^q} \Pr(s < T_* \leq s + w, \delta_* = k | \mathcal{X}_*(s), b_*; \hat{\theta}) \times \\ &\quad f(\mathcal{Y}_*(s) | \mathcal{X}_*(s), b_*; \hat{\theta}) f(b_* | s < T_* \leq s + w, \delta_* = k, \mathcal{Y}_*(s), \mathcal{X}_*(s); \hat{\theta}) db_*}{\int_{\mathbb{R}^q} \Pr(T_* > s | \mathcal{X}_*(s), b_*; \hat{\theta}) f(\mathcal{Y}_*(s) | \mathcal{X}_*(s), b_*; \hat{\theta}) f(b_* | T_* > s, \mathcal{Y}_*(s), \mathcal{X}_*(s); \hat{\theta}) db_*} \\ &= \frac{\int_{\mathbb{R}^q} \left[\int_s^{s+w} \exp\left(-\sum_k \int_0^u \hat{\lambda}_i^k(\nu | \mathcal{X}_*(\nu), b_*; \hat{\theta}) d\nu\right) \hat{\lambda}_i^k(u | \mathcal{X}_*(u), b_*; \hat{\theta}) du \right] \times \\ &\quad f(\mathcal{Y}_*(s) | \mathcal{X}_*(s), b_*; \hat{\theta}) f(b_* | s < T_* \leq s + w, \delta_* = k, \mathcal{Y}_*(s), \mathcal{X}_*(s); \hat{\theta}) db_*}{\int_{\mathbb{R}^q} \exp\left(-\sum_k \int_0^s \hat{\lambda}_i^k(u | \mathcal{X}_*(u), b_*; \hat{\theta}) du\right) f(\mathcal{Y}_*(s) | \mathcal{X}_*(s), b_*; \hat{\theta}) f(b_* | T_* > s, \mathcal{Y}_*(s), \mathcal{X}_*(s); \hat{\theta}) db_*} \end{aligned}$$

Joint model: conditional estimator

$$\begin{aligned} \hat{\pi}_*^k(s, w; \hat{\theta}) &= \frac{\Pr(s < T_* \leq s + w, \delta_* = k | \mathcal{X}_*(s), \hat{b}_*^k; \hat{\theta}) f(\mathcal{Y}_*(s) | \mathcal{X}_*(s), \hat{b}_*^k; \hat{\theta})}{\Pr(T_* > s | \mathcal{X}_*(s), \hat{b}_*^k; \hat{\theta}) f(\mathcal{Y}_*(s) | \mathcal{X}_*(s), \hat{b}_*^k; \hat{\theta})} \\ &= \frac{\left[\int_s^{s+w} \exp\left(-\sum_k \int_0^u \hat{\lambda}_*^k(\nu | \mathcal{X}_*(\nu), \hat{b}_*^k; \hat{\theta}) d\nu\right) \hat{\lambda}_*^k(u | \mathcal{X}_*(u), \hat{b}_*^k; \hat{\theta}) du \right] f(\mathcal{Y}_*(s) | \mathcal{X}_*(s), \hat{b}_*^k; \hat{\theta})}{\exp\left(-\sum_k \int_0^s \hat{\lambda}_*^k(u | \mathcal{X}_*(u), \hat{b}_*^k; \hat{\theta}) du\right) f(\mathcal{Y}_*(s) | \mathcal{X}_*(s), \hat{b}_*^k; \hat{\theta})} \end{aligned}$$

where $\hat{b}_* = \arg \max_b \{\log f(T_* > s, \mathcal{Y}_*(s), b | \mathcal{X}_*(s); \hat{\theta})\}$ and $\hat{b}_*^k = \arg \max_b \{\log f(s < T_* \leq s + w, \delta_* = k, \mathcal{Y}_*(s), b | \mathcal{X}_*(s); \hat{\theta})\}$.

Naive landmark cause-specific proportional hazards model: estimator

$$\begin{aligned}\widehat{\pi}_*^k(s, w; \widehat{\theta}) &= \Pr(s < T_* \leq s + w, \delta_* = k | T_* > s, \{X_{k,*}^E; k = 1, \dots, K\}, Y_*(t_{*n_*(s)}); \widehat{\theta}) \\ &= \int_s^{s+w} \exp\left(-\sum_k \int_s^u \widehat{\lambda}_*^k(\nu | X_{k,*}^E, Y_*(t_{*n_*(s)}); \widehat{\theta}) d\nu\right) \widehat{\lambda}_*^k(u | X_{k,*}^E, Y_*(t_{*n_*(s)}); \widehat{\theta}) du\end{aligned}$$

Two-stage landmark cause-specific proportional hazards model: estimator

$$\begin{aligned}\widehat{\pi}_*^k(s, w; \widehat{\theta}) &= \Pr(s < T_* \leq s + w, \delta_* = k | T_* > s, \mathcal{X}_*(s), \widehat{b}_*; \widehat{\theta}) \\ &= \int_s^{s+w} \exp\left(-\sum_k \int_s^u \widehat{\lambda}_*^k(\nu | \mathcal{X}_*(s), \widehat{b}_*; \widehat{\theta}) d\nu\right) \widehat{\lambda}_*^k(u | \mathcal{X}_*(u), \widehat{b}_*; \widehat{\theta}) du\end{aligned}$$

where $\widehat{b}_* = \mathbb{E}(b_* | \mathcal{Y}_*(s), \mathcal{X}_*(s); \widehat{\theta})$.

5.7.2 Complementary results for Simulation II: robustness assessment

Supplementary case: Weak misspecification of the longitudinal trajectory of the marker

This supplementary case aimed to check the performances of our proposed estimators when the longitudinal trend of the marker was misspecified. In contrast with the extreme case (case 4) in the main manuscript, we considered here a very slight misspecification of the longitudinal trajectory.

Data were generated using a joint model with a longitudinal marker evolution characterized by a combination of cubic B-splines functions:

$$\begin{cases} Y_i(t) &= m_i(t) + \epsilon_i(t) \\ &= (\beta_0 + \beta_{0,X} X_i + b_{i0}) + \sum_{l=1}^3 (\beta_l + \beta_{l,X} X_i + b_{il}) B_l(t, 3) + \epsilon_i(t), \\ \lambda_i^k(t) &= \lambda_{k,0}(t) \exp\left\{\gamma_k X_i + \eta_{1,k} m_i(t) + \eta_{2,k} \frac{\delta m_i(t)}{\delta t}\right\}. \end{cases}$$

As shown in the example of individual trajectory depicted in Figure 5.7, this longitudinal trend slightly differed from the linear evolution assumed in the estimation and prediction steps. The distribution of the covariates and the coefficients used for the data generation are detailed in Section 5.7.3 of this Supplementary Material.

Figure 5.8 shows differences of MSEP for 8 pairs of landmark and horizon times ($s = 1, 3, 5, 8$ and $w = 1.5, 3$). The joint marginal and landmark two-stage estimators were comparable with no significant differences for all the pairs of landmark and horizon

times, except slightly when $w = 1.5$ and $s = 3$ or $s = 5$. Indeed, considering a linear evolution in these times of prediction window does not seem to be incorrect, as illustrated in Figure 5.7. The results confirmed the poor accuracy of the conditional estimator from the joint model in the earlier landmark times, but this estimator became equivalent to the marginal estimator when the landmark time increased. Compared to the superiority of the joint model found in the well-specified case 1, these results confirm the high sensitivity of the estimators from the joint modelling approach to the correct specification of the longitudinal marker trend.

Assessment of model discrimination: computation of AUCs

When we are interested in the predictive accuracy of a model, a popular measure to check its performances in terms of discrimination is the area under the ROC curve, called AUC. We considered the AUC definition adapted to the competing risks setting [Blanche *and others*, 2015]:

$$\text{AUC}_r^k(s, w) = \Pr(\hat{\pi}_{i,r}^k(s, w; \hat{\theta}) > \hat{\pi}_{j,r}^k(s, w; \hat{\theta}) \mid \Delta_i^k(s, w) = 1, T_i > s, \Delta_j^k(s, w) = 0, T_j > s)$$

where $\Delta_i^k(s, w) = \mathbb{1}\{s < T_i \leq s + w, \delta_i = k\}$ with $\delta_i = k$ the cause of event ; i and j two subjects for prediction. Thus, for any subject i at risk at s , $\Delta_i^k(s, w) = 1$ when subject i experiences the event of cause k within the time interval $(s, s + w]$, and $\Delta_i^k(s, w) = 0$ when either he experiences the competing event between s and $s + w$ or is event-free at $s + w$.

Because this was a simulation study, the indicator of event $\Delta_i^k(s, w)$ was known and the estimator could directly be expressed as

$$\widehat{\text{AUC}}_r^k(s, w) = \frac{\sum_{i=1}^{N^{\text{new}}(0)} \sum_{j=1}^{N^{\text{new}}(0)} \mathbb{1}\{T_i > s\} \mathbb{1}\{T_j > s\} \mathbb{1}\{\hat{\pi}_{i,r}^k(s, w; \hat{\theta}) > \hat{\pi}_{j,r}^k(s, w; \hat{\theta})\} \Delta_i^k(s, w) (1 - \Delta_j^k(s, w))}{\sum_{i=1}^N \sum_{j=1}^N \mathbb{1}\{T_i > s\} \mathbb{1}\{T_j > s\} \Delta_i^k(s, w) (1 - \Delta_j^k(s, w))}.$$

Figures 5.9, 5.10, 5.11, 5.12 and 5.13 depict respectively the AUCs for the cases where the joint model is correctly specified (case 1), the dependence function is misspecified (case 2), the proportional hazards assumption is violated (case 3), the longitudinal trend of the marker is severely misspecified (case 4) or slightly misspecified (case 4.bis). Overall, these results were comparable to those obtained in terms of mean squared errors. The joint model discriminated lightly better than the two-stage landmark models when the joint model was well specified (Figure 5.9) or when the dependence function was misspecified (Figure 5.10), whereas in the case of strong violation of the PH assumption, these models

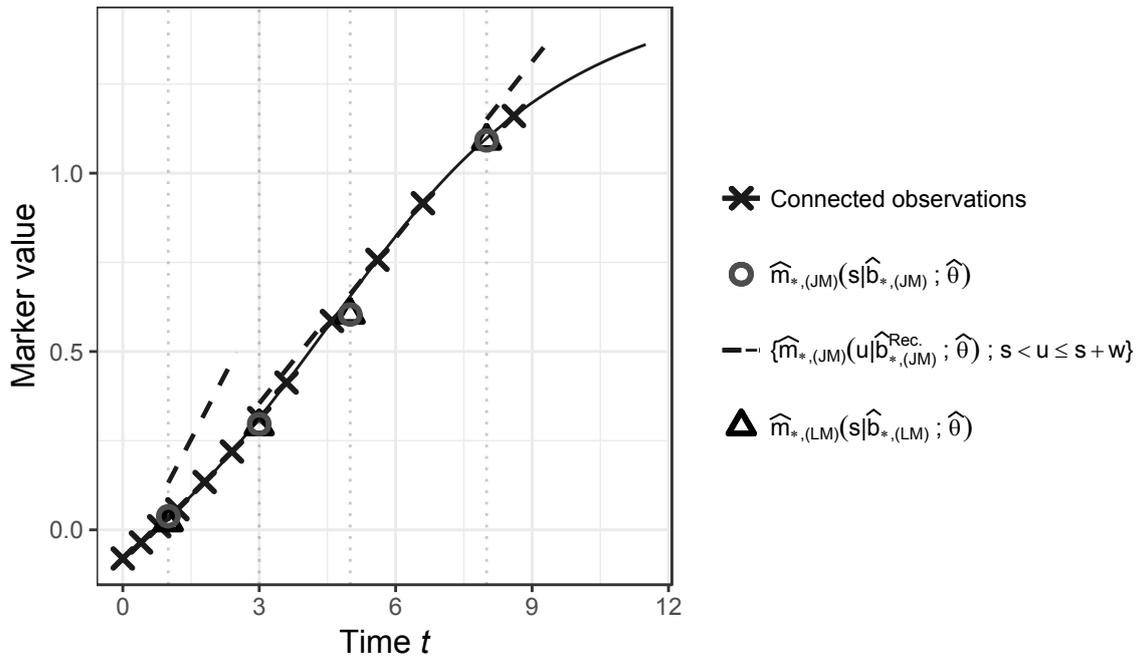


Figure 5.7 – Illustrative example of the model behavior on a randomly selected replicate for misspecified case 4.bis, with a slight nonlinear marker trajectory considered in the generation data for a mean subject \star with $b_\star = 0$ and $X_\star = 2.04$. Are also represented the current values of the marker actually predicted in the estimation models (denoted $\hat{m}_*(JM)$ for the joint model and $\hat{m}_*(LM)$ for the two-stage landmark models) which are based for the joint model on $\hat{b}_{*,(JM)} = \arg \max_b f(b | T_\star > s, \mathcal{Y}_\star(s), \mathcal{X}_\star(s); \hat{\theta})$ in the denominator of the conditional estimator definition and $\hat{b}_{*,(JM)}^{Rec.} = \arg \max_b f(b | T_\star > s, \delta_\star = Rec., \mathcal{Y}_\star(s), \mathcal{X}_\star(s); \hat{\theta})$ in the numerator (see formula in Section 1.2 of the Supplementary Material), and for the two-stage landmark models on $\hat{b}_{*,(LM)} = \mathbb{E}(b_\star | \mathcal{Y}_\star(s), \mathcal{X}_\star(s); \hat{\theta})$.

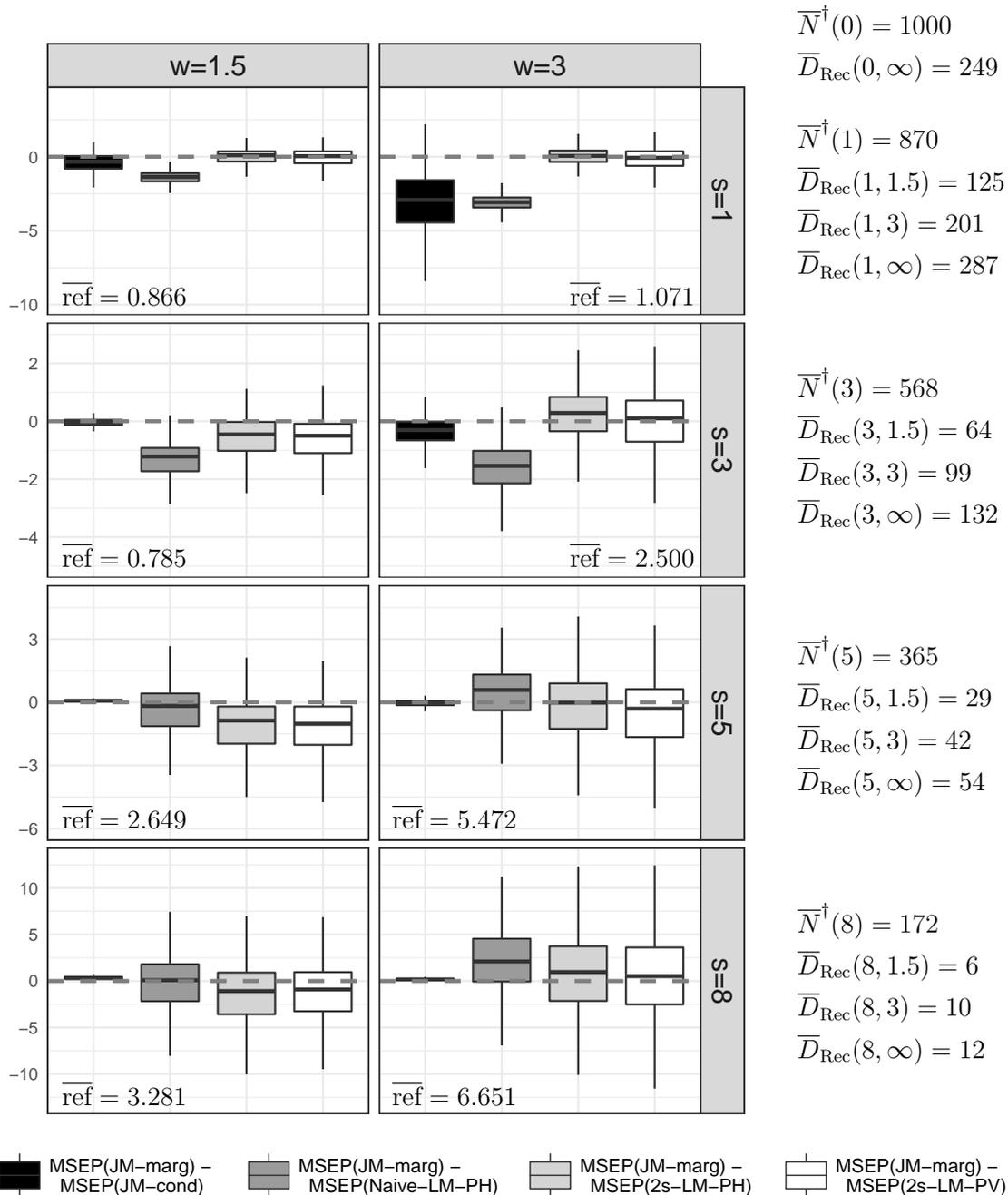


Figure 5.8 – Boxplots of the differences ($\times 1000$) of Mean Square Error of Prediction (MSEP) between the marginal estimator from the joint model (denoted JM-marg) and alternatives in the case of weak misspecification of the longitudinal marker trajectory (case 4.bis). Considered are the conditional estimator from the joint model (JM-cond), the estimators from cause-specific landmark models using a two-stage or naive approach (2s-LM-PH and Naive-LM-PH, respectively) and the two-stage pseudo value model (2s-LM-PV). The distributions are depicted over $R = 499, 497, 479, 439$ replicates for 4 landmark times $s = 1, 3, 5, 8$ respectively, with 2 considered horizons $w = 1.5$ and $w = 3$. $\overline{\text{ref}}$ denotes the mean MSEP ($\times 1000$) using the marginal estimator from the joint model for each (s, w) . $\overline{N}^\dagger(s)$ is the mean number of subjects at risk at s and $\overline{D}_{\text{Rec}}(s, w)$ is the mean number of recurrences occurred between s and $s + w$.

were comparable (Figure 5.11). In the case of a misspecification of the longitudinal marker trajectory, the two-stage landmark models discriminated noticeably better than the joint model when this misspecification was strong (Figure 5.12), but the discrimination abilities of the models were comparable in case of weak misspecification (Figure 5.13).

5.7.3 Simulation data generation

Model of data generation in the Cases 1 and 2

Data were generated according to the joint model:

$$\begin{cases} Y_i(t) &= m_i(t) + \epsilon_i(t) \\ &= (\beta_0 + \beta_{0,X}X_i + b_{i0}) + (\beta_1 + \beta_{1,X}X_i + b_{i1})t + \epsilon_i(t), \\ \lambda_i^k(t) &= \lambda_{k,0}(t) \exp \left\{ \gamma_k X_i + \eta_{1,k} m_i(t) + \eta_{2,k} \frac{\delta m_i(t)}{\delta t} \right\}, \end{cases}$$

where k is the cause of event (Recurrence ; Death), $b_i = (b_{i0}, b_{i1})^\top \sim \mathcal{N}(0, D)$ with $D = \begin{pmatrix} 0.582 & 0.032 \\ 0.032 & 0.061 \end{pmatrix}$ and $\epsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.268$. The other coefficients were $\beta_0 = -1.087$, $\beta_{0,X} = 0.465$, $\beta_1 = -0.066$, $\beta_{1,X} = 0.110$, $\gamma_{\text{Rec}} = 0.064$, $\gamma_{\text{Death}} = 0.208$, $\eta_{1,\text{Rec}} = 0.707$, $\eta_{1,\text{Death}} = 0.023$, $\eta_{2,\text{Rec}} = 2.140$, $\eta_{2,\text{Death}} = -0.462$, and $\log(\lambda_{k,0}(t))$ was a combination of cubic B-splines with the same knot vector $(3 \times 10^{-5}, 6.579, 15.874)^\top$ for all the events and the vector of spline coefficients $(-4.003, -4.107, -4.031, -8.806, -3.643)^\top$ for the recurrence and $(-5.401, -3.484, -3.810, -1.171, -1.263)^\top$ for the death. The times of measurements were $t_{ij} = \sum_{l=0}^j a_{il}$ with $a_{i0} = 0$, $a_{il} = 0.4 + u_{il}$ with $u_{il} \sim \mathcal{U}(-0.15, 0.15)$ for $l \in \{1, 2, 3\}$, $a_{il} = 0.6 + u_{il}$ with $u_{il} \sim \mathcal{U}(-0.2, 0.2)$ for $l \in \{4, 5, 6, 7\}$, $a_{il} = 1 + u_{il}$ with $u_{il} \sim \mathcal{U}(-0.35, 0.35)$ for $l \in \{8, 9, 10\}$, $a_{il} = 2 + u_{il}$ with $u_{il} \sim \mathcal{U}(-0.6, 0.6)$ for $l \in \{11, 12\}$, and $a_{il} = 3.5 + u_{il}$ with $u_{il} \sim \mathcal{U}(-1, 1)$ for $l \in \{13, 14, 15, 16\}$. The covariate X_i corresponded to the PSA level before treatment initiation in the motivating data, and was thus generated following a normal distribution with mean 2.041 and variance 0.503. The censoring time was generated from a uniform distribution on $[1, 15]$.

Model of data generation in the Case 3

Data were generated according to the joint model:

$$\begin{cases} Y_i(t) &= m_i(t) + \epsilon_i(t) \\ &= (\beta_0 + \beta_{0,X}X_i + b_{i0}) + (\beta_1 + \beta_{1,X}X_i + b_{i1})t + \epsilon_i(t), \\ \lambda_i^k(t) &= \lambda_{k,0}(t) \exp \left\{ \gamma_k X_i + \eta_{1,k} * \log(1+t)m_i(t) + \eta_{2,k} * \log(1+t) \frac{\delta m_i(t)}{\delta t} \right\}, \end{cases}$$

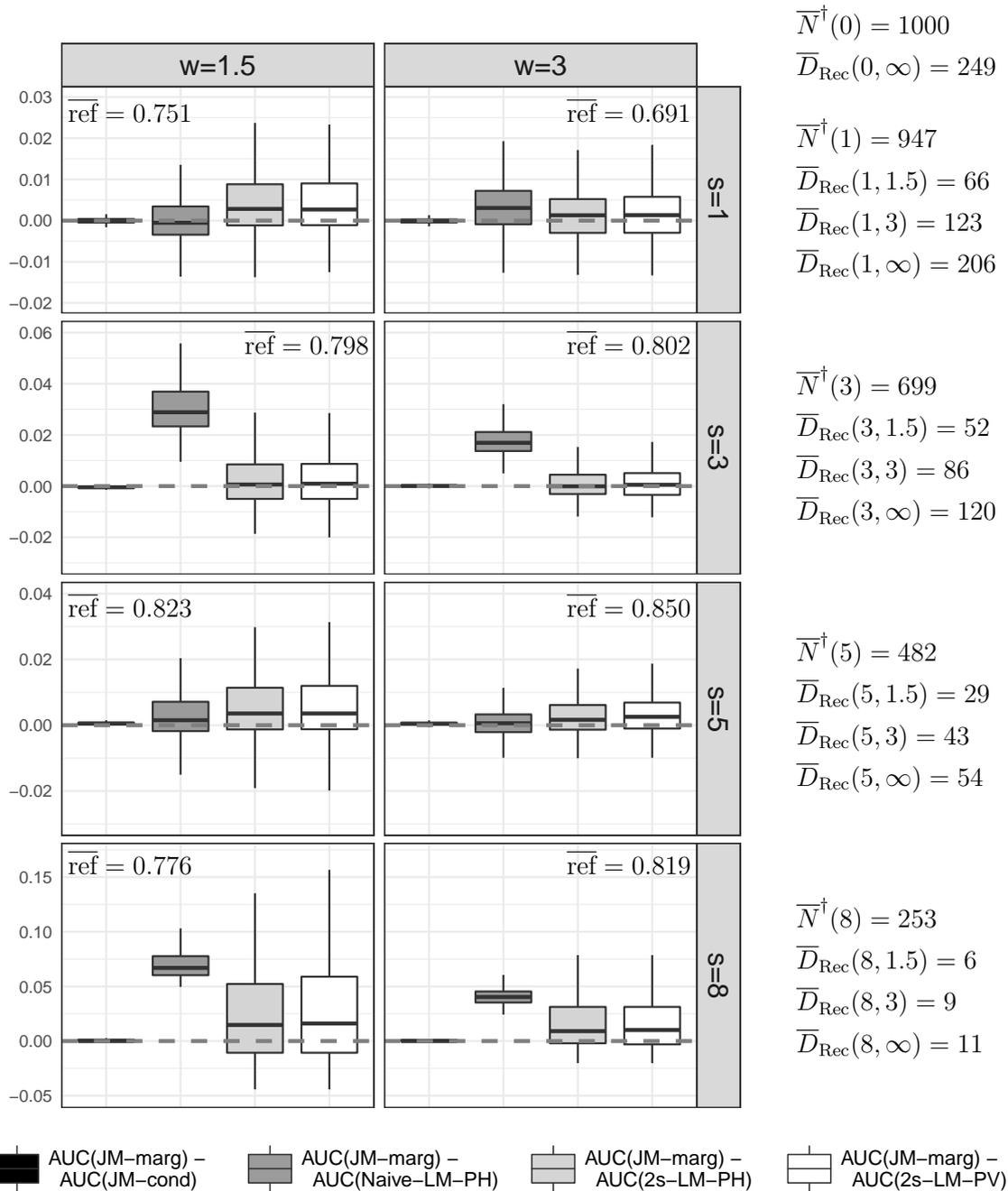


Figure 5.9 – Boxplots of the differences of Area Under ROC Curve (AUC) between the marginal estimator from the joint model (denoted JM-marg) and alternatives in the case of correct specification of the joint model (case 1). Considered are the conditional estimator from the joint model (JM-cond), the estimators from cause-specific landmark models using a two-stage or naive approach (2s-LM-PH and Naive-LM-PH, respectively) and the two-stage pseudo value model (2s-LM-PV). The distributions are depicted over $R = 499, 494, 486, 389$ replicates for 4 landmark times $s = 1, 3, 5, 8$ respectively, with 2 considered horizons $w = 1.5$ and $w = 3$. $\overline{\text{ref}}$ denotes the mean AUC using the marginal estimator from the joint model for each (s, w) . $\overline{N}^\dagger(s)$ is the mean number of subjects at risk at s and $\overline{D}_{\text{Rec}}(s, w)$ is the mean number of recurrences occurred between s and $s + w$.

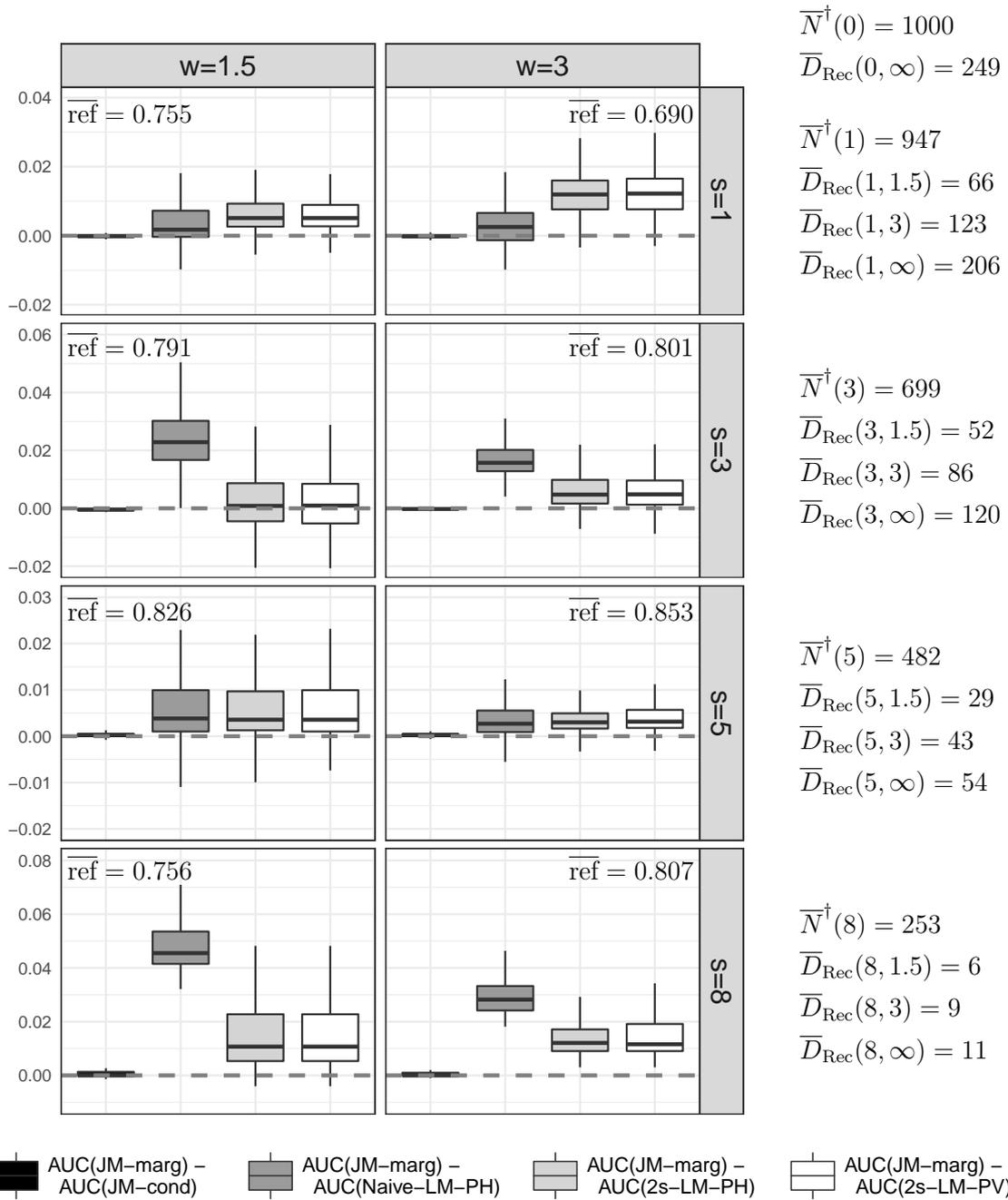


Figure 5.10 – Boxplots of the differences of Area Under ROC Curve (AUC) between the marginal estimator from the joint model (denoted JM-marg) and alternatives in the case of misspecification of the dependence function (case 2). Considered are the conditional estimator from the joint model (JM-cond), the estimators from cause-specific landmark models using a two-stage or naive approach (2s-LM-PH and Naive-LM-PH, respectively) and the two-stage pseudo value model (2s-LM-PV). The distributions are depicted over $R = 499, 498, 497, 428$ replicates for 4 landmark times $s = 1, 3, 5, 8$ respectively, with 2 considered horizons $w = 1.5$ and $w = 3$. $\bar{\text{ref}}$ denotes the mean AUC using the marginal estimator from the joint model for each (s, w) . $\bar{N}^\dagger(s)$ is the mean number of subjects at risk at s and $\bar{D}_{\text{Rec}}(s, w)$ is the mean number of recurrences occurred between s and $s + w$.

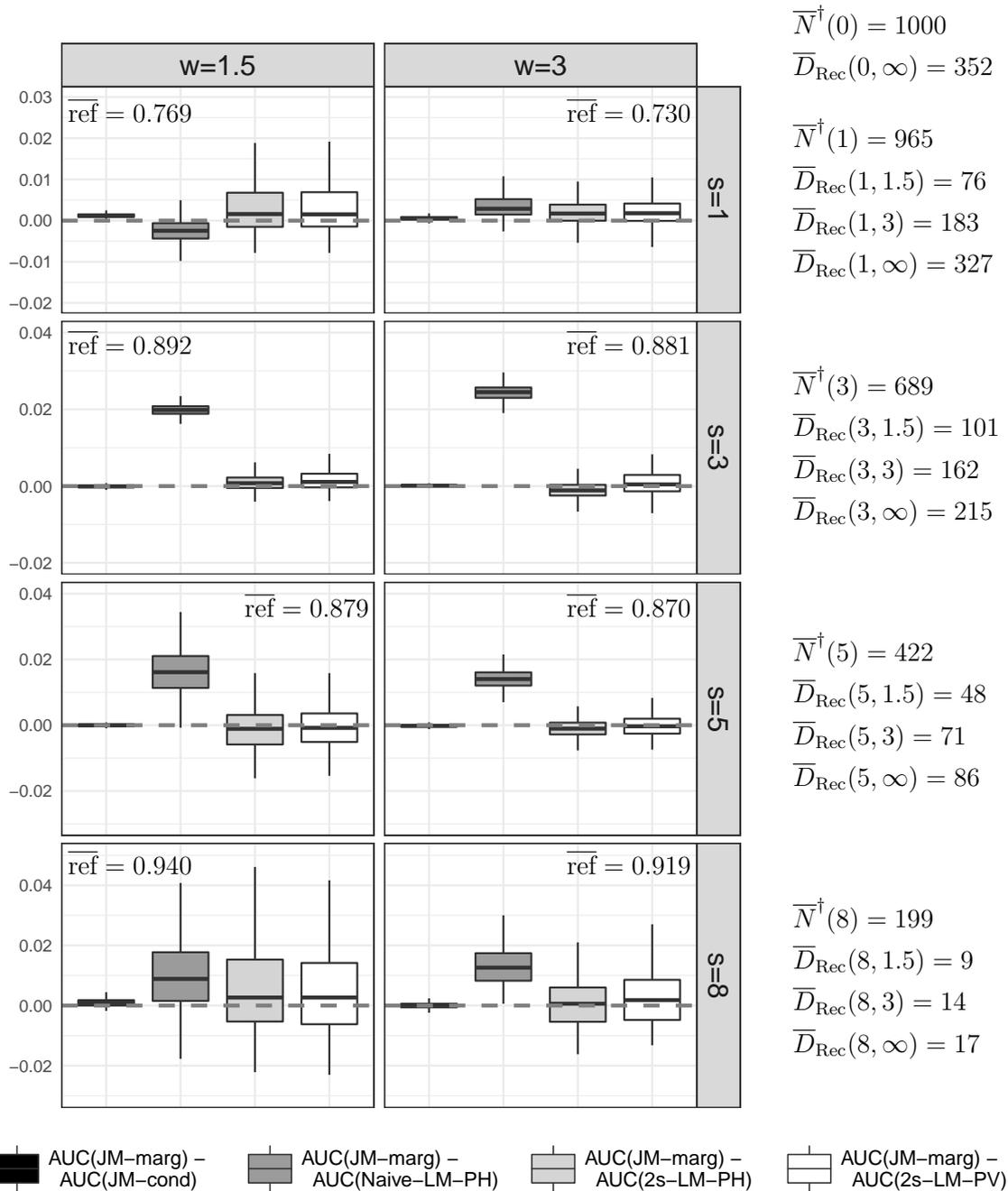


Figure 5.11 – Boxplots of the differences of Area Under ROC Curve (AUC) between the marginal estimator from the joint model (denoted JM-marg) and alternatives in the case of substantial violation of the PH assumption (case 3). Considered are the conditional estimator from the joint model (JM-cond), the estimators from cause-specific landmark models using a two-stage or naive approach (2s-LM-PH and Naive-LM-PH, respectively) and the two-stage pseudo value model (2s-LM-PV). The distributions are depicted over $R = 485, 326, 294, 188$ replicates for 4 landmark times $s = 1, 3, 5, 8$ respectively, with 2 considered horizons $w = 1.5$ and $w = 3$. $\overline{\text{ref}}$ denotes the mean AUC using the marginal estimator from the joint model for each (s, w) . $\overline{N}^\dagger(s)$ is the mean number of subjects at risk at s and $\overline{D}_{\text{Rec}}(s, w)$ is the mean number of recurrences occurred between s and $s + w$.

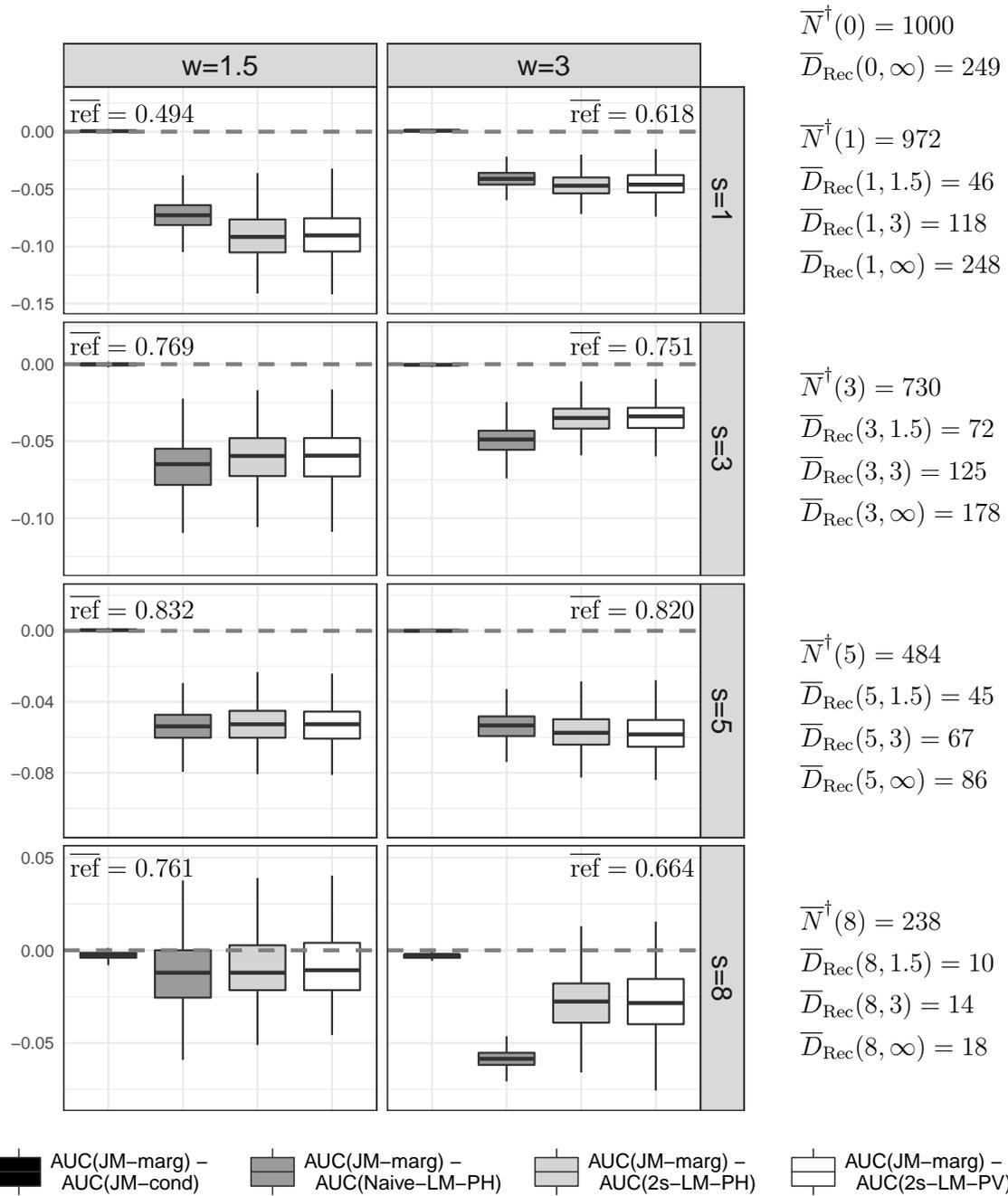


Figure 5.12 – Boxplots of the differences of Area Under ROC Curve (AUC) between the marginal estimator from the joint model (denoted JM-marg) and alternatives in the case of substantial misspecification of the longitudinal marker trajectory (case 4). Considered are the conditional estimator from the joint model (JM-cond), the estimators from cause-specific landmark models using a two-stage or naive approach (2s-LM-PH and Naive-LM-PH, respectively) and the two-stage pseudo value model (2s-LM-PV). The distributions are depicted over $R = 500, 496, 476, 357$ replicates for 4 landmark times $s = 1, 3, 5, 8$ respectively, with 2 considered horizons $w = 1.5$ and $w = 3$. \bar{ref} denotes the mean AUC using the marginal estimator from the joint model for each (s, w) . $\bar{N}^\dagger(s)$ is the mean number of subjects at risk at s and $\bar{D}_{Rec}(s, w)$ is the mean number of recurrences occurred between s and $s + w$.

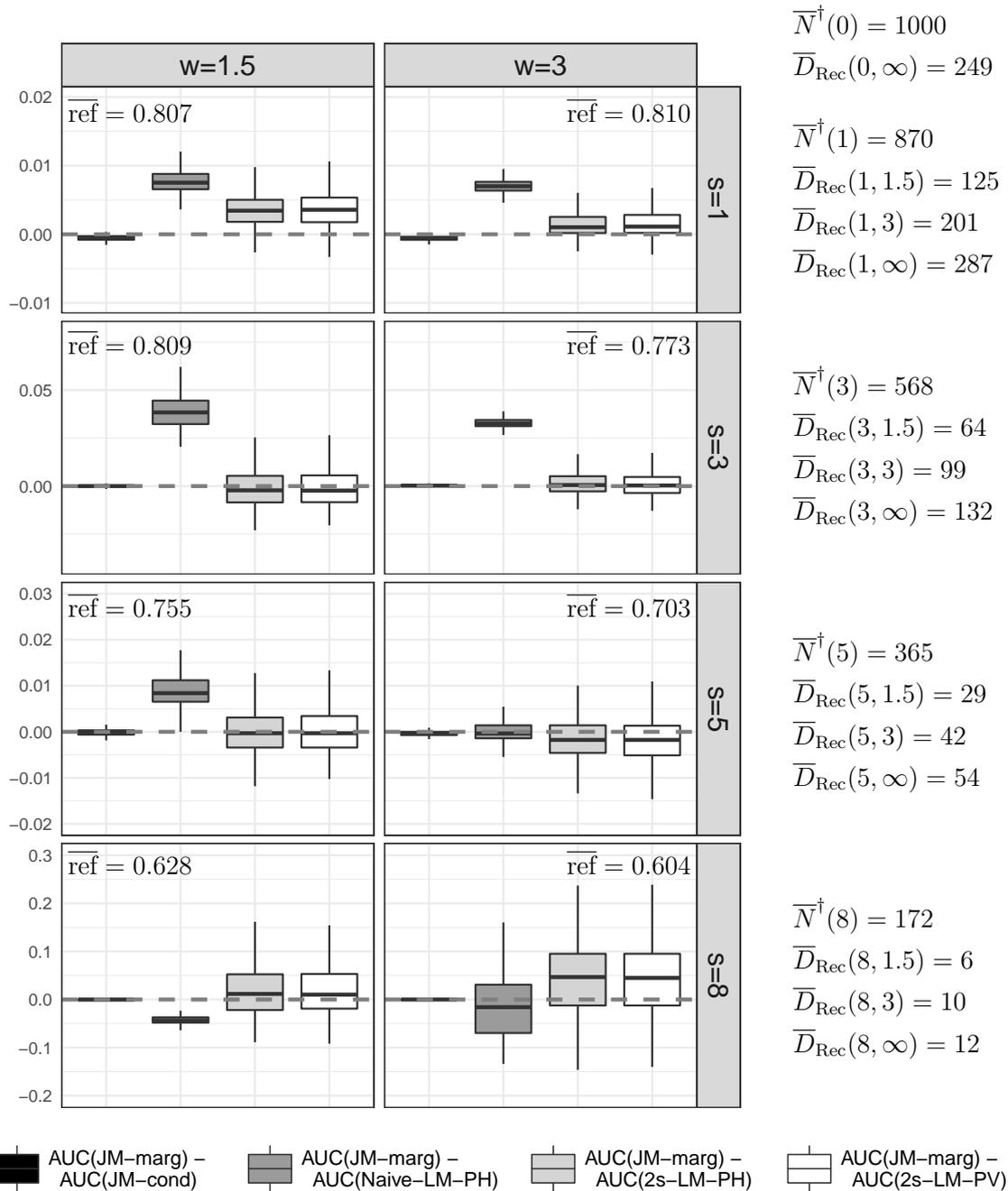


Figure 5.13 – Boxplots of the differences of Area Under ROC Curve (AUC) between the marginal estimator from the joint model (denoted JM-marg) and alternatives in the case of weak misspecification of the longitudinal marker trajectory (case 4.bis). Considered are the conditional estimator from the joint model (JM-cond), the estimators from cause-specific landmark models using a two-stage or naive approach (2s-LM-PH and Naive-LM-PH, respectively) and the two-stage pseudo value model (2s-LM-PV). The distributions are depicted over $R = 499, 497, 479, 439$ replicates for 4 landmark times $s = 1, 3, 5, 8$ respectively, with 2 considered horizons $w = 1.5$ and $w = 3$. $\overline{\text{ref}}$ denotes the mean AUC using the marginal estimator from the joint model for each (s, w) . $\overline{N}^\dagger(s)$ is the mean number of subjects at risk at s and $\overline{D}_{\text{Rec}}(s, w)$ is the mean number of recurrences occurred between s and $s + w$.

with the same covariates, random effects and parameters than in the Cases 1 and 2.

Model of data generation in the Case 4

Data were generated according to the joint model:

$$\left\{ \begin{array}{l} Y_i(t) = m_i(t) + \epsilon_i(t) \\ \quad = (\beta_0 + \beta_{0,X}X_i + b_{i0}) + (\beta_1 + \beta_{1,X}X_i + b_{i1})((1+t)^{-1.2} - 1) + \\ \quad \quad (\beta_2 + \beta_{2,X}X_i + b_{i2})t + \epsilon_i(t), \\ \lambda_i^k(t) = \lambda_{k,0}(t) \exp \left\{ \gamma_k X_i + \eta_{1,k} m_i(t) + \eta_{2,k} \frac{\delta m_i(t)}{\delta t} \right\}, \end{array} \right.$$

where k is the cause of event (Recurrence ; Death), $b_i = (b_{i0}, b_{i1}, b_{i2})^\top \sim \mathcal{N}(0, D)$ with $D =$

$\begin{pmatrix} 0.363 & 0.011 & 0.345 \\ 0.011 & 0.172 & 0.310 \\ 0.345 & 0.310 & 1.746 \end{pmatrix}$ and $\epsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.273$. The other coefficients were $\beta_0 = -0.255, \beta_{0,X} = 0.799, \beta_1 = 0.949, \beta_{1,X} = 0.904, \beta_2 = -0.088, \beta_{2,X} = 0.207, \gamma_{\text{Rec}} = 0.064, \gamma_{\text{Death}} = 0.208, \eta_{1,\text{Rec}} = 0.707, \eta_{1,\text{Death}} = 0.023, \eta_{2,\text{Rec}} = 2.140, \eta_{2,\text{Death}} = -0.462$, and $\log(\lambda_{k,0}(t))$ was a combination of cubic B-splines with the same knot vector $(3 \times 10^{-5}, 6.579, 15.874)^\top$ for all the events and the vector of spline coefficients $(-4.003, -4.107, -4.031, -8.806, -3.643)^\top$ for the recurrence and $(-5.401, -3.484, -3.810, -1.171, -1.263)^\top$ for the death. The times of measurements were $t_{ij} = \sum_{l=0}^j a_{il}$ with $a_{i0} = 0, a_{il} = 0.4 + u_{il}$ with $u_{il} \sim \mathcal{U}(-0.15, 0.15)$ for $l \in \{1, 2, 3\}$, $a_{il} = 0.6 + u_{il}$ with $u_{il} \sim \mathcal{U}(-0.2, 0.2)$ for $l \in \{4, 5, 6, 7\}$, $a_{il} = 1 + u_{il}$ with $u_{il} \sim \mathcal{U}(-0.35, 0.35)$ for $l \in \{8, 9, 10\}$, $a_{il} = 2 + u_{il}$ with $u_{il} \sim \mathcal{U}(-0.6, 0.6)$ for $l \in \{11, 12\}$, and $a_{il} = 3.5 + u_{il}$ with $u_{il} \sim \mathcal{U}(-1, 1)$ for $l \in \{13, 14, 15, 16\}$. The covariate X_i was generated following a normal distribution with mean 2.041 and variance 0.503. The censoring time was generated from a uniform distribution on $[1, 15]$.

Model of data generation in the Case 4.bis

Data were generated according to the joint model:

$$\left\{ \begin{array}{l} Y_i(t) = m_i(t) + \epsilon_i(t) \\ \quad = (\beta_0 + \beta_{0,X}X_i + b_{i0}) + \sum_{l=1}^3 (\beta_l + \beta_{l,X}X_i + b_{il}) B_l(t, 3) + \epsilon_i(t), \\ \lambda_i^k(t) = \lambda_{k,0}(t) \exp \left\{ \gamma_k X_i + \eta_{1,k} m_i(t) + \eta_{2,k} \frac{\delta m_i(t)}{\delta t} \right\}, \end{array} \right.$$

where k is the cause of event (Recurrence ; Death), $b_i = (b_{i0}, b_{i1}, b_{i2}, b_{i3})^\top \sim \mathcal{N}(0, D)$ with

$$D = \begin{pmatrix} 0.536 & 0.084 & 0.012 & -0.275 \\ 0.084 & 4.487 & 4.186 & 2.005 \\ 0.012 & 4.186 & 6.442 & 3.966 \\ -0.275 & 2.005 & 3.966 & 3.968 \end{pmatrix} \text{ and } \epsilon_i(t) \sim \mathcal{N}(0, \sigma^2) \text{ with } \sigma = 0.223. B_l(t, 3) \text{ is}$$

the l^{th} row of the cubic B-spline basis matrix defined with two internal knots placed at $t = 2$ and $t = 7$ and boundary knots at $t = 0$ and $t = 15$. The other coefficients were $\beta_0 = -1.003, \beta_{0,X} = 0.452, \beta_1 = -0.337, \beta_{1,X} = 0.790, \beta_2 = -0.687, \beta_{2,X} = 1.201, \beta_3 = -0.033, \beta_{3,X} = 0.692, \gamma_{\text{Rec}} = 0.500, \gamma_{\text{Death}} = 0.150, \eta_{1,\text{Rec}} = 0.918, \eta_{1,\text{Death}} = 0.068, \eta_{2,\text{Rec}} = 1.566, \eta_{2,\text{Death}} = -0.761$ and, $\log(\lambda_{k,0}(t))$ was a combination of cubic B-splines with the same knot vector $(3 \times 10^{-5}, 15.874)^\top$ for all the events and the vector of spline coefficients $(-3.931, -3.827, -7.564, -4.569)^\top$ for the recurrence and $(-4.549, -2.386, -2.585, -0.209)^\top$ for the death. The times of measurements were $t_{ij} = \sum_{l=0}^j a_{il}$ with $a_{i0} = 0, a_{il} = 0.4 + u_{il}$ with $u_{il} \sim \mathcal{U}(-0.15, 0.15)$ for $l \in \{1, 2, 3\}, a_{il} = 0.6 + u_{il}$ with $u_{il} \sim \mathcal{U}(-0.2, 0.2)$ for $l \in \{4, 5, 6, 7\}, a_{il} = 1 + u_{il}$ with $u_{il} \sim \mathcal{U}(-0.35, 0.35)$ for $l \in \{8, 9, 10\}, a_{il} = 2 + u_{il}$ with $u_{il} \sim \mathcal{U}(-0.6, 0.6)$ for $l \in \{11, 12\},$ and $a_{il} = 3.5 + u_{il}$ with $u_{il} \sim \mathcal{U}(-1, 1)$ for $l \in \{13, 14, 15, 16\}$. The covariate X_i was generated following a normal distribution with mean 2.041 and variance 0.503. The censoring time was generated from a uniform distribution on $[1, 15]$.

5.7.4 Computational time of the procedures

The several prediction models presented in this manuscript notably differ by the considered observed information and the complexity of the computational procedures. Table 5.1 compares the computational times of these procedures to compute individual dynamic predictions for a randomly chosen subject, based on a randomly selected replication. This example was taken from the case 1 of the manuscript, which notably considered two competing events (Recurrence, Death), a linear evolution of the marker with two individual random effects, and 1000 subjects included in the learning sample at baseline. For the joint model, 9 quadrature points were used according to the pseudo-adaptive Gauss-Hermite (GH) technique in the learning step, whereas 9 quadrature points using the adaptive GH technique were used in the validation step. Based on the techniques defined in the manuscript, $L = 1000$ samples of parameters were used to compute the confidence intervals using each approaches. The computing times were obtained on a single CPU of a PC with an Intel i7 processor, 3.40 GHz and 32Go of RAM memory. The R code used for this example can be found at <https://github.com/LoicFerrer/>.

Table 5.1 – Computation time (in seconds) for each step of the procedures to compute individual cumulative incidences of two competing events in the case of correct specification of the joint model (case 1). Considered are the marginal and conditional estimators from the joint model (denoted JM-marg and JM-cond, respectively), the estimators from cause-specific landmark models using a two-stage or naive approach (2s-LM-PH and Naive-LM-PH, respectively) and the two-stage pseudo-value model (2s-LM-PV). Based on a randomly selected replication, the predictions were computed for a randomly chosen subject \star for 2 landmark times $s = 1$ and $s = 5$ and one considered horizon $w = 3$.

	Estimation of the model		Computation of $\{\hat{\pi}_{\star}^k(s, w; \hat{\theta})\}_k$		Computation of the confidence intervals	
	$(s, w) = (1, 3)$	$(5, 3)$	$(1, 3)$	$(5, 3)$	$(1, 3)$	$(5, 3)$
JM-marg	461.329		1.308	1.293	1253.320	1271.876
JM-cond	461.329		0.541	0.525	497.445	523.508
Naive-LM-PH	0.009	0.007	0.018	0.014	172.033	54.724
2s-LM-PH	0.873	0.118	0.022	0.018	538.849	261.779
2s-LM-PV	1.275	0.187	0.004	0.004	1.397	1.428

As expected, the estimation of the joint model was considerably longer than the estimation of the landmark models, notably because the joint likelihood involved an integral over the two individual random effects. Nevertheless the joint model required only one model estimation to predict the individual cumulative incidences at any pair of landmark and horizon times (s, w) , whereas the landmark model based on pseudo-values needed one estimation for each s , and one estimation for each (s, w) was required for the landmark cause-specific proportional hazards model. For the same reasons, the number of subjects at risk and then the computation times of the estimation of the landmark models decreased when the landmark time s increased. Note that the estimation of the naive landmark CS PH model was very fast because the model was only adjusted using the last observation carried forward (LOCF) method.

The computation of the individual cumulative incidences of events and their associated confidence intervals was much longer using the marginal estimator from the joint model than others due to the integral over the individual random effects which was involved in the estimator's definition. Indeed, the conditional estimator from the joint model that approximates the integral by considering the modes of the posterior conditional distributions of the random effects was 2.5 times faster to compute than the marginal

estimator. For the landmark approaches, the computation of the estimator from the landmark pseudo-value model was faster than the computation of the estimator from the cause-specific landmark model because the former directly modelled the quantity of interest and then its expression did not include integral over time. The confidence intervals were substantially faster to compute using the landmark pseudo-value model than others because in addition to the direct modelling, the technique of the former only involved bootstrapped parameters whereas in the landmark CS PH models a perturbation resampling technique was also applied.

5.7.5 Example of R code

The individual dynamic predictions can be computed in clinical practice using R codes. Please see the detailed examples available at <https://github.com/LoicFerrer/>. The functions which implement the estimators defined in the manuscript (the marginal and conditional estimators from the joint model, the naive and two-stage estimators from the cause-specific proportional hazards landmark model and the naive and two-stage estimators from the landmark model based on pseudo-values) and their associated confidence intervals are also included. Simulated data are provided.

Here is given the script used in Section 5.4.1 of the main manuscript to compute the individual cumulative incidences of event for two landmark times $s = 1$, $s = 5$ and a given horizon $w = 3$, in subjects at risk at s . Data were retrieved from a randomly selected replication. Two types of event were considered (Recurrence, Death), the longitudinal evolution of the marker was linear, the same covariate, called X in the code below, impacted the longitudinal and survival processes, and the dependency between these two processes was explained through the true current level and the true current slope of the marker.

```
# Import two lists which contain the learning data and the validation data,
#   at https://github.com/LoicFerrer/
load("data.RData")
ls()
# [1] "data_learn" "data_valid"

str(data_learn)
# List of 2
# $ surv:'data.frame': 1000 obs. of  6 variables:
#   (...)
# $ long:'data.frame': 8319 obs. of  4 variables:
#   (...)
str(data_valid)
```

INDIVIDUAL DYNAMIC PREDICTIONS

```
# List of 3
# $ surv :'data.frame': 500 obs. of 6 variables:
# (... )
# $ long :'data.frame': 4886 obs. of 4 variables:
# (... )
# $ trueT: num [1:500] 11.04 7.67 12.82 ..

head(data_learn$surv, 3)
#   id      X time_of_Death Death time_of_Rec Rec
# 1  1 2.227805      2.106636      0      2.106636  0
# 2  2 1.452869      3.501039      0      3.501039  0
# 3  3 1.004715      2.991233      0      2.991233  1

head(data_learn$long, 3)
#   id      Y      times      X
# 1  1 0.6411969 0.0000000 2.227805
# 2  1 0.8455285 0.3296526 2.227805
# 3  1 1.2596015 0.6912898 2.227805

#####
#### Using the joint model ####
#####

# Load the 'JM' package to fit joint models with shared random effects
library(JM)
# Source the required functions to compute individual cumulative incidences of event.
# The file notably includes the functions 'JMCR.crLong' and 'survfitJMCR'
source("survfitJMCR.R")

# Adaptation of the learning survival database to the competing risks framework
data_learn$surv$Event <- "Alive"
data_learn$surv$Event[data_learn$surv$Death == 1] <- "Death"
data_learn$surv$Event[data_learn$surv$Rec == 1] <- "Rec"
data_learn$surv$time_of_Event <- pmin(data_learn$surv$time_of_Rec,
                                     data_learn$surv$time_of_Death)
data_learn$CR <- JMCR.crLong(data = data_learn$surv,
                             statusVar = "Event",
                             censLevel = "Alive")

head(data_learn$CR, 6)
#   id      X time_of_Death Death time_of_Rec Rec Event time_of_Event strata status2
# 1  1 2.227805      2.106636      0      2.106636  0 Alive      2.106636      Rec      0
# 1.1 1 2.227805      2.106636      0      2.106636  0 Alive      2.106636      Death     0
# 2  2 1.452869      3.501039      0      3.501039  0 Alive      3.501039      Rec      0
# 2.1 2 1.452869      3.501039      0      3.501039  0 Alive      3.501039      Death     0
# 3  3 1.004715      2.991233      0      2.991233  1 Rec        2.991233      Rec      1
# 3.1 3 1.004715      2.991233      0      2.991233  1 Rec        2.991233      Death     0

# Adaptation of the validation survival database to the competing risks framework
data_valid$surv$Event <- "Alive"
data_valid$surv$Event[data_valid$surv$Death == 1] <- "Death"
data_valid$surv$Event[data_valid$surv$Rec == 1] <- "Rec"
data_valid$surv$time_of_Event <- pmin(data_valid$surv$time_of_Rec,
                                     data_valid$surv$time_of_Death)
# WARNING: use the JMCR.crLong function by specifying the 'levels' argument as follows.
```

INDIVIDUAL DYNAMIC PREDICTIONS

```
# The 'strata' argument in the survival validation database has imperatively
# to be organized in the same order as its levels.
data_valid$CR <- JMCR.crLong(data = data_valid$surv,
                           statusVar = "Event",
                           censLevel = "Alive",
                           levels = levels(data_learn$CR$strata)) ## Warning

head(data_valid$CR, 6)
#      id      X time_of_Death Death time_of_Rec Rec Event time_of_Event strata status2
# 1     1 1.476506     5.714061     0     5.714061  0 Alive     5.714061  Death     0
# 1.1   1 1.476506     5.714061     0     5.714061  0 Alive     5.714061   Rec     0
# 2     2 1.797706     7.673092     0     7.673092  1  Rec     7.673092  Death     0
# 2.1   2 1.797706     7.673092     0     7.673092  1  Rec     7.673092   Rec     1
# 3     3 3.262949     9.419707     0     9.419707  0 Alive     9.419707  Death     0
# 3.1   3 3.262949     9.419707     0     9.419707  0 Alive     9.419707   Rec     0

## Learning step

# Estimation of the longitudinal sub-model (independent of the survival sub-model)
lmeFit <- lme(fixed = Y ~ times * X,
             data = data_learn$long,
             random = ~ times | id)
# Estimation of the survival sub-model (independent of the longitudinal sub-model)
coxFit <- coxph(formula = Surv(time_of_Event, status2) ~ X:strata(strata) + strata(strata),
               data = data_learn$CR,
               model = TRUE, x = TRUE)
# Definition of the marker's slope included in the joint model
dForm <- list(fixed = ~ 1 + X,
             indFixed = c(2, 4),
             random = ~ 1,
             indRandom = 2)
# Estimation of the joint model
jointFit <- jointModel(lmeFit,
                     coxFit,
                     timeVar = "times",
                     parameterization = "both",
                     method = "spline-PH-aGH",
                     interFact = list(value = ~ strata(strata) - 1,
                                       slope = ~ strata(strata) - 1,
                                       data = data_learn$CR),
                     derivForm = dForm,
                     control = list(GHk = 9, lng.in.kn = 1),
                     CompRisk = TRUE,
                     verbose = TRUE)

## Computation of the predicted individual cumulative incidences of events
## with 95% confidence intervals based on 1000 Monte Carlo samples

# ID at risk at the landmark times s=1 and s=5
Ri1 <- data_valid$surv$id[which(data_valid$trueT > 1)]
Ri5 <- data_valid$surv$id[which(data_valid$trueT > 5)]

# Longitudinal database and survival database with subjects at risk at s=1
data_valid.s1 <- NULL
```

INDIVIDUAL DYNAMIC PREDICTIONS

```

data_valid.s1$long <- data_valid$long[data_valid$long$times <= 1 &
                                data_valid$long$id %in% Ri1, ]
data_valid.s1$CR <- data_valid$CR[data_valid$CR$id %in% Ri1, ]
# Longitudinal database and survival database with subjects at risk at s=5
data_valid.s5 <- NULL
data_valid.s5$long <- data_valid$long[data_valid$long$times <= 5 &
                                data_valid$long$id %in% Ri5, ]
data_valid.s5$CR <- data_valid$CR[data_valid$CR$id %in% Ri5, ]

## Marginal estimator

# s=1, w=3
P1_3.JMmarg <- survfitJMCR(object = jointFit,
                           newData.long = data_valid.s1$long,
                           newData.surv = data_valid.s1$CR,
                           idVar = "id",
                           formT = ~ X:strata(strata),
                           tLM = rep(1, length(Ri1)),
                           thor = 3,
                           estimator = "marg",
                           simulate = T,
                           M = 1000,
                           CI.levels = c(0.025, 0.975))

P1_3.JMmarg
# Estimator definition: marginal to the individual random effects
#
#
# Predicted individual cumulative incidences of events
#
# $Death
#      tLM thor      Value
# ID 1  1    3 0.07514737
# ID 2  1    3 0.07108905
# ID 3  1    3 0.07597429
# (...)
#
# $Rec
#      tLM thor      Value
# ID 1  1    3 0.1301004
# ID 2  1    3 0.1050290
# ID 3  1    3 0.1726383
# (...)
#
# Predicted individual cumulative incidences of events
# based on 1000 Monte Carlo samples
#
# $Death
#      tLM thor      Mean      Median Lower (2.5%) Upper (97.5%)
# ID 1  1    3 0.07655855 0.07595813 0.05954908 0.09866198
# ID 2  1    3 0.07224854 0.07149081 0.05704207 0.09044535
# ID 3  1    3 0.07769421 0.07656262 0.05441197 0.10801782
# (...)
#

```

INDIVIDUAL DYNAMIC PREDICTIONS

```
# $Rec
#      tLM thor      Mean      Median Lower (2.5%) Upper (97.5%)
# ID 1  1    3 0.1313873 0.1307075  0.10596253  0.1623071
# ID 2  1    3 0.1062217 0.1054323  0.08620265  0.1299271
# ID 3  1    3 0.1741073 0.1724545  0.12964078  0.2298147
# (...)
```

```
# s=5, w=3
P5_3.JMmarg <- survfitJMCR(object = jointFit,
                           newData.long = data_valid.s5$long,
                           newData.surv = data_valid.s5$CR,
                           idVar = "id",
                           formT = ~ X:strata(strata),
                           tLM = rep(5, length(Ri5)),
                           thor = 3,
                           estimator = "marg",
                           simulate = T,
                           M = 1000,
                           CI.levels = c(0.025, 0.975))
```

```
P5_3.JMmarg
# Estimator definition: marginal to the individual random effects
#
#
# Predicted individual cumulative incidences of events
#
# $Death
#      tLM thor      Value
# ID 1  5    3 0.1721675
# ID 2  5    3 0.1259304
# ID 3  5    3 0.1871697
# (...)
```

```
#
# $Rec
#      tLM thor      Value
# ID 1  5    3 0.01770684
# ID 2  5    3 0.24186527
# ID 3  5    3 0.03426441
# (...)
```

```
#
#
# Predicted individual cumulative incidences of events
# based on 1000 Monte Carlo samples
#
# $Death
#      tLM thor      Mean      Median Lower (2.5%) Upper (97.5%)
# ID 1  5    3 0.1729590 0.1724336  0.13513284  0.2145668
# ID 2  5    3 0.1266225 0.1250114  0.09317579  0.1667234
# ID 3  5    3 0.1885321 0.1861479  0.13515246  0.2491214
# (...)
```

```
#
# $Rec
#      tLM thor      Mean      Median Lower (2.5%) Upper (97.5%)
# ID 1  5    3 0.01817897 0.01785144  0.01166856  0.02671285
# ID 2  5    3 0.24515436 0.24351905  0.19302590  0.30336436
# ID 3  5    3 0.03545071 0.03463930  0.02166277  0.05431355
```

INDIVIDUAL DYNAMIC PREDICTIONS

```

# (...)

## Conditional estimator

# s=1, w=3
P1_3.JMcond <- survfitJMCR(object = jointFit,
                           newData.long = data_valid.s1$long,
                           newData.surv = data_valid.s1$CR,
                           idVar = "id",
                           formT = ~ X:strata(strata),
                           tLM = rep(1, length(Ri1)),
                           thor = 3,
                           estimator = "cond",
                           simulate = T,
                           M = 1000,
                           CI.levels = c(0.025, 0.975))

P1_3.JMcond
# Estimator definition: conditional to the MAP of the individual random effects
#
#
# Predicted individual cumulative incidences of events
#
# $Death
#      tLM thor      Value
# ID 1  1    3 0.07798428
# ID 2  1    3 0.07307276
# ID 3  1    3 0.07953390
# (...)
#
# $Rec
#      tLM thor      Value
# ID 1  1    3 0.1828431
# ID 2  1    3 0.1413761
# ID 3  1    3 0.2416013
# (...)
#
# Predicted individual cumulative incidences of events
# based on 1000 Monte Carlo samples
#
# $Death
#      tLM thor      Mean      Median Lower (2.5%) Upper (97.5%)
# ID 1  1    3 0.07932089 0.07874783  0.06098413  0.10237902
# ID 2  1    3 0.07400522 0.07340157  0.05831713  0.09174712
# ID 3  1    3 0.08134418 0.07953099  0.05647367  0.11628135
# (...)
#
# $Rec
#      tLM thor      Mean      Median Lower (2.5%) Upper (97.5%)
# ID 1  1    3 0.1845178 0.1838128  0.1477782  0.2278376
# ID 2  1    3 0.1424508 0.1410770  0.1094119  0.1815720
# ID 3  1    3 0.2432991 0.2410676  0.1678992  0.3242795
# (...)

```

INDIVIDUAL DYNAMIC PREDICTIONS

```
# s=5, w=3
P5_3.JMcond <- survfitJMCR(object = jointFit,
                           newData.long = data_valid.s5$long,
                           newData.surv = data_valid.s5$CR,
                           idVar = "id",
                           formT = ~ X:strata(strata),
                           tLM = rep(5, length(Ri5)),
                           thor = 3,
                           estimator = "cond",
                           simulate = T,
                           M = 1000,
                           CI.levels = c(0.025, 0.975))

P5_3.JMcond
# Estimator definition: conditional to the MAP of the individual random effects
#
#
# Predicted individual cumulative incidences of events
#
# $Death
#      tLM thor      Value
# ID 1  5    3 0.1723353
# ID 2  5    3 0.1273586
# ID 3  5    3 0.1875795
# (...)
#
# $Rec
#      tLM thor      Value
# ID 1  5    3 0.01851501
# ID 2  5    3 0.24754646
# ID 3  5    3 0.03640618
# (...)
#
# Predicted individual cumulative incidences of events
# based on 1000 Monte Carlo samples
#
# $Death
#      tLM thor      Mean      Median Lower (2.5%) Upper (97.5%)
# ID 1  5    3 0.1735219 0.1719451 0.13600771 0.2207591
# ID 2  5    3 0.1280152 0.1269849 0.09377241 0.1677724
# ID 3  5    3 0.1883834 0.1868395 0.13490875 0.2547082
# (...)
#
# $Rec
#      tLM thor      Mean      Median Lower (2.5%) Upper (97.5%)
# ID 1  5    3 0.01902124 0.01876318 0.01242588 0.02763914
# ID 2  5    3 0.24939821 0.24869838 0.19645549 0.31078529
# ID 3  5    3 0.03744895 0.03642236 0.02331171 0.05648034
# (...)
```

INDIVIDUAL DYNAMIC PREDICTIONS

```
#####  
#### Using the two-stage landmark CS PH model ####  
#####  
  
# Load the 'nlme' and 'mstate' packages to fit the model  
library(nlme)  
library(mstate)  
# Source the required functions to compute individual cumulative incidences of event  
# The file notably includes the function 'CumInc'  
source("survfitLMCR.R")  
  
survfitLMCR <- function(tLM, thor, simulate = T, M = 1000, CI.levels = c(0.025, 0.975)) {  
  # tLM: landmark time  
  # thor: horizon window  
  tpred <- tLM + thor  
  
  ## Learning step ##  
  
  # ID at risk at the landmark time  
  Ri <- data_learn$surv[which(data_learn$surv$time_of_Rec > tLM), "id"]  
  nLM <- length(Ri)  
  # One considers all data collected before the landmark time point for subjects still at risk  
  LMLong <- data_learn$long[data_learn$long$times < tLM & data_learn$long$id %in% Ri, ]  
  LMsurv <- data_learn$surv[data_learn$surv$id %in% Ri, ]  
  
  # Estimation of the linear mixed model  
  lmeFit <- lme(fixed = Y ~ times*X,  
               data = LMLong,  
               random = ~ times| id)  
  
  # BLUPs and parameters  
  b <- ranef(lmeFit)  
  sigma <- lmeFit$sigma  
  D <- lapply(pdMatrix(lmeFit$modelStruct$reStruct), "*", sigma^2)[[1]]  
  betas <- fixef(lmeFit)  
  
  # Predicted level and slope of the marker at the landmark time  
  Data.s <- data.frame(id = unique(LMLong$id),  
                      times = tLM,  
                      X = LMLong[!duplicated(LMLong$id), "X"])  
  Xlong.s <- model.matrix( ~ times*X, Data.s)  
  Z.s <- model.matrix( ~ times, Data.s)  
  LMsurv$level <- as.vector(c(Xlong.s %*% betas) + rowSums(Z.s * b))  
  Xlong_deriv.s <- model.matrix( ~ X, Data.s)  
  Zderiv.s <- model.matrix( ~ 1, Data.s)  
  LMsurv$slope <- as.vector(c(Xlong_deriv.s %*% betas[c(2, 4)]) +  
                           rowSums(Zderiv.s * b[ , 2, drop = FALSE]))  
  
  # Administrative censoring at the end of the prediction window  
  LMsurv$tRecAC <- pmin(LMsurv$time_of_Rec, tpred)  
  LMsurv$RecAC <- LMsurv$Rec  
  LMsurv$RecAC[LMsurv$time_of_Rec > tpred] <- 0  
  LMsurv$tDeathAC <- pmin(LMsurv$time_of_Death, tpred)  
  LMsurv$DeathAC <- LMsurv$Death
```

INDIVIDUAL DYNAMIC PREDICTIONS

```
LMSurv$DeathAC[LMSurv$time_of_Death > tpred] <- 0

# Estimation of the Cox model (Recurrence)
coxFit1 <- coxph(Surv(tRecAC, RecAC) ~ X + level + slope, data = LMSurv)
# Estimation of the Cox model (Death)
coxFit2 <- coxph(Surv(tDeathAC, DeathAC) ~ X + level + slope, data = LMSurv)

## Computation of the predicted individual cumulative incidences of events

# Subjects at risk at tLM
Ri_pred <- data_valid$surv$id[which(data_valid$trueT > tLM)]
nLM_pred <- length(Ri_pred)

# Computation of the predicted individual risk scores
LMSurv_pred <- data_valid$surv[Ri_pred, ]
LMlong_pred <- data_valid$long[data_valid$long$times < tLM & data_valid$long$id %in% Ri_pred, ]
Xlong_pred <- split(data.frame(model.matrix( ~ times*X, LMlong_pred)),
  LMlong_pred$id)
Z_pred <- split(data.frame(model.matrix( ~ times, LMlong_pred)),
  LMlong_pred$id)
Y_pred <- split(LMlong_pred$Y, LMlong_pred$id)
V_pred <- lapply(Z_pred, function(x){
  as.matrix(x) %*% D %*% t(as.matrix(x)) + diag(sigma^2, nrow(as.matrix(x)))})
b_pred <- lapply(seq_len(length(Xlong_pred)),
  function(i, x, y, z, v) {
    D %*% t(as.matrix(z[[i]])) %*%
      solve(v[[i]]) %*% (y[[i]] - as.matrix(x[[i]]) %*% fixef(lmeFit))),
    x = Xlong_pred, y = Y_pred, z = Z_pred, v = V_pred)
b_pred <- matrix(unlist(b_pred), ncol = 2, , byrow = T)
Data_pred.s <- data.frame(id = unique(LMlong_pred$id),
  times = tLM,
  X = LMlong_pred[!duplicated(LMlong_pred$id), "X"])
Xlong_pred.s <- model.matrix( ~ times*X, Data_pred.s)
Z_pred.s <- model.matrix( ~ times, Data_pred.s)
LMSurv_pred$level <- as.vector(c(Xlong_pred.s %*% betas) + rowSums(Z_pred.s * b_pred))
Xlong_deriv_pred.s <- model.matrix( ~ X, Data_pred.s)
Zderiv_pred.s <- model.matrix( ~ 1, Data_pred.s)
LMSurv_pred$slope <- as.vector(c(Xlong_deriv_pred.s %*% betas[c(2, 4)]) +
  rowSums(Zderiv_pred.s * b_pred[ , 2, drop = FALSE]))
Xsurv_pred <- model.matrix( ~ 0 + X + level + slope, LMSurv_pred)
HR1 <- as.numeric(exp(Xsurv_pred %*% coxFit1$coef))
HR2 <- as.numeric(exp(Xsurv_pred %*% coxFit2$coef))

# Baseline hazards estimates
bh1 <- basehaz(coxFit1, centered = FALSE)
bh2 <- basehaz(coxFit2, centered = FALSE)

# Reasonably quick function that converts cause-specific hazards to cumulative
# incidence functions and extracts value at horizon
toci <- function(bh1, bh2, HR1, HR2, tpred)
{
  h1 <- bh1
  names(h1)[1] <- "Haz"
  h1$Haz <- h1$Haz * HR1
```

INDIVIDUAL DYNAMIC PREDICTIONS

```

h1$cause <- 1
h2 <- bh2
names(h2)[1] <- "Haz"
h2$Haz <- h2$Haz * HR2
h2$cause <- 2
Haz <- rbind(h1, h2)
CI <- CumInc(Haz)
idx <- sum(CI$time <= tpred)
return(CI[idx, ])
}

# Individual cumulative incidences
ci <- matrix(NA, nLM_pred, 2)
for (i in 1:nLM_pred) {
  ci[i, ] <- as.numeric(toci(bh1, bh2, HR1[i], HR2[i], tpred))[2:3]
}

if (simulate){
  ## Computation of the predicted 95% confidence intervals based on M Monte Carlo samples
  ci.MC <- replicate(M, matrix( , nrow = nLM_pred, ncol = 2), simplify = FALSE)

  # Time-to-event data
  t <- matrix(coxFit1$y, ncol = 2)[,1]
  Delta1 <- matrix(coxFit1$y, ncol = 2)[,2]
  Delta2 <- matrix(coxFit2$y, ncol = 2)[,2]

  # Parametric bootstrap
  aV <- lmeFit$apVar
  Pars <- attr(aV, "Pars")
  varFix <- lmeFit$varFix
  nbetas <- length(betas)
  nP <- length(Pars)
  mat <- matrix(0, nbetas + nP, nbetas + nP)
  mat[seq_len(nbetas), seq_len(nbetas)] <- varFix
  mat[nbetas + seq_len(nP), nbetas + seq_len(nP)] <- aV
  coef_coxFit1.MC <- mvrnorm(M,
                            mu = coxFit1$coef,
                            Sigma = coxFit1$var)
  coef_coxFit2.MC <- mvrnorm(M,
                            mu = coxFit2$coef,
                            Sigma = coxFit2$var)
  coef_long.MC <- mvrnorm(M, c(betas, Pars), Sigma = mat)

  for(l in seq_len(M)){
    betas.MC <- coef_long.MC[l, seq_len(nbetas)]
    sigma.MC <- exp(coef_long.MC[l, nbetas + nP])
    lmeSt <- lmeFit$modelStruct
    lmeSt$reStruct[[1]] <- pdNatural(lmeSt$reStruct[[1]])
    coef(lmeSt) <- coef_long.MC[l, -c(seq_len(nbetas), nbetas + nP)]
    Pars_D.MC <- coef(lmeSt, unconstrained = FALSE)
    D.MC <- matrix(c(Pars_D.MC[1], rep(Pars_D.MC[3], 2), Pars_D.MC[2]), 2, 2)
    diag(D.MC) <- diag(D.MC)^2
    D.MC[upper.tri(D.MC, diag = FALSE)] <-
      D.MC[lower.tri(D.MC, diag = FALSE)] <-
      Reduce('*', Pars_D.MC)
  }
}

```

INDIVIDUAL DYNAMIC PREDICTIONS

```

# Computation of the actualized marker's dynamics in the learning sample
Xlong <- split(data.frame(model.matrix( ~ times*X, LMLong)), LMLong$id)
Z <- split(data.frame(model.matrix( ~ times, LMLong)), LMLong$id)
Y <- split(LMLong$Y, LMLong$id)
V.MC <- lapply(Z, function(x){
  as.matrix(x) %*% D.MC %*% t(as.matrix(x)) + diag(sigma.MC^2, nrow(as.matrix(x)))})
b.MC <- lapply(seq_len(length(Xlong)),
  function(i, x, y, z, v) {
    D.MC %*% t(as.matrix(z[[i]])) %*%
      solve(v[[i]]) %*% (y[[i]] - as.matrix(x[[i]]) %*% betas.MC)},
  x = Xlong, y = Y, z = Z, v = V.MC)
b.MC <- matrix(unlist(b.MC), ncol = 2, , byrow = T)
LMsurv.MC <- LMsurv
LMsurv.MC$level.MC <- as.vector(c(Xlong.s %*% betas.MC) + rowSums(Z.s * b.MC))
LMsurv.MC$slope.MC <- as.vector(c(Xlong_deriv.s %*% betas.MC[c(2, 4)]) +
  rowSums(Zderiv.s * b.MC[ , 2, drop = FALSE]))
Xsurv.MC <- model.matrix( ~ 0 + X + level + slope, LMsurv.MC)

# Computation of the actualized individual risk scores in the validation sample
V_pred.MC <- lapply(Z_pred, function(x){
  as.matrix(x) %*% D.MC %*% t(as.matrix(x)) + diag(sigma.MC^2, nrow(as.matrix(x)))})
b_pred.MC <- lapply(seq_len(length(Xlong_pred)),
  function(i, x, y, z, v) {
    D.MC %*% t(as.matrix(z[[i]])) %*%
      solve(v[[i]]) %*% (y[[i]] - as.matrix(x[[i]]) %*% betas.MC)},
  x = Xlong_pred, y = Y_pred, z = Z_pred, v = V_pred.MC)
b_pred.MC <- matrix(unlist(b_pred.MC), ncol = 2, , byrow = T)
LMsurv_pred.MC <- LMsurv_pred
LMsurv_pred.MC$level_pred.MC <-
  as.vector(c(Xlong_pred.s %*% betas.MC) +
    rowSums(Z_pred.s * b_pred.MC))
LMsurv_pred.MC$slope_pred.MC <-
  as.vector(c(Xlong_deriv_pred.s %*% betas.MC[c(2, 4)]) +
    rowSums(Zderiv_pred.s * b_pred.MC[ , 2, drop = FALSE]))
Xsurv_pred.MC <- model.matrix( ~ 0 + X + level + slope, LMsurv_pred.MC)
HR1.MC <- as.numeric(exp(Xsurv_pred.MC %*% coef_coxFit1.MC[1,]))
HR2.MC <- as.numeric(exp(Xsurv_pred.MC %*% coef_coxFit2.MC[1,]))

# Perturbation-resampling
set.seed(1)
nu_event <- 4 * rbeta(nLM, 1/2, 3/2)
haz1.MCPR <- haz2.MCPR <- N1.PR <- N2.PR <- NULL
for(i in 1:nLM){
  haz1.MCPR[i] <- mean(nu_event * as.numeric(t >= t[i]) *
    exp(c(Xsurv.MC %*% coef_coxFit1.MC[1,])))
  haz2.MCPR[i] <- mean(nu_event * as.numeric(t >= t[i]) *
    exp(c(Xsurv.MC %*% coef_coxFit2.MC[1,])))
  N1.PR[i] <- mean(nu_event * as.numeric(t <= t[i]) * Delta1)
  N2.PR[i] <- mean(nu_event * as.numeric(t <= t[i]) * Delta2)
}
Haz1.MCPR <- data.frame(hazard = cumsum(diff(c(0,N1.PR[order(t)])) /
  haz1.MCPR[order(t)]),
  time = t[order(t)])
Haz2.MCPR <- data.frame(hazard = cumsum(diff(c(0,N2.PR[order(t)])) /

```

INDIVIDUAL DYNAMIC PREDICTIONS

```

                                haz2.MCPR[order(t)]),
                                time = t[order(t)])

# Cumulative incidences after parametric bootstrap and perturbation-resampling
for(i in 1:nLM_pred){
  ci.MC[[1]][i,] <- as.numeric(toci(Haz1.MCPR, Haz2.MCPR, HR1.MC[i], HR2.MC[i], tpred))[2:3]
}

  cat(paste("Monte Carlo sample: ", 1, "/", M, sep = ""), "\n")
}
res.MC <- replicate(2, matrix( , nrow = nLM_pred, ncol = 6), simplify = F)
for(k in 1:2) {
  for (i in seq_len(nLM_pred)) {
    res.MC[[k]][i,] <- c(tLM,
                        thor,
                        mean(sapply(ci.MC, function(x) x[,k])[i,]),
                        median(sapply(ci.MC, function(x) x[,k])[i,]),
                        quantile(sapply(ci.MC, function(x) x[,k])[i,], probs = CI.levels[1]),
                        quantile(sapply(ci.MC, function(x) x[,k])[i,], probs = CI.levels[2]))
  }
  colnames(res.MC[[k]]) <- c("tLM", "thor", "Mean", "Median",
                            paste("Lower", " (", CI.levels[1] * 100, "%)", sep = ""),
                            paste("Upper", " (", CI.levels[2] * 100, "%)", sep = ""))
  rownames(res.MC[[k]]) <- sapply(Ri_pred, function(x) paste("ID", x))
}
res <- replicate(2, matrix( , nrow = nLM_pred, ncol = 3), simplify = F)
for(k in 1:2) {
  for (i in seq_len(nLM_pred)) {
    res[[k]][i,] <- c(tLM, thor, ci[i,k])
  }
  colnames(res[[k]]) <- c("tLM", "thor", "Value")
  rownames(res[[k]]) <- sapply(Ri_pred, function(x) paste("ID", x))
}
names(res) <- names(res.MC) <- c("Rec", "Death")
}
result <- {
  if (simulate)
    list(res = res, res.MC = res.MC, simulate = simulate, M = M)
  else list(res = res, simulate = simulate, M = M)
}
rm(list = ".Random.seed", envir = globalenv())
class(result) <- "survfitLMCR"
return(result)
}

# s=1, w=3
P1_3.2sLMPH <- survfitLMCR(tLM = 1,
                          thor = 3,
                          simulate = T,
                          M = 1000,
                          CI.levels = c(0.025, 0.975))

P1_3.2sLMPH
# Predicted individual cumulative incidences of events
#

```

INDIVIDUAL DYNAMIC PREDICTIONS

```

# $Rec
#       tLM thor      Value
# ID 1   1   3 0.11472941
# ID 2   1   3 0.09965338
# ID 3   1   3 0.16704735
# (...)
#
# $Death
#       tLM thor      Value
# ID 1   1   3 0.07755851
# ID 2   1   3 0.08732394
# ID 3   1   3 0.08023811
# (...)
#
# Predicted individual cumulative incidences of events
# based on 1000 Monte Carlo samples
#
# $Rec
#       tLM thor      Mean      Median Lower (2.5%) Upper (97.5%)
# ID 1   1   3 0.11332436 0.1118789  0.08164853  0.1488563
# ID 2   1   3 0.09858466 0.0980073  0.07424215  0.1257947
# ID 3   1   3 0.16982380 0.1674375  0.10726483  0.2540990
# (...)
#
# $Death
#       tLM thor      Mean      Median Lower (2.5%) Upper (97.5%)
# ID 1   1   3 0.07734816 0.07608985  0.04895068  0.1098262
# ID 2   1   3 0.08598385 0.08609996  0.06279178  0.1119255
# ID 3   1   3 0.08069485 0.07685824  0.04198228  0.1380541
# (...)

# s=5, w=3
P5_3.2sLMPH <- survfitLMCR(tLM = 5,
                           thor = 3,
                           simulate = T,
                           M = 1000,
                           CI.levels = c(0.025, 0.975))

P5_3.2sLMPH
# Predicted individual cumulative incidences of events
#
# $Rec
#       tLM thor      Value
# ID 1   5   3 0.01891257
# ID 2   5   3 0.27924131
# ID 3   5   3 0.03760378
# (...)
#
# $Death
#       tLM thor      Value
# ID 1   5   3 0.2004973
# ID 2   5   3 0.1405373
# ID 3   5   3 0.1595905
# (...)
#

```

```

#
# Predicted individual cumulative incidences of events
# based on 1000 Monte Carlo samples
#
# $Rec
#      tLM thor      Mean      Median Lower (2.5%) Upper (97.5%)
# ID 1   5     3 0.02050010 0.01884418 0.007745837 0.04296857
# ID 2   5     3 0.27184453 0.26945235 0.175979514 0.36897579
# ID 3   5     3 0.03948784 0.03655589 0.015824254 0.08138946
# (...)
#
# $Death
#      tLM thor      Mean      Median Lower (2.5%) Upper (97.5%)
# ID 1   5     3 0.1997381 0.1993476 0.13303412 0.2732692
# ID 2   5     3 0.1409608 0.1388352 0.08532541 0.2102408
# ID 3   5     3 0.1598849 0.1558159 0.08413382 0.2657962
# (...)

```

5.7.6 References

- BLANCHE, PAUL, PROUST-LIMA, CÉCILE, LOUBÈRE, LUCIE, BERR, CLAUDINE, DARTIGUES, JEAN-FRANÇOIS AND JACQMIN-GADDA, HÉLÈNE. Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics* 71(1), 102–113, 2015. 147, 165
- FERRER, LOÏC, RONDEAU, VIRGINIE, DIGNAM, JAMES, PICKLES, TOM, JACQMIN-GADDA, HÉLÈNE AND PROUST-LIMA, CÉCILE. Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer. *Statistics in Medicine* 35(22), 3933–3948, 2013. 70, 106, 107, 109, 112, 114, 115, 116, 119, 122, 130, 139, 145, 163

6 Reflection on the joint modelling of ordinal longitudinal data and time-to-event data

In the literature, joint models have been mainly developed to study a Gaussian longitudinal marker and event history data. Yet many clinical applications concern the use of non-continuous ordinal marker with informative drop-out, such as the repeated measurements of Health Related Quality of Life questionnaires until the occurrence of a clinical event in cancer trials. Some joint models for ordinal outcomes and event history data have been proposed. However, no associated software has been developed so far to analyse this kind of data. In this chapter, we introduce a new joint model with shared random effects for a longitudinal marker with ordinal values and time-to-event data. The model is estimated in the maximum likelihood framework. Based on the developments of Arnold [2009] and Barrett et al. [2015], the event timescale is discretised and a closed-form of the integral over the random effects in the likelihood is found, contrary to the classic inference in joint models with shared random effects which requires numerical approximations of this integral. This work extends the spectrum of the analysis of longitudinal and time-to-event correlated data while proposing a new inference technique which should allow reliable, fast and efficient results. The present chapter introduces only the context and the methodology as this work is still in progress. A simulation study is planned to investigate the performances of the estimation program, and a R package will be probably developed with nice functionalities for a large dissemination.

6.1 Introduction

During my PhD thesis, I carried out a consultancy mission at MAPI Lyon (statistical consulting company for health applications) which aimed to check a secondary endpoint in a phase III clinical trial. The focus was on the post-treatment evolution of quality of life in subjects treated for a specific cancer. To answer to this question, clinicians collected repeated measurements of Health-Related Quality of Life (HRQoL) data over patient's follow-up from the end of treatment until the disease progression or death from any cause (or lost of follow-up). HRQoL data were collected using standard questionnaires such as the EORTC QLQ-C30 questionnaire, which evaluates quality of life in patients with cancer of any type [Aaronson et al., 1993; Osoba et al., 1994], or other questionnaires more specific to the studied disease.

In these questionnaires, subjects respond to several close-ended questions/items, scores are then built from these responses to quantify non-measurable concepts of quality of life such as the subject's perception of its treatment. In practice, these scores are built with a linear interpretation of their possible values, but they are discrete and most of them only consider a few modalities. Thus these discontinuous scores should be treated as ordinal markers measured longitudinally, with possibly informative drop-out.

Based on this experience, we decided to explore another key point in the joint modelling which concerns the distribution of the studied longitudinal marker [Henderson et al., 2000; Tsiatis and Davidian, 2004]. Indeed, the choice of the longitudinal submodel must imperatively be guided by the marker's nature.

In the recent years, joint models implementation has been well addressed with the emergence of multiple software [Rizopoulos, 2010; Philipson et al., 2012; Crowther et al., 2013] which greatly helped the dissemination of the technique to the applied research community and its usage in practice. But most of these software focus on a longitudinal Gaussian marker. More recently, Rizopoulos [2016] developed a R package for fitting joint models for longitudinal and survival data using Markov chain Monte Carlo (MCMC) techniques, which enhances various choices of distributions for the longitudinal marker. But the software still does not manage any ordinal longitudinal markers. Yet, joint models for a longitudinal ordinal response variable have been introduced in the literature [Li et al., 2010; Proust-Lima et al., 2016]. But to our knowledge, no software was ever made available to estimate them.

We thus decided to fill this gap and propose a new joint model with shared random effects for the repeated measurements of an ordinal longitudinal marker and time-to-event

data, and its associated software. However, a well-known complexity in joint models with shared random effects is the hard computational inference. Indeed the likelihood involves an integral over the individual random effects which has no closed form, and thus requires numerical integration algorithms such as Gaussian quadratures. Although powerful developments have been proposed in other software, as for example the use of pseudo-adaptive Gauss-Hermite quadratures [Rizopoulos, 2012a], an accurate estimation of the model parameters may remain costly demanding in practice, and the choices in both the correlation structures of the longitudinal data, and the association structures between longitudinal and survival data, are then restricted.

To circumvent the problem, Barrett et al. [2015] proposed an ingenious trick which consists to discretise the timescale of the event process and replace the commonly used proportional hazards model by a sequential probit model [Albert and Chib, 2001]. Using the results of Arnold [2009] on skew-normal distributions, a likelihood expression was found as freed of any integral over the random effects. Thus, the model allowed the inclusion of complex variance-covariance structure for the data, such as Gaussian autoregressive processes or Brownian motions. The use of complex structure of association between the longitudinal and survival processes was also possible.

We chose to extend this technique for studying a longitudinal marker which is ordinal. Using a cumulative probit mixed model for the marker data and a sequential probit model for the discretised version of the time-to-event data, an exact expression of the likelihood is found.

The rest of the Chapter is as follows. We first introduce the joint model for an ordinal longitudinal marker and time-to-event data in Section 6.2. Details on the formulation of the likelihood, which is freed of any integral over the individual random effects, are also given. As mentioned earlier, the work is still in progress and for instance, the simulation study is not finalized. So the Chapter ends with a discussion on the expected results and the provision for the statistical and clinical communities of such a model, in Section 6.3.

6.2 Methods

6.2.1 Notations

For each subject $i \in \{1, \dots, N\}$, we observe repeated measurements of an ordinal longitudinal marker until the occurrence of an event.

In classical survival analysis, the event timescale is assumed to be continuous: we

observe the event time $T_i^* = \min(T_i, C_i)$ with T_i the true event time and C_i the right censoring time. The indicator of event is denoted $\delta_i = \mathbb{1}\{T_i^* \leq C_i\}$, and equals 1 if the subject experienced the event, 0 if he was censored. In this work we consider instead a discrete time variable S_i which denotes the time interval in which T_i^* occurred. The discrete event timescale decomposes the continuous event timescale into $S + 1$ time intervals. Let us denote l_s and u_s the lower and upper bounds, respectively, of each time interval $s \in \{1, \dots, S + 1\}$, with $l_s = u_{s-1}$ for $s > 1$. The S^{th} first intervals are finite and discretise the observed event timescale, with the lower bound of the first time interval such that $l_1 = 0$, and the upper bound of the S^{th} time interval greater than the last observed time in all subjects, i.e. $u_S > \max_i(T_i^*)$. Finally the $(S + 1)^{\text{th}}$ interval does not contain any observed event time, and $u_{S+1} = +\infty$. In the following, the discrete timescale for the observed events is denoted $\mathcal{S} = \{1, \dots, S\}$, i.e. $\forall i \in \{1, \dots, N\}, S_i \in \mathcal{S}$. We also introduce $\tilde{t}(s) = (l_s + u_s)/2$ as the midpoint of the time interval s , $s \in \mathcal{S}$.

By contrast, the timescale for the longitudinal marker measurements is considered as continuous. We denote $Y_i = (Y_i(t_{i1}), \dots, Y_i(t_{in_i}))^\top$ the vector of the n_i collected measurements for subject i , with $Y_i(t_{ij})$ the marker's observation at time t_{ij} and $t_{in_i} \leq T_i^*$. Without loss of generality, we consider that the longitudinal ordinal marker Y_i is with possible values in the finite state space $\mathcal{K} = \{0, 1, \dots, K\}$, where K denotes the maximum possible value for Y_i .

6.2.2 Joint model

We propose a joint model with shared random effects which is decomposed into two submodels: a cumulative probit mixed model for the repeated measurements of the ordinal longitudinal marker, and a sequential probit model for the discretised time-to-event data; both linked using a function of the individual shared random effects.

Cumulative probit mixed submodel

For the ordinal longitudinal data, we define a cumulative probit mixed submodel which explains the probability that the marker observation be lower or equal to a given value according to covariates at any time t (see Chapter 2 for more details):

$$\Pr(Y_i(t) \leq k | b_i) = \Phi \left(\alpha_{k+1} - X_i^L(t)^\top \beta - Z_i(t)^\top b_i \right) \quad (6.1)$$

for $k \in \{\mathcal{K} - \{K\}\}$, where $X_i^L(t)$ and $Z_i(t)$ denote p -vector and q -vector of (possibly time-dependent) covariates, respectively associated with the fixed effects β and the individual random effects b_i , $b_i \sim \mathcal{N}(0, G)$ with G a general covariance structure. The ordered

thresholds $(\alpha_k)_{k \in \{\mathcal{K}, K+1\}}$ are unknown parameters, except $\alpha_0 = -\infty$ et $\alpha_{K+1} = +\infty$. Finally, the function $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution.

To ensure model's identifiability, two parameter constraints are required. The formulation (6.1) implies that the measurement error terms of the underlying Gaussian marker are independent and with unit variance. Another location constraint can be added by considering for instance that the fixed intercept is null, or by constraining a threshold parameter such as $\alpha_1 = 0$.

Sequential probit submodel

For the survival part, after discretisation of the event time, we are interested in the probability that the subject had the event in time interval $r \in \mathcal{S}$ given that he did not experience the event in previous intervals.

Here the idea is to consider S Gaussian latent variables $\{\xi_{i,r}; r = 1, \dots, S\}$ which are associated to the observed event time S_i such that [Albert and Chib, 2001]:

$$S_i = \begin{cases} 1 & \text{if } \xi_{i,1} \leq 0 \\ 2 & \text{if } \xi_{i,1} > 0, \xi_{i,2} \leq 0 \\ \vdots & \vdots \\ S & \text{if } \xi_{i,1} > 0, \dots, \xi_{i,(S-1)} > 0, \xi_{i,S} \leq 0 \\ S + 1 & \text{if } \xi_{i,1} > 0, \dots, \xi_{i,S} > 0. \end{cases}$$

For identifiability purpose, we fix the variance of $\xi_{i,r}$ as equal to 1. By considering that the expectation of $\xi_{i,r}$ is explained by covariates and random effects such that $\mathbb{E}(\xi_{i,r}) = X_{i,r}^E \top \gamma_r + \eta_r \top W_{i,r} b_i$, we obtain the following sequential probit model:

$$\Pr(S_i = r | S_i \geq r, b_i) = 1 - \Phi \left(X_{i,r}^E \top \gamma_r + \eta_r \top W_{i,r} b_i \right) \quad (6.2)$$

for $r \in \mathcal{S}$ and $\Pr(S_i = S + 1 | S_i \geq S) = 1$. $X_{i,r}^E$ is a \tilde{p} -vector of (possibly time-varying) covariates associated with the vector of (possibly time varying) parameters γ_r . For ease of calculations, we define $W_{i,r} b_i$ rather than $W_i(b_i, r)$ as in the previous chapters, where $W_{i,r} b_i$ is a (possibly time varying) \tilde{q} -vector which depicts the nature of the dependency between the dynamics of the longitudinal marker and the conditional probability of event. It is associated with the vector of (possibly time varying) parameters η_r .

Examples of model specification

As widely developed in Barrett et al. [2015], the model's specification allows various types of random effects structures and association functions. This part is inspired from

this article.

For example, a simple specification considers a linear trajectory of the marker over time with random intercept and slope, i.e. $Z_i(t) = (1, t)^\top$, and the random deviations $b_i = (b_{i,0}, b_{i,1})^\top$ are with variance-covariance matrix $G^{IS} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, where ρ , σ_1 and σ_2 are unknown parameters. In that case, several choices of association structure may be chosen, such as simply the individual deviations ($W_{i,r}b_i = (b_{i,0}, b_{i,1})^\top$) or their interaction with time functions ($W_{i,r}b_i = (b_{i,0}, b_{i,1}t^*(r))^\top$).

In their proposal as in Henderson et al. [2000], Barrett et al. [2015] also considered Gaussian processes which take into account a temporal correlation in the data, such as the continuous auto-regressive process (AR1) or the Brownian motion (BM). In that case, a discretised version of the Gaussian process is considered so that $b_i = (b_{i,1}, \dots, b_{i,S})^\top$ with variance matrix G^{GP} which includes the elements $G_{l,m}^{GP} = \sigma_w^2 \exp(-\rho|\tilde{t}(l) - \tilde{t}(m)|)$ in the case of AR1, or $G_{l,m}^{GP} = \sigma_w^2 \min(\tilde{t}(l), \tilde{t}(m))$ in the case of BM, both for $l, m \in \mathcal{S}$; with σ_w^2 and ρ two unknown parameters to estimate. All the measures collected into the time interval r share the same $b_{i,r}$. Then, at time $r \in S$, the dependence function can take the form $W_{i,r}b_i = b_{i,r}$.

One may also choose more complex specifications for the random effects, such as combinations of the ones defined above, with for example inclusion of both the individual deviations and stochastic deviations for a Gaussian centred process. The random effects b_i are now a vector of size $S + 2$ with variance matrix $G^{IS,GP} = \begin{pmatrix} G^{IS} & 0 \\ 0 & G^{GP} \end{pmatrix}$, and the dependence function can be chosen as $W_{i,r}b_i = (b_{i,0}, b_{i,1}, b_{i,r})^\top$.

In all these examples, note that the associated coefficients η_r may be considered as specific to each time interval r or constant over time intervals.

6.2.3 Estimation

The model is estimated by maximizing the likelihood $L(\theta)$ where θ is the complete set of unknown parameters.

Initial expression of the likelihood

The longitudinal and survival processes are independent conditionally to the random effects. The likelihood is expressed as the product of the individual contributions over the

subjects:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^N L_i(Y_i, S_i, \delta_i; \theta) \\ &= \prod_{i=1}^N \int_{\mathbb{R}^q} f_Y(Y_i|b_i; \theta) f_E(S_i, \delta_i|b_i; \theta) f_b(b_i; \theta) db_i \end{aligned} \quad (6.3)$$

where $f(\cdot)$ denotes a probability density function.

The conditional density of the longitudinal ordinal outcome is

$$f_Y(Y_i|b_i; \theta) = \Phi^{(n_i)}(\zeta_{1,i} - X_i^L \beta - Z_i b_i, \zeta_{2,i} - X_i^L \beta - Z_i b_i; 0, \sigma^2 I_{n_i}) \quad (6.4)$$

where X_i^L is a (n_i, p) matrix with row vectors $X_i^L(t_{ij})^\top$ for $j = 1, \dots, n_i$ and Z_i is a (n_i, q) matrix with row vectors $Z_i(t_{ij})^\top$ for $j = 1, \dots, n_i$. $\zeta_{1,i}$ and $\zeta_{2,i}$ are n_i -vectors, both with elements in $\{\alpha_m; m = 0, \dots, M\}$. Finally, $\Phi^{(d)}(a, b; 0, \Sigma)$ is the cumulative distribution function of a multivariate Gaussian distribution of dimension d , truncated at a , computed at b , with null mean and variance Σ .

The conditional density of the discretised time-to-event outcome is

$$\begin{aligned} f_E(S_i, \delta_i|b_i; \theta) &= \prod_{r=1}^{S_i-1} \Phi \left(X_{i,r}^E \gamma_r + \eta_r^\top W_{i,r} b_i \right) \times \\ &\quad \left[\Phi \left(X_{i,S_i}^E \gamma_{S_i} + \eta_{S_i}^\top W_{i,S_i} b_i \right) \right]^{1-\delta_i} \times \left[1 - \Phi \left(X_{i,S_i}^E \gamma_{S_i} + \eta_{S_i}^\top W_{i,S_i} b_i \right) \right]^{\delta_i}. \end{aligned} \quad (6.5)$$

Finally, the density of the Gaussian random effects is the probability density function of a Gaussian distribution with dimension q , null mean, and variance G , denoted:

$$f_b(b_i; \theta) = \phi^{(q)}(b_i; 0, G). \quad (6.6)$$

Arnold's formula

The likelihood (6.3) involves a q -integral over the individual random effects b_i . To avoid the use of numerical integration techniques, the trick is to use a relation about the skew-normal distributions which is defined in Arnold [2009]:

$$\int \Phi^{(m)}(\nu_1 + \Omega b^*, \nu_2 + \Omega b^*; 0, V) \phi^{(q)}(b^*; 0, I_q) db^* = \Phi^{(m)}(\nu_1, \nu_2; 0, V + \Omega \Omega^\top) \quad (6.7)$$

where ν_1 and ν_2 are m -vectors, b^* is a q -vector, Ω is a (m, q) matrix, and V is a (m, m) matrix.

Final expression of the likelihood

To use the Arnold's formula, we first rescale the random effects as $b_i^* = G^{-1/2}b_i \sim \mathcal{N}(0, I_q)$. Then the density of the random effects becomes

$$f_b(b_i^*; \theta) = \phi^{(q)}(b_i^*; 0, I_q)$$

and the conditional density of the longitudinal outcomes (6.4) is now expressed as

$$f_Y(Y_i|b_i^*; \theta) = \Phi^{(n_i)}(\nu_{1,i}^L + \Omega_i^L b_i^*, \nu_{2,i}^L + \Omega_i^L b_i^*; 0, \sigma^2 I_{n_i})$$

with $\nu_{1,i}^L = \zeta_{1,i} - X_i^L \beta$, $\nu_{2,i}^L = \zeta_{2,i} - X_i^L \beta$ and $\Omega_i^L = -Z_i G^{1/2}$.

Afterwards we rewrite the conditional density of the survival outcomes (6.5) as

$$\begin{aligned} f_E(S_i, \delta_i|b_i^*; \theta) = & \\ & \delta_i \times \Phi^{(S_i-1)} \left(\nu_{1,i,\{1:(S_i-1)\}} + \Omega_{i,\{1:(S_i-1)\}} b_i^*, \nu_{2,i,\{1:(S_i-1)\}} + \Omega_{i,\{1:(S_i-1)\}} b_i^*; 0, I_{S_i-1} \right) + \\ & (-1)^{\delta_i} \times \Phi^{(S_i)} \left(\nu_{1,i,\{1:S_i\}} + \Omega_{i,\{1:S_i\}} b_i^*, \nu_{2,i,\{1:S_i\}} + \Omega_{i,\{1:S_i\}} b_i^*; 0, I_{S_i} \right) \end{aligned}$$

with $\nu_{1,i,\{1:S_i\}}^E$ the column vector of size S_i whose elements are all equal to negative infinity, and $\Omega_{i,\{1:S_i\}}^E$ the (S_i, \tilde{q}) matrix which stacks the row vectors $\Omega_{i,r}^E = \eta_r^\top W_{i,r} G^{1/2}$ for $r = 1, \dots, S_i$. Finally, $\nu_{2,i,\{1:S_i\}}^E$ denotes a S_i -vector with stacked elements $\nu_{2,i,r}^E = X_{i,r}^E \top \gamma_r$, for $r = 1, \dots, S_i$.

By combining the two conditional densities, we obtain a unique Gaussian CDF:

$$\begin{aligned} f_Y(Y_i|b_i^*; \theta) f_E(S_i, \delta_i|b_i^*; \theta) = & \\ & \delta_i \times \Phi^{(n_i+S_i-1)} \left\{ \begin{pmatrix} \nu_{1,i}^L + \Omega_i^L b_i^* \\ \nu_{1,i,\{1:(S_i-1)\}}^E + \Omega_{i,\{1:(S_i-1)\}}^E b_i^* \end{pmatrix}, \begin{pmatrix} \nu_{2,i}^L + \Omega_i^L b_i^* \\ \nu_{2,i,\{1:(S_i-1)\}}^E + \Omega_{i,\{1:(S_i-1)\}}^E b_i^* \end{pmatrix}; \right. \\ & \left. 0, \begin{pmatrix} \sigma^2 I_{n_i} & 0 \\ 0 & I_{S_i-1} \end{pmatrix} \right\} + \\ & (-1)^{\delta_i} \times \Phi^{(n_i+S_i)} \left\{ \begin{pmatrix} \nu_{1,i}^L + \Omega_i^L b_i^* \\ \nu_{1,i,\{1:S_i\}}^E + \Omega_{i,\{1:S_i\}}^E b_i^* \end{pmatrix}, \begin{pmatrix} \nu_{2,i}^L + \Omega_i^L b_i^* \\ \nu_{2,i,\{1:S_i\}}^E + \Omega_{i,\{1:S_i\}}^E b_i^* \end{pmatrix}; \right. \\ & \left. 0, \begin{pmatrix} \sigma^2 I_{n_i} & 0 \\ 0 & I_{S_i} \end{pmatrix} \right\}. \end{aligned}$$

Finally, by applying Arnold's formula, the likelihood can be expressed as

$$\begin{aligned} L(\theta) = \prod_{i=1}^N \left[\delta_i \times \Phi^{(n_i+S_i-1)} \left\{ \nu_{1,i,\{1:(n_i+S_i-1)\}}^*, \nu_{2,i,\{1:(n_i+S_i-1)\}}^*; 0, V_{i,\{1:(n_i+S_i-1)\}}^* \right\} + \right. \\ \left. (-1)^{\delta_i} \times \Phi^{(n_i+S_i)} \left\{ \nu_{1,i,\{1:(n_i+S_i)\}}^*, \nu_{2,i,\{1:(n_i+S_i)\}}^*; 0, V_{i,\{1:(n_i+S_i)\}}^* \right\} \right] \quad (6.8) \end{aligned}$$

where $\nu_{1,i,\{1:(n_i+S_i)\}}^* = (\nu_{1,i}^{L\top}, \nu_{1,i,\{1:S_i\}}^{E\top})^\top$ is a vector of size $(n_i + S_i)$, $\nu_{2,i,\{1:(n_i+S_i)\}}^* = (\nu_{2,i}^{L\top}, \nu_{2,i,\{1:S_i\}}^{E\top})^\top$, is a vector of size $(n_i + S_i)$, and finally $V_{i,\{1:(n_i+S_i)\}}^* = \begin{pmatrix} \sigma^2 I_{n_i} & 0 \\ 0 & I_{S_i} \end{pmatrix} + \begin{pmatrix} \Omega_i^L \\ \Omega_{i,\{1:S_i\}}^E \end{pmatrix} \begin{pmatrix} \Omega_i^L \\ \Omega_{i,\{1:S_i\}}^E \end{pmatrix}^\top$ is a matrix of size $(n_i + S_i, n_i + S_i)$.

Thus the likelihood is finally found as free of any integral over the individual random effects. However, it involves the computation of two cumulative density functions of truncated Gaussian distributions which is not straightforward neither but for which accurate approximations based on numerical integration algorithms exist as explained below.

6.2.4 Implementation

The estimation program of the joint model has been implemented in R but is still not formally validated. Model parameters θ are estimated by maximizing the log term of the likelihood (6.8) using a Marquardt optimisation algorithm [Marquardt, 1963] implemented in the R package `marqLevAlg`.

The likelihood requires the computation of cumulative distribution functions of truncated multivariate Gaussian distributions. In practice, such quantities must be approximated using numerical integration algorithms whose accuracy degree increases with computational time. After investigations, we used the R package `mnormt` when the dimension of the Gaussian CDF was less than 20 and the R package `mvtnorm` otherwise.

To initialize the optimisation algorithm, both submodels are estimated separately. The cumulative probit mixed model is estimated with the same exact likelihood inference which ignores the time-to-event information, while the estimation of the sequential probit model is estimated by ignoring the dependence structure (the vector of association parameters is fixed as null: $\eta_r = 0, \forall r \in \mathcal{S}$).

The perspective is to include this estimation program into a R package. Nice functionalities are already allowed, with notably the inclusion of complex association structures between the longitudinal and survival processes, such as auto-regressive processes or Brownian motions.

6.3 Discussion

Because of the absence of adapted and efficient software, clinicians and statisticians often rely on inappropriate methods to analyse ordinal longitudinal data and time-to-event data, when both are correlated. For example, Ediebah et al. [2015] proposed a joint model for longitudinal HRQoL data and survival data, by focusing on appetite loss score

which involved four possible ordered modalities. But authors did not take into account the ordinal nature of this marker, and only used a joint model adapted for a Gaussian longitudinal marker.

In this sense, we developed and implemented a new joint model with shared random effects for repeated measurements of an ordinal longitudinal marker and time-to-event data which are both correlated. We circumvented the usual inference problems in such models using the results of Arnold [2009] on skew-normal distributions. Based on the papers of Barrett et al. [2015] and Barrett and Su [2017], we defined the event process in discrete time. An exact likelihood -freed of any integral over the individual random effects- was finally found. The model thus allows more complex correlation structures between longitudinal data and association structure between longitudinal and survival data, than the ones usually used. Indeed, it enables for example the inclusion of autoregressive processes while the classical inference technique (which involves an integral over individual random effects in the likelihood) usually restricts the number of random effects to 2 or 3 in practice.

The likelihood involves however the computation of cumulative probability distributions of multivariate skew normal distributions, which may be accurately approximated by much more efficient algorithms [Genz, 1992] than for the usual integral over the random effects. However, such algorithms usually handle the computation of cumulative probability distributions with dimensions less than 20. Above, accuracy might be altered. We initiated a simulation study to investigate the performances of our estimation program with different scenarii of study design. But it was too preliminary to show the results in the present chapter.

Once validated, this methodology will probably find many applications in the analysis of quality of life data in cancer, but also beyond with the rise of subjective data including quality of life or psychometrics in epidemiological research. This is the reason why we want to provide a user-friendly R package.

7 Discussion

This thesis introduced statistical developments concerning two kinds of data currently encountered in longitudinal health studies: repeated marker data and event history data. Many techniques exist in the literature to handle these types of data when they are considered separately, with effective and famous software for most of them. When longitudinal and time-to-event data are correlated, which is frequent in health studies, statistical methods based on joint modelling techniques also exist and some software routines start to be deployed. However, some open issues still arise from the complexity of some disease studies. I addressed some of them in this thesis in the case of cancer progression study.

7.1 Joint multi-state modelling

In the thesis, I was particularly interested in the progression of subjects with a localized prostate cancer who were monitored from the end of their treatment using repeated measures of the prostate specific antigen (PSA). These patients can experience -possibly successively- several types of clinical recurrences (loco-regional or distant) and death. They can also receive an additional treatment which was unplanned at baseline, but associated with the disease evolution. In the data I had access to, the patients were initially treated by radiation therapy and might initiate an hormonal therapy at any time.

In order to model the disease in its whole, a proper solution was to not only distinguish the several clinical stages that may experience the patient, but also characterize the transitions between them, using a multi-state process. To account for the trajectory of PSA which is central in prostate cancer follow-up, I also had to model the repeated measures of PSA jointly. However, no solution existed in the literature to perform such an analysis when the longitudinal process and the multi-state process were correlated. Joint models had been extended to competing risks setting [Elashoff et al., 2008; Rizopoulos, 2012b; Andrinopoulou et al., 2014] but not multi-state data.

In Chapter 3, I thus proposed a joint model with shared random effects for a Gaussian longitudinal process and a Markov non-homogeneous multi-state process, which properly takes into account the correlation between the longitudinal process and the multi-state

process. The model provides a quantification of the impact of the PSA dynamics and other prognostic factors measured at the end of treatment on each transition intensity between clinical health states.

When developing sophisticated models, one recurrent question regards the underlying assumptions and the goodness of fit evaluation. In particular, in this joint multi-state model, one could wonder whether, after adjustment on the individual covariates, there subsisted a correlation between the individual transition times which would violate the Markov assumption of the multi-state process. Such a residual correlation could be modelled by an additional individual random effect in the survival model, called frailty term. But the inference in a joint multi-state frailty model would be very complex, and interpretation of the model parameters would be not effortless. Instead, I proposed in Chapter 4 a score test for the inclusion of a log-normal frailty term in the multi-state part of a joint multi-state model, which avoids the estimation of the model when including the frailty. Interestingly, I showed that such a test not only tested the Markovian assumption of the multi-state process, but also constituted a useful goodness-of-fit tool, for example to detect outlier subjects or more generally a lack of fit of the model. In the application on the prostate cancer data, the score test allowed us to improve the model specification and define a relevant and efficient joint multi-state model. The main results confirmed the significant deleterious impact of PSA increase on the hazard to experience a disease recurrence (loco-regional recurrence, distant recurrence, hormonal therapy administration) from the end of the radiotherapy. However the PSA dynamics were not significantly associated to the risk of death without recurrence after radiotherapy. Finally, after a first disease recurrence higher PSA dynamic features were not necessarily associated to a negative evolution of the disease. It supported that in the advanced stages of the disease, other criteria than the PSA dynamics should be considered.

An essential aspect of statistical research consists of supplying portable statistical software associated with statistical developments for the community. I provided a R function for the estimation of the joint multi-state model which pre-processes the data with `mstate` package and then exploits all the facilities of `JM` package for the estimation of shared random effect models. I also programmed the score test into a function which can be used on an estimated joint multistate model object. I included in these functions additional functionalities, such as the inclusion of non-linear marker dynamics in the association structure between the longitudinal and the multi-state processes, and techniques for accuracy enhancement in the model estimates.

GENERAL DISCUSSION

Indeed, most of the time, the trajectory of the biomarker is included in the event history model through a couple of features, mostly its current level or slope, without further exploration of the log-linearity assumption. Yet, in prostate cancer, it has been shown the current level of PSA did not have a loglinear relationship with any type of cancer recurrence and that mostly PSA levels between 0 and 4 ng/mL were associated with loglinear increase in the risk of recurrence [Sène et al., 2014]. In the joint multi-state model estimation function, any transformation of the current level can now be considered. In addition, we showed in the application of the score test, that such departure from the loglinearity assumption could translate in a higher rejection of the Markov assumption.

Regarding estimation accuracy, we found out when developing the joint multi-state model that the type of numerical integration over the random effects considered for the likelihood computation could substantially impact the estimation quality. In JM package, a pseudo-adaptive quadrature was used [Rizopoulos, 2012a]. It exploits the idea of adaptive Gaussian quadrature [Lesaffre and Spiessens, 2001] with a recentering and standardization of the integral around the random effects without recalculating this standardisation at each iteration but using a pseudo recentering and standardization around the predicted random effects in the linear mixed model. I extended this technique to a multi-step pseudo-adaptive Gauss–Hermite rule. The procedure is then repeated: the joint model is estimated once using the pseudo-adaptive technique, and then it can be reestimated several times by starting from the previously estimated parameters and centering the integral on the predicted random effects derived from the joint model rather than on the linear mixed model. With this technique, estimates were found as more accurate while using a relatively small number of quadrature points and thus reducing computational time.

I applied this work in the specific context of progression of cancer in men with a localized prostate cancer treated by external beam radiotherapy only. However, men with a localized prostate cancer are more and more treated by other therapies, the use of radiotherapy alone becoming somewhat obsolete in such a population [Laverdière et al., 1997; Mottet et al., 2017]. Thus, expansions of this work could be to focus on subjects with localized disease treated by:

- combined radiotherapy and hormonotherapy. However, PSA trajectories after such treatment may be heterogeneous and a joint model considering latent classes would possibly be more adapted to this population. Such a development was done in a very recent paper of Rouanet et al. [2016], who considered jointly a Gaussian

longitudinal marker and a three-states multi-state process;

- surgery (ablation of the prostate). However, PSA are supposed to be close to zero after such intervention with possibly an increase when eventually the disease evolves again. The linear mixed model as considered for now may not apply in that context with many zeros, and a joint model with a tobit mixed sub-model for the observed PSA might be required [Pike and Weissfeld, 2013].

As in the context of radiotherapy alone, it would be interesting in these population to:

- describe the disease evolution in its whole (by considering several kinds of clinical recurrence and additional treatments unplanned at baseline, such as hormonotherapy) from the PSA dynamics and other prognostic factors;
- provide individual dynamic predictions of a given quantity of interest such as cumulative incidences of clinical recurrence.

The joint model including a multi-state submodel applies far beyond localized prostate cancer. In cancer research, it could also be used at more advanced stages, and in different locations. For instance, it could be used to jointly model clinical progressions and tumor size. I already contributed to such an analysis in Król et al. [2016], which modelled simultaneously tumor size, (possibly repeated) disease progressions and death, in subjects with metastatic colorectal cancer. However this paper did not distinguish the type of clinical progressions that could experience the subject. The joint multi-state model we proposed would be thus particularly relevant in this kind of application. Finally, the model could also be applied to other chronic diseases where multiple clinical progressions are of interest.

7.2 Joint modelling of ordinal longitudinal data and survival data

In observational cohorts or clinical studies, whether for describing the marker evolution, quantifying its impact on the disease evolution, or predicting a clinical quantity of interest from the marker data, a crucial point is to take care of the nature of the marker. A consultancy for a pharmaceutical laboratory during my thesis confronted me with the modelling of Health-Related Quality of Life (HRQoL) data in cancer clinical trials. These data were: 1) ordinal, because they are built from summary scores with few levels which characterize latent domains such as fatigue, nausea, pain, or others; 2) collected until the disease progression which stopped the collection of the HRQoL measurements.

In practice, no software was available to analyse such data, and the classical inference

proposed in the literature involved hard computational procedures with the computation of integrals over each individual random effects performed through costly numerical integration algorithms. We thus proposed in Chapter 6 a new joint model for an ordinal longitudinal marker and survival data which is decomposed into two sub-models: a cumulative probit mixed model for the ordinal measures of the longitudinal marker and a sequential probit model for the time-to-event data after discretisation of their time-scale. The asset of this technique, which was initially proposed for a joint model with a longitudinal Gaussian marker [Barrett et al., 2015] is that the likelihood is freed of any integral over the individual random effects from the mixed model. Indeed, the simplifications in the likelihood come from the combination of Gaussian density functions and Gaussian probability functions in the integrand, which is also the case with ordinal longitudinal markers.

At the end of the thesis, this work was only initiated with the writing of the model and the complete implementation of its estimation procedure. This work needs to be continued and finalized after the thesis with in particular the validation of the estimation program in a simulation study.

A first application is planned with data from the EORTC (European Organisation for Research and Treatment of Cancer) with the objective to illustrate the usefulness of the model to study HRQoL repeated measures in cancer clinical trials. It will allow both to properly specify the HRQoL evolution over time and quantify its impact on the clinical event occurrence, while properly taking into account the correlation between these two processes.

Overall, more and more subjective measures are included in health studies, and many are ordinal. As a result, this work may be useful in many areas. For example, in clinical trials that deal with repeated subjective data, and for which the evolution may be interrupted by an event. But also in epidemiological cohorts which include data from health measurement scales and which are interested in the link with a health incident. An example is disability in neurodegenerative diseases (e.g., Alzheimer's disease, multiple system atrophy.)

7.3 Individual dynamic predictions

Two main objectives are identified in epidemiological studies: studying the aetiology of the disease or predicting a clinical event of interest at the individual level. In fact, prediction becomes more and more central in many clinical applications. To compute

these predictions in a dynamic and individualized way, a natural wish is to take into account the complete history of the subject until the prediction time or, more realistically the complete history about specific information collected on the subject. Based on it, we would expect to accurately predict the future evolution of a patient.

Various contributions were published in biostatistics about individual dynamic prediction. In this work we were interested in cumulative incidences of event, computed from a prediction time and in the context of competing risks. Such a quantity could be computed from several approaches, notably using joint models [Rizopoulos, 2011], partly conditional models [Maziarz et al., 2017] or landmark models [van Houwelingen, 2007; Nicolaie et al., 2013b]. However, I felt that in this setting, there remained some vagueness on the correct definition of the quantity of interest, the validation of its proposed estimators, and the definition and their variability. In addition, there was a lack of comparison in terms of predictive abilities between developed estimators. Theoretical limits of each method were put forward but they were not supported by realistic or fair comparisons.

I filled this gap in Chapter 5 by introducing, validating and comparing dynamic tools for predicting cancer progression from biomarker data. The quantity of interest was the cumulative incidence of a competing event, computed according to the individual longitudinal information (including biomarker data) collected until a prediction time, and I compared the joint modelling approach and the landmarking approach with different specifications of models. Two definitions of the variability in the predictions are proposed, one which only takes into account the variability due to the learning sample, and another which adds the variability due to the marker's observations of the subject to predict. Finally the estimators detailed in this chapter are individualized, powerful and adapted to meet the needs of patients and their clinicians. They are also implemented in R, with guidelines provided for a correct practical use by clinicians and statisticians.

In summary, the extensive simulation study we carried out confirmed that the joint model is usually a more efficient method than the landmarking approach to compute individual predictions. However, the simulations also warned us on the necessity to:

- accurately specify the structure of dependency that links the longitudinal marker with the event model, whatever the approach;
- accurately specify the longitudinal trajectory of the marker, especially in joint models, in order to obtain accurate predictions.

Before deriving a predictive tool from the joint model, it is thus essential to use diagnostic tools for the goodness-of-fit of the longitudinal outcomes. Dobson and Henderson

GENERAL DISCUSSION

[2003] and Rizopoulos et al. [2010] proposed diagnostic tools based on conditional residuals and multiple imputation residuals, respectively, both based on informal graphical procedures. But it would be useful to develop goodness-of-fit tests, such as done in standard linear mixed models [Claeskens and Hart, 2009], to correctly evaluate the fit of the longitudinal sub-model in the joint models.

The misspecification problem of the longitudinal marker trend in the usual joint models leads us to envisage other ways to model the longitudinal marker trajectory. In system biology, mechanistic models are used to directly translate an hypothesized biological mechanism into a system of differential equations that are then linked with the observation of biomarkers [Antia et al., 2005]. By integrating biomarkers measurements into biological insights, such models are expected to better fit the longitudinal data. Mostly developed in HIV [Perelson et al., 1993; Prague et al., 2013] or in infectious diseases [Lee et al., 2009], some first examples exist in prostate cancer [Desmée et al., 2017] and might be a direction of improvement.

In prostate cancer prediction of clinical relapse as in other diseases, there is a gap between biostatistical developments and clinical practice. Indeed, in treated subjects with a localized prostate cancer, clinicians usually use naive methods to account for the PSA dynamics and guide their clinical decision. For instance, predictions of clinical relapse are done according to the last observed PSA value, the observed PSA doubling time (expected duration to double the observed PSA value) or the observed PSA velocity (slope of the observed PSA over time) [Ng et al., 2009]. Yet, these dynamics do not consider that the marker is measured with error, at discrete times, may be subject to informative censoring, and they do not make use of information from an entire population. The purpose of joint models in that context is to take into account all these aspects to provide a validated quantification of individual predictions.

The simulations already demonstrated the superiority of the joint approach and two-stage landmark approach over the naïve approach that only considers the last PSA measure available. However, it would be of particular interest to expand our comparison of predictive tools to compare predictive performances of the tools used in practice by the clinicians, and the tools derived from the joint modelling or two-stage landmarking approaches not in simulations but on real data with both training data to develop the prediction tools and validation data to assess their performances.

Once the best prediction model selected, a software solution might be developed to propose a reliable and efficient risk-calculator to clinicians, incorporating variability of

the predictions, and with a user-friendly interface. The objective of such a tool would not obviously be to replace the clinician but rather to help him during its decision making.

Our work considered only one longitudinal marker. However, it would certainly be relevant when developing individual dynamic predictions to account for multiple markers, possibly correlated and with various natures, such as the tumor size, blood-based biomarkers or HRQoL markers for example [Andrinopoulou et al., 2014]. With multiple longitudinal markers, estimation of joint models might become infeasible since consideration of multiple markers induces a large increase in the number of random effects, and thus in the dimension of the numerical integrals. In that context, two-stage landmark models constitute a serious alternative. These landmark models could be estimated at predefined landmark times or their supermodel version might be explored [Nicolai et al., 2013a]. Such a model would allow obtaining predictions at any prediction time, by combining different landmark models in one and thus smoothing the predictions over the prediction times.

Individual predictions derived from the joint multi-state model would also be reliable, allowing new quantities of interest to be predicted. This could include the time spent in a specific clinical state or the number of years gained with an additional treatment. This was explored in standard multi-state models in dementia context [Wanneveich et al., 2016] and might be expanded in the presence of a biomarker with the joint model extension. From the joint modelling perspective, other structures than the one we developed in Chapter 3 with transition-specific proportional intensity model could be relevant for prediction. For example, considering a sequential probit sub-model for the survival sub-part (as considered in Chapter 6 for the joint ordinal model) would directly model the probability of having the event in a given time interval. The pseudo-value approach in its joint version might properly account for the longitudinal information brought by the marker and directly model various quantities of interest depending on the link function considered.

Finally, we argued in Chapter 5 that the use of individual dynamic predictions in clinical practice could lead to a dynamic personalized medicine, whether for adapting the treatment strategies according to the updated individual probabilities [Sène et al., 2016], or optimizing the monitoring for each patient [Rizopoulos et al., 2015]. The later might be useful for the planning of clinical trials, notably with the computation of the required number of subjects in the trial, while integrating the individual and dynamic aspects of

the screening planning.

7.4 General conclusion

In longitudinal cancer studies, biomarker data of various natures may be used to either detect the disease, describe its evolution over time, or predict the subject's future. Based on prostate cancer research, we generally argued in this thesis that a disease evolution should be considered in its entirety, by characterizing the set of clinical events of interest, focusing on the (possibly multiple) transition(s) between them instead of only considering the first occurrence of a unique type of clinical event, and properly account for longitudinal biomarker data.

To conclude, this thesis opens up both clinical and statistical prospects, with the challenge to answer new questions asked by the clinicians, while including new longitudinal data such as HRQoL data or genomic data, for example. Novel statistical models will have to meet these needs while adapting to the features of the data and providing valid estimates with reliable and efficient inference methods as well as user-friendly interfaces to guarantee a certain practical use.

Bibliography

Note: page(s) referring to each article is (are) given after the reference.

- Aalen O. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4): 701–726, 1978. 50
- Aaronson N. K., Ahmedzai S., Bergman B., Bullinger M., Cull A., Duez N. J., Filiberti A., Flechtner H., Fleishman S. B., de Haes J. C., et al. The european organization for research and treatment of cancer qlq-c30: a quality-of-life instrument for use in international clinical trials in oncology. *JNCI: Journal of the National Cancer Institute*, 85(5): 365–376, 1993. 28, 192
- Abrahamowicz M., Bartlett G., Tamblyn R., and du Berger R. Modeling cumulative dose and exposure duration provided insights regarding the associations between benzodiazepines and injuries. *Journal of Clinical Epidemiology*, 59(4): 393–403, 2006. 63
- Albert J. H. and Chib S. Sequential ordinal modeling with applications to survival data. *Biometrics*, 57(3): 829–836, 2001. 59, 193, 195
- Allard W. J., Matera J., Miller M. C., Repollet M., Connelly M. C., Rao C., Tibbe A. G., Uhr J. W., and Terstappen L. W. Tumor cells circulate in the peripheral blood of all major carcinomas but not in healthy subjects or patients with nonmalignant diseases. *Clinical Cancer Research*, 10(20): 6897–6904, 2004. 29
- Andersen P. K. Life years lost among patients with a given disease. *Statistics in Medicine*, 36(22): 3573–3582, 2017. 47
- Andersen P. K. and Gill R. D. Cox’s regression model for counting processes: a large sample study. *The Annals of Statistics*, 10(4): 1100–1120, 1982. 34
- Andersen P. K. and Keiding N. Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11(2): 91–115, 2002. 53, 112
- Andersen P. K. and Pohar Perme M. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, 19(1): 71–99, 2010. 52, 134, 143, 144
- Andersen P. K., Borgan O., Gill R. D., and Keiding N. *Statistical Models Based On Counting Processes*. Springer, 1993. 47, 53, 56, 69, 143
- Anderson J. R., Cain K. C., and Gelber R. D. Analysis of survival by tumor response. *Journal of Clinical Oncology*, 1(11): 710–719, 1983. 59
- Andrinopoulou E.-R., Rizopoulos D., Takkenberg J. J., and Lesaffre E. Joint modeling of two longitudinal outcomes and competing risk data. *Statistics in Medicine*, 33(18): 3167–3178, 2014. 67, 201, 208
- Antia R., Ganusov V. V., and Ahmed R. The role of models in understanding cd8+ t-cell memory. *Nature Reviews Immunology*, 5(2): 101–111, 2005. 207
- Arnold B. C. Flexible univariate and multivariate models based on hidden truncation. *Journal of Statistical Planning and Inference*, 139(11): 3741–3749, 2009. 66, 191, 193, 197, 200

- Barrett J. and Su L. Dynamic predictions using flexible joint models of longitudinal and time-to-event data. *Statistics in Medicine*, 36(9): 1447–1460, 2017. 157, 200
- Barrett J., Diggle P., Henderson R., and Taylor-Robinson D. Joint modelling of repeated measurements and time-to-event outcomes: flexible model specification and exact likelihood inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1): 131–148, 2015. 66, 191, 193, 195, 196, 200, 205
- Barry M. J. Prostate-specific antigen testing for early diagnosis of prostate cancer. *New England Journal of Medicine*, 344(18): 1373–1377, 2001. 28
- Breslow N. E. Discussion of professor cox’s paper. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34: 216–217, 1972. 51, 142
- Brown E. R. Assessing the association between trends in a biomarker and risk of event with an application in pediatric hiv/aids. *The Annals of Applied Statistics*, 3(3): 1163, 2009. 63
- Chi Y.-Y. and Ibrahim J. G. Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, 62(2): 432–445, 2006. 67
- Claeskens G. and Hart J. D. Goodness-of-fit tests in mixed models. *Test*, 18(2): 213–239, 2009. 207
- Colosimo E., Ferreira F. v., Oliveira M., and Sousa C. Empirical comparisons between kaplan-meier and nelson-aalen survival function estimators. *Journal of Statistical Computation and Simulation*, 72(4): 299–308, 2002. 50
- Cooper G. M. *The Cell: A Molecular Approach*. Sinauer Associates, 2000. 27
- Cox D. R. Regression models and life-tables. volume 34, pages 187–220. 1972. 34, 51, 141
- Crowther M. J. and Lambert P. C. Simulating biologically plausible complex survival data. *Statistics in Medicine*, 32(23): 4118–4134, 2013. 112, 145
- Crowther M. J., Abrams K. R., Lambert P. C., et al. Joint modeling of longitudinal and survival data. *The Stata Journal*, 13(1): 165–184, 2013. 64, 192
- Dantan E., Joly P., Dartigues J.-F., and Jacqmin-Gadda H. Joint model with latent state for longitudinal and multistate data. *Biostatistics*, 12(4): 723–736, 2011. 35, 67, 106
- De Wreede L. C., Fiocco M., and Putter H. The mstate package for estimation and prediction in non-and semi-parametric multi-state and competing risks models. *Computer Methods and Programs in Biomedicine*, 99(3): 261–274, 2010. 57, 69
- Desmée S., Mentré F., Veyrat-Follet C., Sébastien B., and Guedj J. Using the saem algorithm for mechanistic joint models characterizing the relationship between nonlinear psa kinetics and survival in prostate cancer patients. *Biometrics*, 73(1): 305–312, 2017. 207
- Diggle P. and Kenward M. G. Informative drop-out in longitudinal data analysis. *Applied Statistics*, pages 49–93, 1994. 46, 61
- Diggle P. J., Heagerty P. J., Liang K.-Y., and Zeger S. L. *Analysis of Longitudinal Data*. Oxford University Press, 2002. 40
- Dobson A. and Henderson R. Diagnostics for joint longitudinal and dropout time modeling. *Biometrics*, 59(4): 741–751, 2003. 122, 206
- Duchateau L., Janssen P., Kezic I., and Fortpied C. Evolution of recurrent asthma event rate over time in frailty models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(3): 355–363, 2003. 48

BIBLIOGRAPHY

- Ediebah D. E., Galindo-Garre F., Uitdehaag B. M., Ringash J., Reijneveld J. C., Dirven L., Zikos E., Coens C., van den Bent M. J., Bottomley A., et al. Joint modeling of longitudinal health-related quality of life data and survival. *Quality of Life Research*, 24(4): 795–804, 2015. 199
- Eisenhauer E., Therasse P., Bogaerts J., Schwartz L., Sargent D., Ford R., Dancey J., Arbuck S., Gwyther S., Mooney M., et al. New response evaluation criteria in solid tumours: Revised recist guideline (version 1.1). *European Journal of Cancer*, 45(2): 228–247, 2009. 28
- Elashoff R. M., Li G., and Li N. A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics*, 64(3): 762–771, 2008. 35, 67, 106, 201
- Faucett C. L. and Thomas D. C. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, 15(15): 1663–1685, 1996. 35
- Ferlay J., Soerjomataram I., Dikshit R., Eser S., Mathers C., Rebelo M., Parkin D. M., Forman D., and Bray F. Cancer incidence and mortality worldwide: Sources, methods and major patterns in globocan 2012. *International Journal of Cancer*, 136(5), 2015. 27
- Ferrer L., Rondeau V., Dignam J., Pickles T., Jacqmin-Gadda H., and Proust-Lima C. Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer. *Statistics in Medicine*, 35(22): 3933–3948, 2016. 70, 106, 107, 109, 112, 114, 115, 116, 119, 122, 130, 139, 145, 163
- Fine J. P. and Gray R. J. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446): 496–509, 1999. 52, 134
- Fisher L. D. and Lin D. Y. Time-dependent covariates in the Cox proportional-hazards regression model. *Annual Review of Public Health*, 20(1): 145–157, 1999. 34
- Fitzmaurice G. M., Molenberghs G., and Lipsitz S. R. Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 691–704, 1995. 46
- Fitzmaurice G. M., Laird N. M., and Ware J. H. *Applied Longitudinal Analysis*. John Wiley & Sons, 2012. 43
- Follmann D. and Wu M. An approximate generalized linear model with random effects for informative missing data. *Biometrics*, pages 151–168, 1995. 63
- Genz A. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2): 141–149, 1992. 200
- Giobbie-Hurder A., Gelber R. D., and Regan M. M. Challenges of guarantee-time bias. *Journal of Clinical Oncology*, 31(23): 2963–2969, 2013. 58
- Goldstein B. A., Pomann G. M., Winkelmayr W. C., and Pencina M. J. A comparison of risk prediction methods using repeated observations: an application to electronic health records for hemodialysis. *Statistics in Medicine*, 36: 2750—2763, 2017. 35, 136, 158
- Greenwood M. The natural duration of cancer. *Reports on Public Health and Medical Subjects*, 33: 1–26, 1926. 49
- Hahn E. D. and Soyer R. Probit and logit models: Differences in the multivariate realm. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, pages 1–12, 2005. 45
- Hamdy F. C., Donovan J. L., Lane J. A., Mason M., Metcalfe C., Holding P., Davis M., Peters T. J., Turner E. L., Martin R. M., et al. 10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. *New England Journal of Medicine*, 375(15): 1415–1424, 2016. 29

- Harville D. A. Bayesian inference for variance components using only error contrasts. *Biometrika*, 61(2): 383–385, 1974. 42
- Henderson R., Diggle P., and Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4): 465–480, 2000. 46, 63, 192, 196
- Hougaard P. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73(2): 387–396, 1986. 60
- Hougaard P. Frailty models for survival data. *Lifetime Data Analysis*, 1(3): 255–273, 1995. 59
- Hougaard P. Multi-state models: a review. *Lifetime Data Analysis*, 5(3): 239–264, 1999. 53
- Hsieh F., Tseng Y.-K., and Wang J.-L. Joint modeling of survival and longitudinal data: Likelihood approach revisited. *Biometrics*, 62(4): 1037–1043, 2006. 63
- Jewell N. P. and Nielsen J. P. A framework for consistent prediction rules based on markers. *Biometrika*, 80: 153–164, 1993. 133, 136
- Joly P. and Commenges D. A penalized likelihood approach for a progressive three-state model with censored and truncated data: Application to aids. *Biometrics*, 55(3): 887–890, 1999. 57
- Joly P., Touraine C., Georget A., Dartigues J.-F., Commenges D., and Jacqmin-Gadda H. Prevalence projections of chronic diseases and impact of public health intervention. *Biometrics*, 69(1): 109–117, 2013. 47
- Jones R. H. and Boadi-Boateng F. Unequally spaced longitudinal data with ar (1) serial correlation. *Biometrics*, pages 161–175, 1991. 40
- Kalbfleisch J. D. and Prentice R. L. *The Statistical Analysis of Failure Time Data*, volume 360. John Wiley & Sons, 2011. 34, 60, 136
- Kaplan E. L. and Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282): 457–481, 1958. 49
- Król A., Ferrer L., Pignon J.-P., Proust-Lima C., Ducreux M., Bouché O., Michiels S., and Rondeau V. Joint model for left-censored longitudinal data, recurrent events and terminal event: Predictive abilities of tumor burden for cancer evolution with application to the ffcd 2000–05 trial. *Biometrics*, 72(3): 907–916, 2016. 35, 67, 106, 138, 157, 204
- Kurland B. F. and Heagerty P. J. Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. *Biostatistics*, 6(2): 241–258, 2005. 46
- Laird N. M. and Ware J. H. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982. 40
- Laverdière J., Gomez J., Cusan L., Suburu E. R., Diamond P., Lemay M., Candas B., Fortin A., and Labrie F. Beneficial effect of combination hormonal therapy administered prior and following external beam radiation therapy in localized prostate cancer. *International Journal of Radiation Oncology* Biology* Physics*, 37(2): 247–252, 1997. 203
- Lawrence Gould A., Boye M. E., Crowther M. J., Ibrahim J. G., Quartey G., Micallef S., and Bois F. Y. Joint modeling of survival and longitudinal non-survival data: current methods and issues. report of the dia bayesian joint modeling working group. *Statistics in Medicine*, 34(14): 2181–2195, 2015. 61
- Lee H. Y., Topham D. J., Park S. Y., Hollenbaugh J., Treanor J., Mosmann T. R., Jin X., Ward B. M., Miao H., Holden-Wiltse J., et al. Simulation and prediction of the adaptive immune response to influenza a virus rnfecion. *Journal of Virology*, 83(14): 7151–7165, 2009. 207

BIBLIOGRAPHY

- Lesaffre E. and Spiessens B. On the effect of the number of quadrature points in a logistic random effects model: an example. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(3): 325–335, 2001. 65, 203
- Li N., Elashoff R. M., Li G., and Saver J. Joint modeling of longitudinal ordinal data and competing risks survival times and analysis of the ninds rt-pa stroke trial. *Statistics in Medicine*, 29(5): 546–557, 2010. 36, 192
- Liang K.-Y. and Zeger S. L. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1): 13–22, 1986. 53
- Liquet B., Timsit J.-F., and Rondeau V. Investigating hospital heterogeneity with a multi-state frailty model: application to nosocomial pneumonia disease in intensive care units. *BMC Medical Research Methodology*, 12(1): 79, 2012. 60
- Little R. J. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421): 125–134, 1993. 46, 61
- Little R. J. and Rubin D. B. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2014. 34, 46
- Liu L., Wolfe R. A., and Huang X. Shared frailty models for recurrent events and a terminal event. *Biometrics*, 60(3): 747–756, 2004. 60
- Marquardt D. W. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2): 431–441, 1963. 199
- Maziarz M., Heagerty P., Cai T., and Zheng Y. On longitudinal prediction with time-to-event outcome: Comparison of modeling options. *Biometrics*, 73(1): 83–93, 2017. 35, 36, 134, 137, 138, 157, 206
- Meira-Machado L., de Uña-Álvarez J., Cadarso-Suárez C., and Andersen P. K. Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research*, 18(2): 195–222, 2009. 53
- Michalski J. M., Winter K., Purdy J. A., Parliament M., Wong H., Perez C. A., Roach M., Bosch W., and Cox J. D. Toxicity after three-dimensional radiotherapy for prostate cancer on RTOG 9406 dose Level V. *International Journal of Radiation Oncology* Biology* Physics*, 62(3): 706–713, 2005. 36
- Michiels B., Molenberghs G., Bijmens L., Vangeneugden T., and Thijs H. Selection models and pattern-mixture models to analyse longitudinal quality of life data subject to drop-out. *Statistics in Medicine*, 21(8): 1023–1041, 2002. 46
- Molenberghs G. and Verbeke G. A review on linear mixed models for longitudinal data, possibly subject to dropout. *Statistical Modelling*, 1(4): 235–269, 2001. 46
- Molenberghs G., Verbeke G., and Demétrio C. G. An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis*, 13(4): 513–531, 2007. 43
- Mottet N., Bellmunt J., Bolla M., Briers E., Cumberbatch M. G., De Santis M., Fossati N., Gross T., Henry A. M., Joniau S., et al. Eau-estro-siog guidelines on prostate cancer. part 1: screening, diagnosis, and local treatment with curative intent. *European Urology*, 71(4): 618–629, 2017. 203
- National Collaborating Centre for Cancer (UK). *Prostate Cancer: diagnosis and treatment*, volume 175. Cardiff (UK): National Collaborating Centre for Cancer (UK), 2014. 28
- Nelson W. Hazard plotting for incomplete failure data(multiply censored data plotting on various type hazard papers for engineering information on time to failure distribution). *Journal of Quality Technology*, 1: 27–52, 1969. 50

- Ng M. K., Van As N., Thomas K., Woode-Amisshah R., Horwich A., Huddart R., Khoo V., Thompson A., Dearnaley D., and Parker C. Prostate-specific antigen (psa) kinetics in untreated, localized prostate cancer: Psa velocity vs psa doubling time. *BJU International*, 103(7): 872–876, 2009. 207
- Nicolaie M., Houwelingen J., Witte T., and Putter H. Dynamic prediction by landmarking in competing risks. *Statistics in Medicine*, 32(12): 2031–2047, 2013a. 208
- Nicolaie M., van Houwelingen J., de Witte T., and Putter H. Dynamic pseudo-observations: A robust approach to dynamic prediction in competing risks. *Biometrics*, 69(4): 1043–1052, 2013b. 36, 137, 143, 206
- Nielsen G. G., Gill R. D., Andersen P. K., and Sørensen T. I. A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, pages 25–43, 1992. 59
- Osoba D., Zee B., Pater J., Warr D., Kaizer L., and Latreille J. Psychometric properties and responsiveness of the eortc quality of life questionnaire (qlq-c30) in patients with breast, ovarian and lung cancer. *Quality of Life Research*, 3(5): 353–364, 1994. 192
- Perelson A. S., Kirschner D. E., and De Boer R. Dynamics of hiv infection of cd4+ t cells. *Mathematical Biosciences*, 114(1): 81–125, 1993. 207
- Philipson P., Diggle P., Sousa I., Kolamunnage-Dona R., Williamson P., and Henderson R. Joiner: Joint modelling of repeated measurements and time-to-event data. Available from: <http://cran.r-project.org/>, 2012. 64, 192
- Pickles T., Kim-Sing C., Morris W. J., Tyldesley S., and Paltiel C. Evaluation of the houston biochemical relapse definition in men treated with prolonged neoadjuvant and adjuvant androgen ablation and assessment of follow-up lead-time bias. *International Journal of Radiation Oncology* Biology* Physics*, 57(1): 11–18, 2003. 36
- Pike F. and Weissfeld L. Joint modeling of censored longitudinal and event time data. *Journal of Applied Statistics*, 40(1): 17–27, 2013. 204
- Prague M., Commenges D., and Thiébaud R. Dynamical models of biomarkers and clinical progression for personalized medicine: The hiv context. *Advanced Drug Delivery Reviews*, 65(7): 954–965, 2013. 207
- Prentice R. L., Kalbfleisch J. D., Peterson Jr A. V., Flournoy N., Farewell V. T., and Breslow N. E. The analysis of failure times in the presence of competing risks. *Biometrics*, pages 541–554, 1978. 53
- Proust-Lima C. and Blanche P. *Dynamic Predictions*. In Wiley StatsRef : Statistics Reference Online. John Wiley & Sons, Ltd, 2015. 31, 35
- Proust-Lima C. and Taylor J. M. 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, 2009. 134, 136, 137, 138, 145, 157
- Proust-Lima C., Taylor J. M., Williams S. G., Ankerst D. P., Liu N., Kestin L. L., Bae K., and Sandler H. M. Determinants of change in prostate-specific antigen over time and its association with recurrence after external beam radiation therapy for prostate cancer in five large cohorts. *International Journal of Radiation Oncology* Biology* Physics*, 72(3): 782–791, 2008. 32, 115, 117, 145
- Proust-Lima C., Séne M., Taylor J. M., and Jacqmin-Gadda H. Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research*, 23(1): 74–90, 2014. 62, 105
- Proust-Lima C., Dartigues J.-F., and Jacqmin-Gadda H. Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: a latent process and latent class approach. *Statistics in Medicine*, 35(3): 382–398, 2016. 36, 106, 192

BIBLIOGRAPHY

- Putter H., Fiocco M., and Geskus R. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11): 2389–2430, 2007. 48
- Putter H. and van Houwelingen H. C. Frailties in multi-state models: Are they identifiable? do we need them? *Statistical Methods in Medical Research*, 24(6): 675–692, 2011. 60, 103
- Rizopoulos D. JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9): 1–33, 2010. 61, 64, 69, 70, 192
- Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3): 819–829, 2011. 36, 134, 137, 139, 157, 206
- Rizopoulos D. Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive gaussian quadrature rule. *Computational Statistics & Data Analysis*, 56(3): 491–501, 2012a. 69, 193, 203
- Rizopoulos D. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC Press, 2012b. 63, 65, 110, 139, 201
- Rizopoulos D. The r package jmbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *Journal of Statistical Software*, 72(7): 1–46, 2016. 64, 192
- Rizopoulos D., Verbeke G., and Molenberghs G. Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics*, 66(1): 20–29, 2010. 122, 207
- Rizopoulos D., Taylor J. M., Van Rosmalen J., Steyerberg E. W., and Takkenberg J. J. Personalized screening intervals for biomarkers using joint models for longitudinal and survival data. *Biostatistics*, 17(1): 149–164, 2015. 32, 136, 208
- Rondeau V., Mathoulin-Pelissier S., Jacqmin-Gadda H., Brouste V., and 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. 60
- Rouanet A., Joly P., Dartigues J.-F., Proust-Lima C., and Jacqmin-Gadda H. Joint latent class model for longitudinal data and interval-censored semi-competing events: Application to dementia. *Biometrics*, 72(4): 1123–1135, 2016. 106, 107, 203
- Rouanet A., Helmer C., Dartigues J.-F., and Jacqmin-Gadda H. Interpretation of mixed models and marginal models with cohort attrition due to death and drop-out. *Statistical Methods in Medical Research* (in press), 2017. 34, 46
- Scheike T. H., Zhang M.-J., and Gerds T. A. Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, 95(1): 205–220, 2008. 52, 134, 143
- Sène M., Bellera C. A., and Proust-Lima C. Shared random-effect models for the joint analysis of longitudinal and time-to-event data: application to the prediction of prostate cancer recurrence. *Journal de la Société Française de Statistique*, 155(1): 134–155, 2014. 117, 203
- Sène M., Taylor J. M., Dignam J. J., Jacqmin-Gadda H., and Proust-Lima C. Individualized dynamic prediction of prostate cancer recurrence with and without the initiation of a second treatment: Development and validation. *Statistical Methods in Medical Research*, 25(6): 2972–2991, 2016. 32, 36, 134, 136, 137, 208
- Stiratelli R., Laird N., and Ware J. H. Random-effects models for serial observations with binary response. *Biometrics*, pages 961–971, 1984. 44
- Struthers C. A. and Kalbfleisch J. D. Misspecified proportional hazard models. *Biometrika*, 73(2): 363–369, 1986. 51
- Suissa S. Immortal time bias in pharmacoepidemiology. *American Journal of Epidemiology*, 167(4): 492–499, 2007. 58

- Suresh K., Taylor J. M., Spratt D. E., Daignault S., and Tsodikov A. Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model. *Biometrical Journal* (in press), 2017. 133, 136
- Sweeting M. J., Barrett J. K., Thompson S. G., and Wood A. M. The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the aric study. *Statistics in Medicine* (in press), 2016. 134, 137, 157, 158
- Sylvestre M.-P., Huszti E., and Hanley J. A. Do oscar winners live longer than less successful peers? a reanalysis of the evidencedo oscar winners live longer than less successful peers? *Annals of Internal Medicine*, 145(5): 361–363, 2006. 58
- Taylor J. M., Park Y., Ankerst D. P., Proust-Lima C., Williams S., Kestin L., Bae K., Pickles T., and Sandler H. Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics*, 69(1): 206–213, 2013. 115, 138, 145, 157
- Temel J. S., Greer J. A., Muzikansky A., Gallagher E. R., Admane S., Jackson V. A., Dahlin C. M., Blinderman C. D., Jacobsen J., Pirl W. F., et al. Early palliative care for patients with metastatic non-small-cell lung cancer. *New England Journal of Medicine*, 363(8): 733–742, 2010. 29
- Tsiatis A., Degruittola V., and Wulfsohn M. Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association*, 90(429): 27–37, 1995. 59, 62
- Tsiatis A. A. and Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3): 809–834, 2004. 35, 61, 62, 106, 136, 192
- van der Vaart A. W. *Asymptotic Statistics*, volume 3. Cambridge University Press, 1998. 63
- van Houwelingen H. C. From model building to validation and back: a plea for robustness. *Statistics in Medicine*, 33(30): 5223–5238, 2014. 58, 133
- van Houwelingen H. C. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1): 70–85, 2007. 36, 59, 136, 206
- Verbeke G. and Molenberghs G. *Linear Mixed Models for Longitudinal Data*. Springer, 2000. 40
- Verbeke G., Fieuws S., Molenberghs G., and Davidian M. The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research*, 23(1): 42–59, 2014. 43
- Vonesh E. F., Greene T., and Schluchter M. D. Shared parameter models for the joint analysis of longitudinal data and event times. *Statistics in Medicine*, 25(1): 143–163, 2006. 63
- Wanneveich M., Jacqmin-Gadda H., Dartigues J.-F., and Joly P. Impact of intervention targeting risk factors on chronic disease burden. *Statistical Methods in Medical Research*, 2016. 208
- Wulfsohn M. S. and Tsiatis A. A. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1): 330–339, 1997. 35, 59, 61, 62
- Ye W., Lin X., and Taylor J. M. Semiparametric modeling of longitudinal measurements and time-to-event data—a two-stage regression calibration approach. *Biometrics*, 64(4): 1238–1246, 2008. 63
- Young Jr J., Roffers S., Ries L., Fritz A., and Hurblut A. *SEER Summary Staging Manual - 2000: Codes and Coding Instructions*. National Cancer Institute, 2001. 27
- Yu M., Taylor J. M. G., and Sandler H. M. Individual prediction in prostate cancer studies using a joint longitudinal survival–cure model. *Journal of the American Statistical Association*, 103(481): 178–187, 2008. 63, 139, 157
- Zhang D., Chen M.-H., Ibrahim J. G., Boye M. E., and Shen W. Jmfit: a sas macro for joint models of longitudinal and survival data. *Journal of Statistical Software*, 71(3), 2016. 64

Modélisation conjointe et prédiction des différents risques de progression de cancer à partir des mesures répétées de biomarqueurs

Dans les études longitudinales en cancer, une problématique majeure est la description de l'évolution de la maladie d'un patient ou la prédiction de son état futur, à partir de mesures répétées d'un marqueur biologique. La modélisation conjointe permet de répondre à ces objectifs, mais elle a principalement été développée pour l'étude simultanée d'un marqueur longitudinal Gaussien et d'un unique temps d'événement. Afin de caractériser les transitions entre événements successifs qu'un patient peut connaître, nous étendons la méthodologie classique en introduisant un modèle conjoint pour un processus longitudinal Gaussien et un processus multi-états Markovien non homogène. Le modèle suppose que les temps de transition individuels sont indépendants conditionnellement aux covariables incluses. Nous proposons aussi un score test afin de tester cette hypothèse. Ces développements sont appliqués à deux cohortes d'hommes avec un cancer de la prostate localisé traité par radiothérapie. Le modèle permet de quantifier l'impact des dynamiques de l'antigène spécifique de la prostate, et d'autres facteurs pronostiques mesurés à la fin du traitement, sur chaque intensité de transition entre états cliniques prédéfinis. Cette thèse fournit ensuite des outils statistiques et des lignes directrices pour le calcul de prédictions dynamiques individuelles d'événements cliniques, dans le cadre de risques compétitifs. Enfin, un dernier travail amène une réflexion sur la modélisation conjointe de données longitudinales ordinales et de données de survie, avec une technique d'inférence innovante. Ainsi, ce travail introduit des méthodes statistiques adaptées à divers types de données longitudinales et d'histoire d'événements, qui permettent de répondre aux besoins des cliniciens. Des recommandations méthodologiques et des outils logiciels sont associés à chaque développement, pour une utilisation pratique par les communautés clinique et statistique.

Mots clés : Biomarqueurs ; Cancer de la prostate ; Données de marqueurs longitudinaux ; Données d'histoire d'événements ; Effets aléatoires partagés ; Modèle conjoint ; Modèle landmark ; Prédiction dynamique individuelle ; Terme de fragilité.

Joint modelling and prediction of several risks of cancer progression from repeated measurements of biomarkers

In longitudinal studies in cancer, a major problem is the description of the patient's disease evolution or the prediction of his future state, based on repeated measurements of a biological marker. Joint modelling enables to meet these objectives but it has mainly been developed for the simultaneous study of a Gaussian longitudinal marker and a single event time. In order to characterize the transitions between successive events that a patient may experience, we extend the classical methodology by introducing a joint model for a Gaussian longitudinal process and a non-homogeneous Markovian multi-state process. The model assumes that individual transition times are independent conditionally to included covariates. We also propose a score test to assess this assumption. These developments are applied on two cohorts of men with localized prostate cancer treated with radiotherapy. The model quantifies the impact of prostate specific antigen dynamics, and other prognostic factors measured at the end of treatment, on each transition intensity between predefined clinical states. This thesis then provides statistical tools and guidelines for the computation of individual dynamic predictions of clinical events in the context of competitive risks. Finally, a last work leads to a reflection on joint modelling of longitudinal ordinal data and survival data with an innovative inference technique. To conclude, this work introduces statistical methods adapted to various types of longitudinal data and event history data, which meet the needs of clinicians. Methodological recommendations and software tools are associated with each development, for practical use by the clinical and statistical communities.

Key words: Biomarkers; Event history data; Frailty term; Individual dynamic prediction; Joint model; Landmark model; Longitudinal marker data; Prostate cancer; Shared random effects.

Spécialité : Santé publique – option : Biostatistique

Laboratoire : Unité INSERM U1219 - Université de Bordeaux - ISPED
146 rue Léo Saignat, 33076 Bordeaux Cedex, FRANCE