



HAL
open science

Joint models for a longitudinal semicontinuous biomarker and a terminal event with application to cancer clinical trials

Denis Rustand

► **To cite this version:**

Denis Rustand. Joint models for a longitudinal semicontinuous biomarker and a terminal event with application to cancer clinical trials. Human health and pathology. Université de Bordeaux, 2020. English. NNT : 2020BORD0252 . tel-03245705

HAL Id: tel-03245705

<https://theses.hal.science/tel-03245705v1>

Submitted on 2 Jun 2021

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

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

THÈSE PRÉSENTÉE
POUR OBTENIR LE GRADE DE
DOCTEUR DE L'UNIVERSITÉ DE BORDEAUX

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

Thèse préparée dans le cadre du Réseau doctoral en santé publique animé par l'EHESP

Par Denis RUSTAND

**MODÈLES CONJOINTS POUR UN BIOMARQUEUR
SEMI-CONTINU ET UN ÉVÈNEMENT TERMINAL AVEC
APPLICATION AUX ESSAIS CLINIQUES EN
CANCÉROLOGIE**

**JOINT MODELS FOR A LONGITUDINAL SEMICONTINUOUS
BIOMARKER AND A TERMINAL EVENT WITH APPLICATION
TO CANCER CLINICAL TRIALS**

Sous la direction de Virginie RONDEAU
Co-directeur: Laurent BRIOLLAIS

Soutenue le 10 décembre 2020

Membres du jury

Mme.	PROUST-LIMA Cécile	Dr, INSERM U1219, Bordeaux	Présidente
M.	SWEETING Michael	Pr, University of Leicester, Leicester	Rapporteur
M.	GUEDJ Jérémie	Dr, Université Paris Diderot, Paris	Rapporteur
Mme.	BELLERA Carine	Dr, INSERM U1219, Bordeaux	Examinatrice
M.	BRIOLLAIS Laurent	Dr, Lunenfeld-Tanenbaum Research Institute Sinai Health System, Toronto	Co-directeur de thèse
Mme	RONDEAU Virginie	Dr, INSERM U1219, Bordeaux	Directrice de thèse
M.	RUE Hâvard	Pr, KAUST University, Thuwal	Invité

Nothing in Biology Makes Sense Except in the Light of Evolution.

Theodosius Dobzhansky

Contents

Contents	4
Acknowledgements	7
Scientific valorisation	9
Résumé substantiel en français	11
1 Introduction	21
1.1 Definition of Cancer	21
1.2 Cancer epidemiology	22
1.2.1 Risks factors	22
1.2.2 Treatment	24
1.2.3 Clinical trials	26
1.2.4 Evaluation of cancer therapeutics	26
1.2.5 RECIST criteria	27
1.2.6 Surrogates endpoints	28
1.3 Datasets availability	29
1.4 Statistical challenges	30
1.5 Thesis structure	30
2 Theoretical background	33
2.1 Analysis of repeated measurements	33
2.1.1 Linear mixed effects model	35
Description	35
Estimation	36
2.1.2 Generalized linear mixed effects model	37
Logarithm link function	38
Logit link function	38
2.1.3 Non-linear mixed effects models	39
2.1.4 Modeling strategies for a semicontinuous outcome	39

	Tobit model	39
	Two-part model	40
2.2	Survival analysis	42
2.2.1	Outcome of interest	42
2.2.2	The proportional hazards model	44
2.2.3	Baseline hazard approximation	45
2.2.4	Time-dependent covariates	45
2.3	Joint modeling for longitudinal data and a terminal event	46
2.3.1	Shared random effects joint model	47
	Association structures	48
	Likelihood expression	48
2.3.2	Extensions of the standard joint model	49
2.4	Computational aspects	49
2.4.1	Frequentist inference	49
2.4.2	Bayesian inference	50
2.4.3	Integrals over the random effects	51
2.4.4	Available programs and packages	51
3	Two-part joint model for a longitudinal semicontinuous outcome and a terminal event with application to metastatic colorectal cancer data	53
3.1	Introduction	53
3.2	Article	54
3.3	Additional remarks	80
3.3.1	On the left-censoring	80
3.3.2	On the numerical integration	80
3.3.3	Erratum in the likelihood function	81
4	A marginal two-part joint model for a longitudinal biomarker and a terminal event with application to head and neck cancer data	83
4.1	Introduction	83
4.2	Article	84
4.3	Additional remarks	126
4.3.1	On the software	126
4.3.2	On the association structure	126
5	Bayesian estimation of two-part joint models with R-INLA	127
5.1	Introduction	127
5.2	Article	128
5.3	Additional remarks	146
5.3.1	On the association structure	146
5.3.2	On the marginal two-part joint model	146
5.3.3	On the prior distributions	146
6	General discussion	147

6.1	Conclusion on the thesis work	147
6.2	Critical insights and perspectives	149
6.3	General conclusion	150
	Bibliography	151
	Appendices	161
	R code for the estimation of the conditional TPJM with frailtypack	161
	R code for the estimation of the marginal TPJM with frailtypack	165
	R code for the estimation of the conditional TPJM with R-INLA	168

Acknowledgements

It would not have been possible to complete this doctoral thesis without the help and support of several people around me. First of all, I would like to thank my thesis director, Dr. Virginie Rondeau, for giving me the opportunity to carry out this doctoral research. You trusted me and gave me the freedom and the tools to make the project my own, your deep insights helped me through all the stages of my research.

I would like to thank my co-director, Dr. Laurent Briollais, who introduced me to research three years ago. You gave me a desire to do this thesis and I always felt at home when visiting your lab in Toronto. Thank you for your availability and all the knowledge and practical advice you shared with me.

I would like to thank the members of the jury, Dr. Jérémie Guedj and Pr. Michael Sweeting for having accepted to review this thesis, your developments in joint modeling are a great inspiration for me. I would like to thank Dr. Carine Bellera and Dr. Cécile Proust-Lima for having accepted to examine this thesis, your contributions to cancer research and survival analysis gave me a solid foundation on which I could build this project. I would also like to thank Pr. Håvard Rue for our promising collaboration and for accepting to be an invited member of this jury. I am honored that you all take part in my jury.

Thanks also go to members of the Biostatistics team of the Bordeaux Population Health Centre in Bordeaux. In particular the director of the team, Dr. Hélène Jacqmin-Gadda for having made me very welcome in the team. It was a pleasure to work in your team. I would also like to thank all my former and present colleagues, who showed me support during these three years, your help in understanding some of the methodological problems is invaluable.

I also acknowledge my gratitude to Pr. Christophe Tournigand, Dr. Sébastien Branchoux and Dr. Janet van Niekerk for an insightful collaboration all along this thesis.

Of course, I am also very grateful to all the funding agencies which supported this thesis, including the EHESP doctoral network, the Académie Française and the University of Toronto.

Finally, I would like to express my gratitude to my family and friends for their continuous support.

Scientific valorisation

Articles

Thesis publications

- [Denis Rustand](#), Laurent Briollais, Christophe Tournigand, Virginie Rondeau. Two-part joint model for a longitudinal semicontinuous marker and a terminal event with application to metastatic colorectal cancer data, *Biostatistics*, 2020, kxaa012.
- [Denis Rustand](#), Laurent Briollais, Virginie Rondeau. A marginal two-part joint model for a longitudinal biomarker and a terminal event with application to advanced head and neck cancers. *Submitted*.
- [Denis Rustand](#), Janet Van Niekerk, Håvard Rue, Christophe Tournigand, Virginie Rondeau, Laurent Briollais. Bayesian Estimation of Two-Part Joint Models for a Longitudinal Semicontinuous Biomarker and a Terminal Event with R-INLA: Interests for Cancer Clinical Trial Evaluation. *Submitted* (arXiv:2010.13704).

Related articles

- Branchoux S, Bellera C, Italiano A, [Rustand D](#), Gaudin AF, Rondeau V. Immune-checkpoint inhibitors and candidate surrogate endpoints for overall survival across tumour types: A systematic literature review. *Crit Rev Oncol Hematol*. 2019;137:35-42.


Communications

Oral presentations at conferences

- [Denis Rustand](#), Laurent Briollais, Virginie Rondeau. The use of joint modeling to analyze tumor dynamics and immunotherapies. *SMAC 2019 scientific days*, January 2019, Bordeaux, France.

- Denis Rustand, Laurent Briollais, Virginie Rondeau. Joint modeling of a semicontinuous longitudinal biomarker and a terminal event. *Survival analysis for junior researchers*, April 2019, Copenhagen, Denmark.
- Denis Rustand, Laurent Briollais, Virginie Rondeau. Joint modeling of a semicontinuous longitudinal biomarker and a terminal event with application to metastatic colorectal cancer data. 11^e *Rencontres scientifiques du Réseau doctoral en santé publique*, March 2020, Marseille, France.
- Denis Rustand, Laurent Briollais, Virginie Rondeau. Joint modeling of a semicontinuous longitudinal biomarker and a terminal event with application to colorectal metastatic cancer data. *The 30th International Biometrics Conference*, August 2020, Seoul, South Korea (Virtual conference).
- Denis Rustand, Laurent Briollais, Virginie Rondeau. Joint modeling of a semicontinuous longitudinal biomarker and a terminal event. *The 41st Annual Conference of the International Society for Clinical Biostatistics*, August 2020, Kraków, Poland (Virtual conference).

R package

Development of the conditional and marginal two-part joint models in the function *longiPenal* of the  package *frailtypack*.

Résumé substantiel en français

Introduction

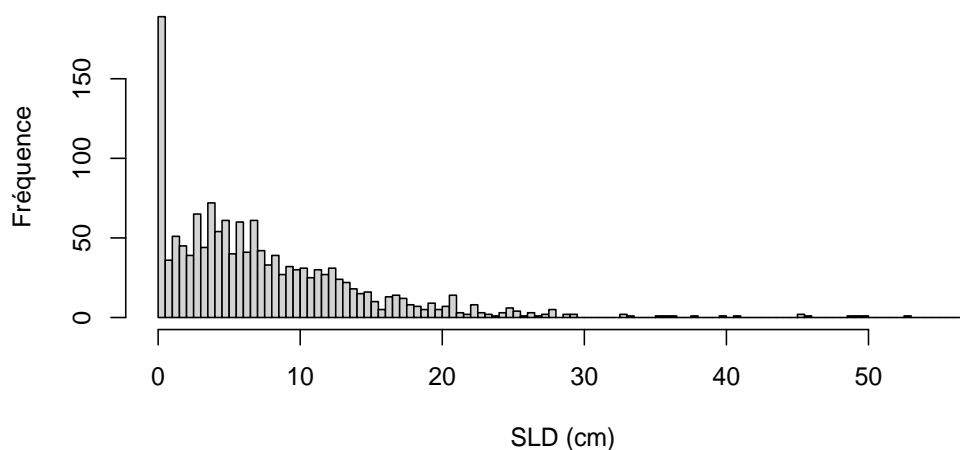
Le mot “Cancer” regroupe plusieurs maladies caractérisées par la prolifération excessive de cellules anormales potentiellement invasives. A la différence de la plupart des maladies, le cancer n’est pas lié à une bactérie ou un virus, il vient de nos propres cellules. Les nouvelles cellules sont généralement produites pour développer de nouveaux tissus ou remplacer des cellules mortes car endomagées ou vieillissantes. Les cellules saines cessent de se diviser lorsque ce n’est pas nécessaire tandis que les cellules cancéreuses prolifèrent de manière incontrôlable. La rapide prolifération des cellules cancéreuses produit des tumeurs, une accumulation de cellules organisées qui agit comme un organe. Les tumeurs sont capables de manipuler les cellules saines environnantes pour accéder au système sanguin en créant des vaisseaux sanguins. Elles sont par conséquent capables de se nourrir et peuvent disséminer des cellules cancéreuses dans d’autres parties du corps, ce processus est appelé angiogénèse. Lorsque les cellules cancéreuses migrent, elles peuvent produire des tumeurs secondaires nommées métastases, qui sont le plus souvent la cause des décès du cancer. Le type de cancer peut être défini par le tissu cellulaire où les premières cellules anormales se développent, avec deux principales catégories. Les cancers à tumeurs solides représentent environ 90% des cancers humains, ils sont caractérisés par une masse locale de tissu cellulaire anormal. Les cancers à tumeur liquide, comme les leucémies ou les lymphomes, correspondent aux cellules anormales présentes dans les fluides (e.g., le sang). D’autres critères permettent de catégoriser les différents types de cancer: leur localisation, le type de cellule affectée ou les caractéristiques génétiques des tumeurs (Lin et al. (2008)). La réponse au traitement peut notamment dépendre de ces profils génétiques (Chan et al. (2019)). De nombreuses maladies différentes avec des caractéristiques variées sont par conséquent classifiées comme cancer (Hanahan and Weinberg (2011)). Les différents types de cancer nécessitent généralement une prise en charge thérapeutique spécifique. De nombreux nouveaux traitements sont candidats pour améliorer la prise en charge thérapeutique des cancers, ces traitements font l’objet d’essais cliniques pour évaluer leur efficacité en comparaison avec l’approche thérapeutique standard (i.e., si le nouveau traitement n’existait pas), généralement en terme de survie. Le critère de référence est en effet la survie globale, qui correspond au temps écoulé entre la randomisation des lignes de traitement et le décès, quelle qu’en soit la cause. C’est une quantité facile à mesurer et précise

mais lorsque les patients ont un bon pronostic (faible risque de décès), un nombre important de patients doivent participer à l'étude pour avoir suffisamment de puissance statistique pour distinguer les lignes de traitement, entraînant une augmentation des coûts des essais cliniques. De plus, à mesure que les traitements gagnent en efficacité, les patients survivent plus longtemps, entraînant une augmentation des coûts de l'essai clinique mais également la prolongation des délais de distribution des nouveaux traitements. Malgré ces limitations, la survie globale reste la mesure la plus pertinente cliniquement pour évaluer les traitements contre le cancer. Dans cette situation, on recherche généralement un substitut à la survie afin de comparer l'efficacité des traitements, cependant en oncologie aucun marqueur n'a été démontré suffisamment corrélé à la survie pour servir de substitut fiable (Prasad et al. (2015)). Toutefois les durées de vie ne sont pas la seule information reflétant l'efficacité d'un traitement dans les données d'essais cliniques en oncologie. En effet, les lésions tumorales représentent un symptôme direct de la maladie et reflètent l'état du patient au cours du suivi de l'essai clinique. L'information sur l'évolution des tailles tumorales est systématiquement récoltée lors d'essais cliniques pour des cancers à tumeurs solides dans le cadre des critères d'évaluation RECIST (Litière et al. (2017)). La somme des plus longs diamètres des lésions cibles (SLD) est un biomarqueur calculé en prenant la somme des plus longs diamètres des principales lésions tumorales du patient à l'initiation de l'essai clinique. Ensuite, les mêmes lésions sont mesurées à chaque visite de suivi du patient jusqu'à la date de point. Ainsi on obtient une représentation longitudinale de l'évolution de la charge tumorale. Ce biomarqueur capture de l'information sur l'efficacité du traitement (réduction de la charge tumorale) ainsi que sur les éventuels effets délétères tels que l'hyperprogression, observée avec de nouvelles classes thérapeutiques telles que les immunothérapies (Champiat et al. (2017)) ou encore le développement d'une résistance au traitement, souvent observé avec des traitements tels que la chimiothérapie (Hansen et al. (2017)). Cette information est généralement utilisée seulement à l'échelle individuelle et résumée par des indicateurs à l'échelle de l'essai, distinguant les patients avec une diminution, un état stable ou une progression des tailles tumorales. Cette approche résume l'information précise récoltée et résulte en une perte d'information. De plus, l'évaluation de l'efficacité d'un traitement repose généralement sur un modèle de survie ne tenant pas compte de l'information sur les mesures de tailles tumorales. Cette méthode ne permet pas d'efficacement prendre en compte l'hétérogénéité individuelle de la population et l'ensemble de l'information issue de l'essai clinique dans l'analyse statistique. De plus, il est primordial d'éviter toute stratégie thérapeutique provoquant une progression de la maladie, et l'analyse des temps de survie passe généralement à côté de cette subtilité, en offrant une vision moyenne de l'effet du traitement, sans tenir compte des variations individuelles.

Dans ce cadre, il semble optimal de prendre en compte les durées de survie ainsi que l'information sur les tailles tumorales lorsque l'on souhaite évaluer l'efficacité d'un traitement ou comparer deux stratégies thérapeutiques. Une méthode efficace pour analyser simultanément des durées de survie et un biomarqueur longitudinal utilise des "modèles conjoints" qui permettent de prendre en compte l'association entre ces deux marqueurs (en effet, l'association entre ces deux marqueurs est forte car la taille des tumeurs affecte directement le risque de décès et le décès implique que l'on n'observera plus de mesures de la SLD). Une subtilité des mesures de tailles tumorales est la présence d'un excès de zéros dû aux patients ayant une disparition des

symptômes de la maladie sous l'effet du traitement.

Distribution de la SLD dans l'essai clinique GERCOR



Histogramme de la distribution de la SLD dans l'essai clinique GERCOR. La distribution est caractérisée par un excès de zéros et une queue de distribution lourde à droite.

Les méthodes de régression couramment utilisées pour modéliser l'évolution d'un marqueur longitudinal supposent une distribution Gaussienne, ignorent l'excès de zéros et l'absence de valeurs négatives, ce qui n'est pas réaliste. Une transformation non-linéaire (e.g., logarithme) permet de contraindre à la positivité des mesures et corrige la queue de distribution mais l'excès de zéros est plus difficile à prendre en compte. Une méthode permettant de prendre en compte cet excès de zéros dans le modèle de régression considère que la mesure des tailles tumorales est sujette à une censure à gauche en raison de la limite de détectabilité du matériel de mesure (on suppose alors que l'on n'observe pas de "vrais zéros", mais des valeurs trop petites pour être mesurées). Cette approche a été proposée pour l'analyse des mesures répétées de la SLD dans de précédents travaux de recherche (Król et al. (2016)). Lorsque de vrais zéros sont observés, les modèles two-part ont été proposés, ils sont particulièrement utiles lorsque les zéros sont informatifs. Dans notre contexte, les zéros correspondent à la disparition complète des lésions tumorales mesurées sous l'effet du traitement et sont par conséquent informatifs vis à vis de l'effet du traitement. C'est pourquoi nous proposons le développement d'une extension des modèles conjoints pour prendre en compte la distribution semi-continue des mesures des tailles tumorales (la notion de semi-continuité implique ici des valeurs non négatives ainsi qu'un excès de zéros). Nous proposons de remplacer le modèle de régression linéaire à effets mixtes généralement utilisé pour décrire les trajectoires individuelles du biomarqueur par un modèle two-part. Le modèle two-part décompose la distribution du biomarqueur en deux parties afin de la décrire sous deux angles complémentaires. La première partie décrit l'effet de variables d'ajustement sur la probabilité d'observer une valeur positive de la SLD (i.e., une valeur non-nulle). La seconde partie peut être spécifiée sous plusieurs formes, la principale étant la forme "conditionnelle" du modèle two-part qui décrit l'effet de variables d'ajustement sur la valeur moyenne du biomarqueur à condition qu'elle soit positive. Ces deux parties utilisent chacune un modèle de régression à effets mixtes,

elles sont liées par la corrélation de leurs effets aléatoires. Les effets aléatoires capturent donc la corrélation entre les mesures répétées d'un patient et la corrélation entre les deux parties du modèle two-part.

La modélisation conjointe consiste à connecter des modèles pour former un modèle plus complexe. Dans ce travail de thèse, nous nous focalisons sur la modélisation conjointe d'un biomarqueur longitudinal et un événement terminal. Des modèles conjoints ont par ailleurs été proposés pour analyser plusieurs biomarqueurs longitudinaux avec un événement terminal ainsi que des événements récurrents (i.e., plusieurs temps d'événements, généralement modélisés par un modèle de survie à fragilité). Les motivations derrière la modélisation conjointe pour un biomarqueur longitudinal et un événement terminal sont multiples:

- L'analyse des mesures répétées du biomarqueur sans tenir compte des décès pourrait être biaisée car le décès représente une forme de censure informative (un patient décédé ne produira plus de mesures du biomarqueur).
- Le modèle de survie peut être ajusté sur des facteurs de confusion mesurés mais ignore les facteurs qui affectent le risque de décès lorsqu'ils ne sont pas mesurés. Les mesures répétées du biomarqueur permettent de capturer l'hétérogénéité individuelle de la population et la modélisation conjointe permet d'ajuster le modèle de survie sur cette hétérogénéité individuelle (qui représente l'effet de facteurs de confusion non mesurés).
- Le fait d'utiliser simultanément l'information sur le biomarqueur et la survie améliore l'estimation de l'effet du traitement (Rizopoulos (2012)).

Les modèles conjoints permettent donc d'étudier: l'évolution d'un biomarqueur longitudinal censuré par l'événement terminal, le risque de décès ajusté sur l'hétérogénéité individuelle de la population capturée par les effets aléatoires, le risque de décès ajusté sur un biomarqueur endogène avec une erreur de mesure (i.e., précision des outils de mesure), explorer l'association entre le biomarqueur et le risque d'événement et prédire le risque de l'événement sachant les mesures répétées du biomarqueur. La forme standard du modèle conjoint utilise des effets aléatoires pour connecter les modèles entre eux, permettant ainsi de tenir compte de la corrélation entre le processus longitudinal et le temps d'événement. Un modèle de régression à effets mixtes est généralement utilisé pour les mesures répétées du biomarqueur, avec souvent une fonction de lien non linéaire permettant de lier le marqueur au prédicteur linéaire en corrigeant une éventuelle queue de distribution lourde et l'hétéroscédasticité des mesures (i.e., pour se ramener à une distribution Gaussienne). Le modèle de survie est un modèle de Cox à hazards proportionnels, avec possiblement une dépendance temporelle des variables d'ajustement. Lorsque seuls les effets aléatoires sont partagés entre le modèle de régression à effets mixtes et le modèle de survie, on parle de l'association "effets aléatoires partagés" et elle n'est pas dépendante du temps. Lorsque l'ensemble du prédicteur linéaire (possiblement retransformé en cas de fonction de lien non linéaire avec le biomarqueur) est partagé, le modèle de survie est ajusté sur la valeur individuelle du biomarqueur estimée par le modèle à effets mixtes. On parle alors d'une association "niveau courant" du biomarqueur, et cette association est dépendante du temps. Cela complexifie le modèle car la densité de probabilité des durées de survie qui doit être calculée dans la fonction de vraisemblance du modèle conjoint nécessite un calcul d'intégrale pour lequel

une solution analytique est uniquement disponible lorsque le modèle de survie est ajusté sur des variables indépendantes du temps. Une approximation numérique de cette intégrale (unidimensionnelle) est donc nécessaire lors du calcul de la vraisemblance des données au modèle avec la structure d'association "niveau courant". La quadrature de Gauss-Kronrod est généralement utilisée pour approximer cette intégrale. De nombreuses fonctions de lien (aussi appelées structures d'association) ont été proposées pour lier le modèle pour le biomarqueur au modèle de survie. Dans cette thèse, nous nous focalisons sur les deux principales formulations que sont les effets aléatoires partagés et le niveau courant du biomarqueur. De plus nous proposons une nouvelle fonction de lien, spécifique au modèle two-part dans notre premier article.

L'expression de la fonction de vraisemblance du modèle conjoint dépend des modèles choisis pour le biomarqueur et la survie. Soit $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{ij})^\top$ le vecteur des mesures répétées du biomarqueur pour l'individu $i = (1, \dots, n)$ aux visites $j = (1, \dots, n_i)$, T_i correspond au temps d'évènement et δ_i l'indicateur de censure associé. La contribution à la vraisemblance pour un individu i peut s'écrire

$$\begin{aligned} L_i(\cdot) &= p(\mathbf{Y}_i, T_i, \delta_i), \\ &= \int_{\mathbf{b}_i} p(\mathbf{Y}_i|\mathbf{b}_i)p(T_i, \delta_i|\mathbf{b}_i)p(\mathbf{b}_i)d\mathbf{b}_i, \end{aligned}$$

où

$$p(\mathbf{Y}_i|\mathbf{b}_i) = \prod_{j=1}^{n_i} p(Y_{ij}|\mathbf{b}_i).$$

Les effets aléatoires corrélés \mathbf{b}_i définissent la relation entre le modèle pour le biomarqueur et le modèle de survie. Les deux modèles sont donc indépendants conditionnellement aux effets aléatoires et la vraisemblance du modèle se calcule par l'intégrale sur la distribution des effets aléatoires du produit des densités de probabilité conditionnelles aux effets aléatoires du modèle pour le biomarqueur et du modèle de survie.

Article 1. Modèle two-part conjoint pour un biomarqueur longitudinal semi-continu et un évènement terminal.

Dans un premier article, nous avons développé le modèle two-part dans le contexte de la modélisation conjointe pour un biomarqueur longitudinal semi-continu et des durées de survie. Ce travail a été motivé par l'essai GERCOR, un essai clinique randomisé de phase 3 comparant deux stratégies de traitement pour des patients atteints d'un cancer colorectal métastatique. L'idée générale de ce travail consiste à proposer un modèle qui utilise directement les mesures de tailles tumorales plutôt qu'un critère de substitution tout en prenant en compte l'excès de zéros dans la distribution des mesures de tailles tumorales. Ces zéros correspondent aux patients pour qui les tumeurs mesurées ont disparu sous l'effet du traitement. Nous avons évalué la relation entre les tailles tumorales et le risque de décès et nous avons comparé le nouveau modèle conjoint two-part avec deux approches alternatives. La première est un modèle conjoint standard qui ignore l'excès de zéros et considère que la distribution du biomarqueur est continue à l'échelle logarithmique. La seconde approche suppose que la mesure du biomarqueur est sujette à une limite de détection et considère les zéros observés comme des valeurs censurées. Avec le nouveau modèle conjoint

two-part, on considère que de vrais zéros (i.e., non censurés) peuvent être observés. C'est une approche qui a l'avantage de décrire l'effet de variables explicatives sur la probabilité d'observer une valeur positive du biomarqueur ainsi que sur la distribution des valeurs positives. Nous avons proposé une formulation générale du modèle conjoint two-part estimé par la méthode du maximum de vraisemblance obtenu par l'algorithme Levenberg-Marquardt. Nous avons adapté les structures d'association "effets aléatoires partagés" et "niveau courant" couramment utilisées pour un modèle conjoint standard au nouveau modèle conjoint two-part. De plus, nous proposons une nouvelle association propre au modèle two-part où l'effet de la probabilité d'observer une valeur positive du biomarqueur et l'effet de l'espérance du biomarqueur à condition d'observer une valeur positive sur le risque de décès sont évalués séparément dans le modèle de survie. La comparaison de l'effet joint des zéros et des valeurs positives du biomarqueur avec leur effet indépendant sur le risque de décès représente un intérêt clinique considérable, on peut en particulier distinguer les patients avec une réponse complète des tumeurs (i.e., $SLD=0$) des patients avec une réponse partielle ou une progression de la maladie (i.e., $SLD>0$).

Dans une étude de simulation, nous avons évalué les performances du modèle estimé en termes de biais et de taux de couverture. Nous avons par ailleurs évalué les conséquences en cas d'erreur de spécification du modèle (i.e., sous l'hypothèse que les données sont générées selon un modèle conjoint standard ou un modèle conjoint avec censure à gauche du biomarqueur). Cette étude de simulation a permis de mettre en évidence que lorsque de vrais zéros sont observés, l'effet du traitement donné par les modèles conjoints standard et avec censure à gauche du biomarqueur peut être biaisé, en particulier si l'effet du traitement est différent entre les deux composantes du modèle two-part. Ce biais dans le modèle pour le biomarqueur se répercute sur le modèle de survie au travers de la fonction de lien et peut ainsi biaiser les conclusions sur l'effet du traitement sur le risque de décès. L'application aux données de l'essai clinique GERCOR était motivée par la comparaison de l'effet du traitement capturé par les deux composantes du biomarqueur (zéros et valeurs positives) ainsi que par la relation de cet effet du traitement avec le risque de décès. Le cancer colorectal fait partie des principales causes de décès du cancer, approximativement la moitié des patients développent des métastases et la chimiothérapie palliative est souvent utilisée pour prolonger la survie. Dans ce contexte, une disparition des tumeurs (i.e., $SLD=0$) n'est pas fréquente et seulement une fraction des patients observeront une telle réponse au traitement. Dans l'étude GERCOR, 12% des mesures répétées de la SLD sont des zéros. Nous avons montré que les deux composantes du modèle two-part (i.e., probabilité d'observer une valeur positive et espérance du biomarqueur à condition d'observer une valeur positive) sont significativement associées au risque de décès. De plus, les modèles conjoints standards et avec censure à gauche du biomarqueur ont une moins bonne discrimination des lignes de traitement vis à vis du risque de décès car ils expliquent moins bien les variations du biomarqueur que le modèle conjoint two-part. Ce nouveau modèle possède cependant des limites, notamment lorsque l'on s'intéresse à l'effet du traitement sur la moyenne (inconditionnelle) du biomarqueur, telle que donnée par les modèles conjoints standard et avec censure à gauche du biomarqueur. Cet article a été publié dans la revue *Biostatistics* (Rustand et al. (2020)) et le modèle conjoint two-part a été implémenté dans la fonction *longiPenal* du package R **frailtypack**. Une interaction entre les langages de programmation Fortran 90 et R permet d'estimer les paramètres de ce modèle statistique.

Article 2. Extension du modèle conjoint conditionnel two-part au modèle marginal two-part avec application au cancer de la tête et du cou.

Le modèle conjoint two-part proposé dans le premier article utilise la formulation conditionnelle du modèle two-part, telle que proposée initialement dans la littérature (Olsen and Schafer (2001); Tooze et al. (2002)). Une formulation alternative a récemment été proposée: le marginal two-part model (Smith et al. (2014)). Une reformulation de la fonction de vraisemblance du modèle permet d'obtenir l'effet de variables d'ajustement sur la moyenne marginale (i.e., inconditionnelle) du biomarqueur dans la partie continue du modèle two-part, à la place de la moyenne à condition d'observer une valeur positive. Le modèle conjoint marginal two-part est par conséquent un mélange entre l'approche par censure à gauche et le modèle conditionnel two-part qui permet de prendre en compte des vrais zéros. Une étude de simulation évalue chaque formulation du modèle conjoint two-part (i.e., conditionnelle et marginale) ainsi que le modèle conjoint standard avec censure à gauche du biomarqueur sous l'hypothèse de chacun de ces modèles pour la génération des données. Elle révèle notamment que la formulation conditionnelle est biaisée lorsque les données sont générées selon la formulation marginale du modèle conjoint two-part. Des problèmes de convergence ont été observés avec la formulation conditionnelle tandis que la formulation marginale est robuste quelque soit le scénario de simulation des données, excepté lorsque le biomarqueur a une trajectoire non-linéaire. Dans cette situation, la formulation conditionnelle est plus adaptée mais des trajectoires plus complexes que la trajectoire log-linéaire (e.g., fonctions paramétriques, splines) peuvent aussi donner plus de flexibilité à la formulation marginale, au prix d'une interprétation plus complexe des paramètres du modèle. De plus, tout comme observé dans notre premier article, le modèle conjoint avec censure à gauche du biomarqueur est biaisé lorsque de vrais zéros sont observés. En outre, nous avons illustré comment le modèle conjoint marginal two-part facilite l'interprétation clinique des résultats de l'essai SPECTRUM, un essai clinique randomisé de phase 3 qui évalue l'efficacité d'un anticorps monoclonal (panitumumab) en complément de la chimiothérapie pour des patients atteints de cancer récurrent ou métastatique de la tête et du cou. Le choix entre la formulation conditionnelle et marginale du modèle conjoint two-part dépend de la question clinique d'intérêt. Lorsque l'on s'intéresse à l'espérance du biomarqueur parmi les valeurs positives, la formulation conditionnelle est préférable tandis que la formulation marginale est plus adaptée lorsque l'on s'intéresse à la moyenne marginale du biomarqueur. Cet article confirme que la prise en compte de vrais zéros est importante et que l'approche par censure à gauche peut être limitée pour évaluer l'effet du traitement sur la moyenne marginale du biomarqueur. Ces zéros correspondent aux patients avec une réponse complète des tumeurs mesurées au traitement et sont par conséquent d'intérêt. La formulation conditionnelle du modèle conjoint two-part peut être instable lorsque la proportion de zéros est faible et nous avons montré que la formulation marginale est plus stable car l'association entre la partie binaire et continue du modèle est prise en compte dans la vraisemblance du modèle en plus de la corrélation des effets aléatoires. Les auteurs de l'étude initiale de ces données d'essai clinique ont conclu à une absence de différence entre les deux lignes de traitement vis à vis de la survie globale. Cependant, le panitumumab est associé à une meilleure survie sans progression (la progression du cancer est définie par les critères RECIST), lorsqu'il est ajouté à la chimiothérapie standard (Vermorken et al. (2013)). Notre nouveau modèle indique un possible effet indirect du

traitement sur la survie, le panitumumab est en effet associé à une probabilité significativement plus élevée d'observer une disparition des tumeurs au cours du suivi de l'étude. Un critère de validation statistique (likelihood cross-validation criterion) montre que la formulation marginale du modèle conjoint two-part offre un meilleur ajustement aux données de cet essai clinique que la formulation conditionnelle.

Article 3. Estimation Bayésienne du modèle conjoint two-part avec INLA: Intérêts pour l'évaluation des essais cliniques sur le cancer

Nous avons ensuite développé la méthode d'estimation Bayésienne du modèle conjoint conditionnel two-part. L'approche fréquentiste consiste à maximiser la vraisemblance des données au modèle proposé, les paramètres estimés représentent ainsi uniquement la distribution des données observées. Cette approche est limitée pour plusieurs raisons. Lorsque la taille d'échantillon est faible, l'estimation des paramètres peut être instable et provoquer des problèmes de convergence de l'algorithme d'optimisation (qui repose sur les dérivées de la fonction de vraisemblance pour faire évoluer les paramètres à chaque itération et approcher du maximum de vraisemblance). De plus, la fonction de vraisemblance contient une intégrale multidimensionnelle correspondant à la distribution des effets aléatoires corrélés. N'ayant pas de solution analytique permettant de calculer cette intégrale, elle doit être approximée numériquement. La méthode utilisée dans le package R **frailtypack** pour un modèle conjoint standard (i.e., un unique modèle de régression pour le biomarqueur) est la quadrature de Gauss-Hermite. Cette méthode s'est toutefois révélée limitée car elle devient très couteuse en temps de calcul à mesure que la dimension de l'intégration augmente, donnant ainsi lieu à des temps de calculs prohibitifs pour le modèle conjoint two-part qui possède deux modèles de régression pour le biomarqueur. Dans ce cadre, nous avons développé pour les deux premiers articles de cette thèse une méthode de Monte-Carlo pour approximer cette intégrale multidimensionnelle dans **frailtypack**, pour laquelle les temps de calculs ne dépendent pas de la dimension de l'intégration.

La méthode Bayésienne repose sur les probabilités conditionnelles, et en particulier le théorème de Bayes pour définir la distribution de probabilité des paramètres du modèle sachant les données observées. Cette distribution est nommée la distribution *a posteriori* des paramètres, elle est obtenue à partir de la fonction de vraisemblance du modèle ainsi qu'une distribution *a priori* des paramètres du modèle. Cette distribution *a priori* reflète les connaissances sur la valeur possible des paramètres avant d'observer les données. L'*a priori* doit être spécifié, il peut être informatif ou non-informatif, par exemple on pourrait avoir un *a priori* informatif sur le paramètre qui reflète la taille tumorale au début de l'essai clinique si les connaissances biologiques le permettent. Le plus souvent, on souhaite laisser parler les données et donner un *a priori* non-informatif, c'est à dire que toutes les valeurs sont plausibles (en pratique on définit généralement une distribution de probabilité avec une variance très large, de manière à ce que la vraie valeur du paramètre soit nécessairement plausible). Le fait de définir un *a priori* non-informatif permet d'obtenir une approximation fidèle du maximum de vraisemblance obtenu par la méthode fréquentiste. L'approche Bayésienne est plus robuste aux problèmes de convergence, en particulier pour les petites tailles d'échantillons car si les données ne sont pas suffisamment informatives, les paramètres du modèle reflèteront l'*a priori*. La méthode couramment utilisée

pour obtenir le maximum *a posteriori* des paramètres du modèle est la méthode de Markov Chain Monte-Carlo (MCMC). C'est une méthode qui repose sur l'échantillonnage et la loi des grands nombres, elle est généralement associée à des temps de calculs conséquents, en particulier pour les modèles simples pour lesquels la méthode fréquentiste est généralement plus rapide. Cependant pour les modèles complexes (nombre de paramètres élevé et/ou dimension des effets aléatoires corrélés élevée), cette méthode est plus efficace que la méthode fréquentiste et permet de développer des modèles plus complexes. Par exemple, des modèles avec plusieurs biomarqueurs longitudinaux ont été proposés avec cette approche (Brown et al. (2005); Rizopoulos and Ghosh (2011)). C'est la raison pour laquelle nous avons souhaité développer l'estimation Bayésienne du modèle conjoint two-part. En effet, le modèle two-part décompose le biomarqueur avec deux modèles de régression, ce qui en pratique est similaire à la modélisation conjointe de deux biomarqueurs longitudinaux en termes de complexité du modèle. Une alternative prometteuse à MCMC est l'algorithme INLA (Integrated Nested Laplace Approximation), implémentée dans le package R **R-INLA**. Cet algorithme permet d'estimer les modèles qui peuvent s'exprimer sous la forme d'un modèle à processus Gauss-Markov latents. Il permet d'obtenir une estimation rapide et précise du maximum *a posteriori* des paramètres et a récemment été introduit pour estimer des modèles conjoints (Van Niekerk et al. (2019)). Nous avons proposé une estimation du modèle conjoint conditionnel two-part avec **R-INLA** et l'avons comparée avec l'estimation fréquentiste initialement proposée dans **frailtypack**. Une étude de simulations démontre une réduction des temps de calcul au-delà de nos attentes et une meilleure précision de l'estimation des effets fixes du modèle avec **R-INLA**. En particulier, les paramètres associés aux effets aléatoires du modèle two-part dans le modèle de survie (qui quantifient l'association entre la survie et le biomarqueur) sont estimés avec une variance nettement réduite avec **R-INLA**, tandis que les paramètres associés au risque de base, à la variance des effets aléatoires et à l'erreur résiduelle de mesure du biomarqueur ont une estimation plus proche de la vraie valeur avec **frailtypack** mais restent dans l'intervalle de crédibilité avec **R-INLA**. Une application aux données de l'essai clinique GERCOR illustre les différences entre les deux méthodes d'estimation et montre notamment que les deux composantes du modèle two-part (zéros et valeurs positives) sont significativement associées au risque de décès avec **R-INLA**, tandis qu'aucune association significative n'est trouvée avec **frailtypack**, à cause de la forte variabilité des paramètres d'association constatée dans les simulations. En outre, **R-INLA** permet d'obtenir l'estimation du modèle en moins d'une minute tandis que **frailtypack** nécessite jusqu'à une heure de calculs (sans supercalculateur). Une seconde application aux données de l'étude PRIME, un essai clinique randomisé de phase 3 pour évaluer l'efficacité de l'addition de panitumumab à la chimiothérapie pour traiter le cancer colorectal métastatique illustre la robustesse de l'estimation Bayésienne aux problèmes de convergence. C'est une étude pour laquelle la présence (ou l'absence) d'une mutation génétique (gène KRAS) a été mesurée pour chaque patient. Cette mutation est connue pour altérer la réponse des patients au traitement (Van Cutsem et al. (2008); Normanno et al. (2009); Bokemeyer et al. (2008)). Il y a un intérêt particulier à distinguer les patients qui bénéficieront du traitement de ceux pour qui il pourrait avoir un effet délétère, en tenant compte de cette mutation génétique. Ainsi, le modèle de régression approprié inclut de nombreuses interactions pour évaluer les différentes sous-populations qui ont reçu ou non le traitement testé et qui ont un gène KRAS

muté ou non. Dans ce cadre, l'estimation fréquentiste proposée dans **frailtypack** conduit à des problèmes de convergence dûs à la complexité du modèle. Nous montrons comment l'estimation Bayésienne proposée dans **R-INLA** permet d'estimer ce modèle complexe et a notamment permis d'identifier une sous-population de patients pour qui l'interaction entre la mutation KRAS et le traitement panitumumab est associée à une décroissance significative de la taille tumorale (comparé aux patients ayant uniquement la mutation ou le traitement), suggérant un possible effet sur la survie. L'estimation Bayésienne permet donc de s'affranchir des limites imposées par l'approche fréquentiste en termes de complexité des modèles (nombre de paramètres, dimension des effets aléatoires) et en termes de temps de calcul. Ce travail reste toutefois limité, en particulier nous n'avons pas proposé l'association "niveau courant (du biomarqueur)", ni la formulation marginale du modèle two-part conjoint car ces modèles ne peuvent pas s'exprimer directement sous la forme d'un processus Gauss-Markov latent et nécessitent des développements supplémentaires pour être estimés avec **R-INLA**.

Conclusion

Dans cette thèse, nous avons étendu les méthodes statistiques disponibles pour l'analyse conjointe d'un biomarqueur longitudinal semi-continu et un évènement terminal en proposant le modèle conjoint two-part. Nos développements ont été motivés par l'évaluation de thérapies anti-tumorales dans le cadre d'essais cliniques sur le cancer pour lesquels la survie et la charge tumorale sont deux mesures d'intérêt. La relation entre ces deux mesures peut être prise en compte par la modélisation conjointe de manière à tirer parti de ces deux sources d'information sur l'effet du traitement simultanément. Cette approche méthodologique permet une meilleure compréhension de la relation entre la réponse des tumeurs au traitement et le risque de décès. Cela contribue ainsi à la recherche clinique en pourvoyant une méthode innovante pour l'évaluation des traitements dans les essais cliniques sur le cancer, s'affranchissant des limites des critères de réponse standards (i.e., RECIST). Ce nouveau modèle statistique est applicable au delà de la recherche sur le cancer car les biomarqueurs longitudinaux semi-continus sont fréquents dans divers domaines de la recherche scientifique (e.g., précipitations quotidiennes, consommations ou dépenses pour des biens, des médicaments ou de la nourriture, données d'expression génétique ou de composition du microbiote).

Chapter 1

Introduction

1.1 Definition of Cancer

The generic term “Cancer” involves a number of diseases defined by the excessive proliferation of abnormal cells and their potential for invasiveness. As opposed to most known diseases, cancer is not a bacteria or a virus, it originates from inside our own cells. New cells are usually produced to build new tissue or to replace cells that have died because of aging or damages. Healthy cells stop dividing when this is not required while cancerous cells proliferate uncontrollably. The rapid proliferation of cancerous cells produces tumors, an accumulation of organized cells acting like an organ. Tumors are able to manipulate surrounding healthy cells to get access to the blood system by creating blood vessels. It is therefore able to feed itself and can spread cancerous cells in other parts of the body, this process is referred to as angiogenesis. When cancerous cells migrate, they can produce secondary tumors named metastases which are most often the cause of death from cancer. In humans, the type of cancer can be defined by the cell tissue where the primary abnormal cells develop, with two main categories:

- Solid tumor cancers represent about 90% of human cancers (e.g., carcinomas, sarcomas), they are characterized by a local abnormal mass of tissue.
- Liquid tumor cancers, such as leukemias or lymphomas, correspond to abnormal cells present in body fluids (e.g., blood, bone marrow).

In this thesis work, we focus on the measurements of the size of solid tumors defined as a biomarker. The different types of cancer are also characterized by their location (e.g., breast, lung) and cell type (squamous, myeloid, lymphoid, adenomatous). Moreover, cancers can also be categorized according to some genetic features (Lin et al. (2008)), which can modulate the response to treatment (Chan et al. (2019)). Many different diseases with various characteristics are classified as cancer (Hanahan and Weinberg (2011)).

The cell proliferation and apoptosis are well-oiled mechanisms, especially due to the intervention of specific proteins. The alteration of genes involved in these mechanisms can allow cells to proliferate out of control or suppress apoptosis. Beyond genetic mutations, epigenetic mechanisms can modify the genes expression without altering the genetic sequence. Some cancers have only epigenetic mutations and no genetic mutations but in most cases, the genetic epimutations appear in addition to the genetic mutations and participate to the extraordinary diversity of cancerous cells in a tumor. Mutations are essential to produce genetic variations that fuels natural selection. A breed that does not produce mutations is condemned to extinction because it is unable to adapt to constant changes in the environment. A mutation can result from a DNA replication error or environmental factors (exposition to mutagens) and does not always imply a risk of cancer. In this context, it is difficult to define when a cancer begins since mutations and pre-cancerous lesions (genetic changes associated to an increased risk of cancer than can change the function of the cells but not enough to cause a cancerous behaviour) are usually contained by the environment long enough to not result in the development of invasive metastases. A high number of microscopic clinically unapparent tumors were found in several autopsy studies. For instance, a study of 110 medicolegal autopsies from young and middle-aged women between 20 and 54 years old, of which only one received breast cancer diagnosis, found cancerous cells in 22 of them (20%), with a significant effect of age (Nielsen et al. (1987)). Another study of 152 prostate glands from male patients between 10 and 49 years old identified microtumors in 0%, 9%, 20% and 44% in the second, third, fourth and fifth decade of age (Sakr et al. (1993)). Despite this proportion of individuals with microtumors, the lifetime diagnosis for breast or prostate cancer is below 2% (Kareva (2018)). As explained by Stephen C. Stearns, professor of evolutionary medicine at Yale University, we all have thousands of pre-cancerous mutant lesions and we would probably all die from cancer if we were to live long enough (Stearns and Medzhitov (2015)).

1.2 Cancer epidemiology

Cancer is an increasing source of mortality, especially in developed countries. It is estimated that more than 18 million new cases of cancer were diagnosed and more than 9 million death in 2018 (Bray et al. (2018)). There is a great variability of the risk of developing or dying from cancer according to the type of cancer, genotype and phenotype of the individuals, environmental exposures and geographic area.

1.2.1 Risks factors

Risk factors associated with cancer can be identified through epidemiological studies where individuals who developed cancer are compared to those who did not. These studies entail behaviors, exposure to substances and characteristics associated with an increased risk of cancer. However, these studies often induce some uncertainty and cannot prove on their own that the observed increased risk is due to the risk factor and not the result of chance or something else than the suspected risk factor. Many studies are required to confirm a similar association between the risk factor and the risk of developing cancer and in addition, a plausible mechanism is needed

to explain how the risk factor causes cancer to conclude on a causal association. The most well-known risks factors according to the U.S. National Cancer Institute (www.cancer.gov) are:

- Age (advancing age is the most important risk factor for cancer, overall).
- Tobacco use along with environmental tobacco smoke cause many types of cancer.
- Alcohol consumption increases the risk of cancer and it is much higher for those who both drink alcohol and use tobacco.
- Sunlight exposure and sunlamps cause early aging of the skin that can lead to skin cancer due to ultraviolet radiation.
- Obesity is associated with an increased risk of several types of cancer.
- Ionizing radiation can damage DNA and cause cancer.
- Exposure to carcinogens (e.g., heavy metals, second hand tobacco smoke).
- Chronic inflammation can cause DNA damage and lead to cancer.
- Diet: there is some uncertainty but specific dietary components could be associated with an increased (or decreased) risk of developing cancer.
- Hormones such as estrogens are human carcinogens and can increase a woman's risk of cancer.
- Immunosuppression provoked by immuno-suppressive drugs or infection to HIV can cause cancer.
- Infectious agents: Some viruses, bacteria and parasites can cause cancer or increase the risk of developing cancer.

Some risk factors, such as aging, cannot be controlled and others involving behavior or exposure could potentially be controlled to reduce the incidence of cancers, although most cancer cases cannot be avoided just by controlling the known risk factors. Several studies showed that around 30-40% of all cancer cases are attributable to potentially modifiable risk factors in the United kingdom (Brown et al. (2018)), Australia (Whiteman et al. (2015)) or the United States (Islami et al. (2018)).

An analysis of the role of hereditary factors on the risk of developing cancer, based on 44788 pairs of twins, found an effect of heritability on the risk of developing certain types of cancer. Heritable factors were estimated to account for 42% of prostate cancer risk, 27% of breast cancer risk and 35% of colon cancer risk (Lichtenstein et al. (2000)). Recent advances in the past 20 years have made possible the sequencing of the human genome at a high depth, the use of common genetic polymorphisms to characterize common human diseases (catalogued in the International HapMap Project, see Gibbs et al. (2003)), the application of high-throughput genotyping of millions of polymorphisms simultaneously and the development of new statistical methods to interrogate the massive amounts of data generated (e.g., genome-wide association studies). These advances have led to the identification of novel cancer causing variants (e.g.,

Al-Tassan et al. (2015)). It is now an important focus of cancer research to identify these genetic risk factors (Pomerantz and Freedman (2011)).

1.2.2 Treatment

The earlier the cancer is diagnosed and treated, the better the prognosis. When the primary tumor does not develop into metastases, the tumor is resected whenever possible and the cure rate remains high. When metastases have developed, the treatment is more challenging as the cancerous cells have circulated into the body and are not localized anymore. The treatment option then depends on multiple factors and the main types of treatments are, according to the U.S. National Cancer Institute (www.cancer.gov):

- **Surgery:** Widely used to resect primary tumors, surgery attempts to remove the entire tumor and can be followed by other treatments to avoid resurgence of cancer. Sometimes removing an entire tumor might damage an organ or the body and only part of the cancer tumor is removed. Finally, surgery can ease cancer symptoms by removing tumors that are causing pain or pressure.
- **Radiation therapy:** High doses of radiations are used to damage the DNA of cancer cells, these cells are not killed right away but when sufficiently damaged, they cannot spread anymore and are removed by the body. It is used to treat cancer (curative) or to ease cancer symptoms (palliative). It can be combined with surgery either to shrink the tumor before resection, to target cancer cells during surgery (avoids radiations passing through the skin), or to kill the remaining cancer cells after surgery. The radiations also affect nearby healthy cells that can cause serious side effects.
- **Chemotherapy:** It targets cancer cells that grow and divide quickly with chemical drugs to stop or slow their growth. As with radiotherapy, it is used for both curative and palliative cares. It can be used before surgery or radiotherapy to reduce the tumor size (referred to as “neoadjuvant chemotherapy”) or after, to destroy remaining cancer cells (referred to as “adjuvant chemotherapy”). It can also improve the efficacy of other treatments and can be used for non-local cancers, as it targets all fast growing cells in the body. Cancer cells are not the only fast growing cells, healthy cells in the mouth, intestines or those responsible of growing hair and nails are also damaged by chemotherapy, causing side effects. However, these side effects usually disappear once the treatment is over. Recent work has shown that it is sometimes better to maintain the tumor size instead of trying to remove it entirely because some cancer cells can develop some resistance to chemotherapy. When all the other cancer cells are removed by chemotherapy, only the resistant cells proliferate and the treatment can be ineffective in contrast to when the tumor is contained. Resistant cells are in competition with non-resistant cells and survival can be improved, although there are situations where the containment will make a bad prognosis even worse (Hansen et al. (2017)).
- **Hormone therapy:** It is a more specific treatment that slows or stops cancers that use hormones to grow, such as prostate or breast cancers. It can also be combined with other

treatments and cause side effects due to the interference with the hormones' behavior.

- Stem cells transplant: It is used after a first line treatment (e.g., radiation therapy, chemotherapy) in order to restore blood-forming stem cells destroyed by these treatments.
- Targeted therapy: This type of treatment targets proteins related to cancer cell dynamics (growth, division and spread). They are either molecules small enough to enter cells or monoclonal antibodies, a type of antibodies designed to attach to specific targets found on cancer cells. Most of the time, a biopsy is required because the treatment efficacy is subject-specific.
- Immunotherapy: This class of treatments helps the immune system fighting against cancer. The immune system usually detects and destroys abnormal cells but cancer cells often bypass this destruction either because of genetic changes that prevent the immune system to detect them or because of specific proteins on their surface that turn off the immune system or interfere with surrounding healthy cells to block the immune system response. Among immunotherapies, the Immune Checkpoint Inhibitors (ICI) have received a lot of attention recently (Pardoll (2012)). Several clinical trials suggest that blockade of checkpoint inhibitors (e.g., PD1 pathway) induces sustained tumour regression in various tumour types (Hargadon et al. (2018)). However, this new class of treatments was shown to induce some new patterns of responses such as the pseudoprogression where the tumor burden or number of tumor lesions increases initially before decreasing, or the hyperprogression, which is a phenomenon reflecting a very rapid tumor progression following immunotherapy and suggesting that ICI could impact detrimentally on a small subset of patients (Borcoman et al. (2019); Wang et al. (2018); Kamada et al. (2019); Fuentes-Antrás et al. (2018)).

Next generation sequencing technology has made possible the identification of somatic mutation in the tumor that can modify a patient's response to treatment (Dancey et al. (2012); Xing et al. (2011)). For example the KRAS gene mutation is predictive of a poor response to EGFR inhibiting drugs (e.g., panitumumab and cetuximab) in colorectal cancer (Lievre et al. (2006)). Such mutational features are particularly of interest for ICIs because this class of treatments directly targets gene expression (e.g., PD-L1). Several potential biomarkers based on DNA, RNA or protein features have been proposed to predict the response to an ICI (e.g., microsatellite instability, mismatch repair deficiency, $IFN\gamma$ expression), see Galluzzi et al. (2018). Because cancer cells are mutated cells, their gene sequences differ from healthy cells. It is therefore of interest to focus on the genetic of the tumor but in a single tumor, the mutations accumulate as the tumor grows which makes difficult the analysis of all these mutations that are accumulating over time. The tumor mutational burden (TMB) is a measure of the total amount of somatic coding mutations in a tumor. It has been proposed to distinguish hypo-mutated tumors from hyper-mutated tumors. Hyper-mutated tumors are more likely than hypo-mutated tumors to generate tumor-specific peptides (neoantigens) recognized by the immune system when treated with an ICI (Strickland et al. (2016)).

1.2.3 Clinical trials

To propose a new treatment or a new combination of treatments, a series of steps (called phases) are required, each of them requiring independent clinical trials for their evaluation. The first phase (phase I) aims to verify the drug safety, the appropriate dose and look for side effects. It is sometimes preceded by a phase 0 trial, which is a very small trial that help researchers decide whether a new agent should be tested in a phase I trial or not. Phase I involves only around 15 to 30 patients. A new treatment must be successful in each phase in order to proceed to the next one. In the second phase (phase II), the drug is tested on a larger number of patients (e.g., up to 100), with a focus on treatment effect (usually on the tumor dynamics). Then in phase III, the treatment is compared to a standard therapy in terms of efficacy and safety, it includes a large number of patients to make sure that the result is valid (e.g., from 100 to several thousands). Sometimes, a later phase (phase IV) is conducted after giving the drug a license for further evaluation of the effectiveness and safety. Our methodological developments focus on trials that aims at comparing a new treatment to the standard of care and concern mostly phase III cancer clinical trials.

1.2.4 Evaluation of cancer therapeutics

The evaluation of new therapeutics requires criteria in order to compare them to a placebo or standard of care (Fiteni et al. (2014)). The gold standard is the overall survival (OS), corresponding to the time from randomization until death from any cause. It is precise and easy to measure but has several limitations:

- In case the enrolled patients have a good prognosis (low death rate), a large sample size is required to get a sufficient power to distinguish the treatment lines, thus increasing the cost of the trial.
- As new treatments gain efficacy, patients survive longer and the follow-up required to evaluate treatments gets longer too, thus inducing an increased cost of the trial as well as an increased delay in the availability of the new treatment when it is better than the standard of care.

Despite these limitations, OS remains the most clinically relevant endpoint in cancer clinical trials. In the context of cancer clinical trials, the disease manifests itself through tumors, which are measurable entities that can reflect the disease severity and its evolution over time. Tumor-centered clinical surrogates for overall survival are increasingly studied. Ideally, the three-dimensional volume of each tumor should be considered but it requires highly sophisticated equipment to be measured and there could be too many tumors to account for all of them. In the 1980s, the World Health Organization proposed a set of criteria to report and categorize the tumor response in cancer clinical trials (Miller et al. (1981)). The main idea was to have a “common language”, able to describe and compare the results of cancer treatments. It became the standard method for evaluating the tumor response. However, these criteria have been criticized because of inconsistent methods of measurement including errors in tumor measurements, errors in selection of measurable target lesions, intercurrent diseases and radiologic technical problems

(Thiesse et al. (1997)). To remedy these problems, a new set of criteria were introduced in 2001: the Response Evaluation Criteria in Solid Tumors (RECIST).

1.2.5 RECIST criteria

The RECIST criteria have been developed to evaluate the tumor response to treatment in solid tumor cancer clinical trials. They are comparable to the WHO criteria but have more simple and reproducible guidelines (Choi et al. (2005)). A major update of these guidelines has been proposed in 2009 (Eisenhauer et al. (2009)) and several other updates have been proposed recently because of the non usual response observed for the new classes of treatments. In practice, tumors are classified as “target” and “non-target” lesions at baseline. Only the target lesions are measured at each follow-up visit of the patient while other non-target lesions are only evaluated qualitatively. Based on the RECIST criteria, the number of target lesions is limited, they are chosen based on their size (lesions with the longest diameter) and their suitability for accurate repeated measurements (either by imaging techniques or clinically). A biomarker that reflects the tumor burden and its evolution over time is the sum of the longest diameter of the target lesions (SLD). The RECIST criteria relies on categorization of the response of the target lesions:

- Complete Response (CR): Disappearance of all target lesions. Based on the criteria, it can be required that the disappearance persists for a certain duration (e.g., 1 month).
- Partial Response (PR): At least a 30% decrease in the SLD, taking as reference the baseline SLD.
- Progressive Disease (PD): At least a 20% increase in the SLD, taking as reference the smallest SLD recorded since the treatment started.
- Stable Disease (SD): Neither sufficient decrease to qualify for PR nor sufficient increase to qualify for PD.

Similarly, the response of other non-target lesions are also classified:

- Complete Response: Disappearance of all non-target lesions.
- Progressive disease: Appearance of one or more new lesions and/or unequivocal progression of existing non-target lesions (since non-target lesions are not measured, progression can be difficult to assess and usually requires confirmation)
- Stable Disease: Persistence of one or more non-target lesions.

The overall response is defined based on these categories, giving a global summary of the response to treatment:

New patterns of responses were recently reported for the immunotherapies, for example the “pseudoprogression” or “flare effect”, where the tumor burden grows for a limited time (usually for a few weeks, sometimes up to a couple of months) and then shrinks during a successive phase, resulting in a good efficacy of the treatment. The initial progression was considered as a failure of treatment according to RECIST criteria, therefore requiring the criteria to be updated

Table 1.1: Evaluation of overall response

Target lesions	Non-Target Lesions	New Lesions	Overall response
CR	CR	No	CR
CR	SD	No	PR
PR	Non-PD	No	PR
SD	Non-PD	No	SD
PD	Any	Yes or No	PD
Any	PD	Yes or No	PD
Any	Any	Yes	PD

in order to better capture this delayed treatment effect (Tazdait et al. (2018)). The RECIST working group (which comprises representatives of the European Organization for Research and Treatment of Cancer (EORTC), the National Cancer Institute (NCI) of the United States and the Canadian Cancer Trials Group (CCTG), as well as several pharmaceutical companies) published a proposition for a new criteria called iRECIST, with the goal to standardise the response assessment in immunotherapy clinical trials. With this version of RECIST criteria, stable disease (SD), partial response (PR) and complete response (CR) remain identical but the definition of a progressive disease (PD) changes. A confirmation of a progression at a second visit later in the follow-up (4-8 weeks) is required to confirm an observed progression. Moreover, death or immunotherapy discontinuation due to clinical progression is considered as confirmation of progression (Seymour et al. (2017)).

The RECIST criteria provide a simple and pragmatic methodology to evaluate the activity and efficacy of new cancer therapeutics in solid tumors, using validated and consistent criteria to assess changes in tumor burden. Moreover, the standardization of tumor response facilitates the definition of tumor-based candidate surrogate endpoints. However, these candidate surrogate endpoints are often limited due to the loss of information driven by the categorization of the response. The great variability of cancer cells makes each tumor unique and individual responses are difficult to categorize (Fisher et al. (2013)). These criteria are widely used to categorize tumor response in chemotherapy trials but raise some questions for new treatment evaluation such as immunotherapies because of the large inter-individual variability of the response.

1.2.6 Surrogates endpoints

A surrogate endpoint is a biomarker able to predict clinical benefits, harm, or lack of those in clinical trials (De Gruttola et al. (2001); Prentice (1989)). The research of alternative endpoints that can predict the response to treatment is an important focus of cancer research. They are usually evaluated in terms of predictive performance of the overall survival but the validation of a surrogate endpoint requires an appropriate methodology (Sofeu et al. (2019, 2020)). However, the evidence supporting the use of surrogate endpoints in oncology is limited (Prasad et al. (2015)). Other endpoints are sometimes of interest (and replace the OS as the primary endpoint), such as the quality of life measured by questionnaires. This latter allows for patient-centred clinical decision making but this is out of the scope of this thesis. Several surrogate endpoints for cancer clinical trials have been proposed in the literature and are used in clinical trials despite the low evidence supporting their use (Kemp and Prasad (2017); Piedbois and Croswell (2008)).

For example, an examination of 54 marketing approvals of new cancer drugs by the Food and Drug Administration between January 1, 2008 and December 31, 2012 shows that 36 (57%) were approved on the basis of a surrogate endpoint. Out of these 36 drugs' approval, 19 were based on a reduction in tumor size or in volume and 17 were based on a progression-free survival (progression being defined by the RECIST criteria). An analysis of the postmarketing studies shows that with several additional years of follow-up, 31 of these 36 approvals based on surrogate endpoints have in fact unknown effects on the overall survival or fail to show any gain in survival (Kim and Prasad (2015)). It is even more challenging to use surrogates for the new class of therapeutics (e.g., ICIs), I was involved in a literature review of the studies assessing the surrogacy of candidate endpoints for ICI and we found no evidence for a surrogate endpoint for overall survival (Branchoux et al. (2019)).

The main drawback for these candidate surrogates is that the categorization of individual responses results in a loss of information (individual baseline value of the biomarker and evolution over time). Clinical trials are very expensive and methodological developments should take advantage of all the available information to answer the clinical question of interest. The continuous longitudinal process of the tumor burden is of clinical interest because it captures information about the primary target of most cancer treatments. The development of new statistical methodologies makes possible the joint analysis of multiple endpoints, such as survival times and the repeated measurements of a biomarker.

1.3 Datasets availability

Clinical trials are usually funded by the pharmaceutical industry to provide evidence of the effectiveness of the new drugs they develop. Because they have financial interest in the outcome of clinical trials, they must follow strict clinical practice guidelines and publish their results in peer-reviewed journals (Chopra (2003)). The data from cancer clinical trials is rarely shared with the world's research community because of its ownership by the industry and because it usually contains very detailed information on each participant (Tucker et al. (2016)). However, data anonymisation is sometimes used to share the data without privacy risk. In this context, we got access to the results of the GERCOR clinical trial to apply the new statistical model that we developed (two-part joint model) and illustrate its relevance. We were initially supposed to work with the results of an innovative immunotherapy trial, for which our new model is particularly of interest as discussed in this thesis. However the pharmaceutical company that conducts this trial required dissuasive conditions such as the right to deny the publication of our work. We were also supposed to analyze the ARCAD database (Franko et al. (2016)), a large database composed of multiple clinical trials data for colorectal cancer which would have been interesting in particular to evaluate the relative performances of **frailtypack** and **R-INLA** for large sample sizes, with a focus on the computation time (see Chapter 5). Similarly, the conditions required to get access to this database were dissuasive. In this context, we illustrated our new methods with data from the project data sphere platform (www.projectdatasphere.org) which is a nonprofit, open-access cancer research platform for data sharing.

1.4 Statistical challenges

Several statistical challenges arise when analyzing cancer clinical trial data. We are interested in the analysis of the repeated measurements of the SLD in addition to the standard analysis of survival times. These two outcomes are linked because the measurement of the SLD is censored by death (i.e., informative censoring that can bias the analysis of the SLD). For example, a patient could have a rapid tumor progression that provokes his death before he/she could have any visit, therefore the tumor progression does not appear in the measurements. The SLD value is only recorded at each visit of the patient, an appropriate regression model must be defined to make inference about the longitudinal trajectory of the biomarker. The joint analysis of survival times and a longitudinal biomarker is useful in the context of complex cancer dynamics to explore the relationship between the evolution of the biomarker and the time to a clinical event. In the context of cancer clinical trials, the SLD is a biomarker that directly captures information about the disease at baseline and its evolution over time. Moreover, it captures the individual heterogeneity in the patient population. Several applications have been proposed to use the biomarker's information to predict the occurrence of the event using joint modeling techniques (Król et al. (2016, 2018)). However, in this thesis, we are interested in the joint modeling of the SLD with survival time to evaluate new therapeutics, thus an inferential framework. Flexible models are used to produce accurate predictions, there is no need of an explicit interpretation of the model as long as it performs well in terms of prediction. When evaluating new treatments, the model is developed in order to have a clear interpretation of the differences between treatment arms. The measurements of the SLD is a mixture of discrete and continuous measures because some patients will have a complete shrinkage of their SLD upon treatment effect while other patients will have positive continuous measurements. Such distribution requires an appropriate methodology to be fitted, in particular because we are interested in the characterization of patients with a complete disappearance of their tumors after treatment (corresponding to CR of target lesions according to RECIST criteria) along with those with a partial response or a stable disease (PR, SD). The characterization of patients with progressive disease (PD) is also of interest when progression or hyperprogression of the tumors is observed. This is particularly challenging from a statistical point of view because we analyze multiple outcomes, which require complex models. The recent development of joint modeling in biomedical research has rapidly reached the limit of the available algorithms in terms of computational burden and model complexity. There is a great need for developing further statistical methods focusing on the joint analysis of longitudinal tumor size and death and for providing efficient tools for the analysis of data from cancer clinical trials.

1.5 Thesis structure

Our main objective in this thesis is to develop a general methodology to analyze jointly the longitudinal measurements of the SLD and the survival times accounting for the excess of zeros of the SLD. The longitudinal analysis of the SLD overcomes the limitation of the RECIST criteria, which categorize the tumor response. The particular distribution of the measurements of the SLD (i.e., mixture of zeros and positive values) is of particular interest. Previous analyses

of the SLD used censoring techniques to account for the excess of zeros in the distribution of this biomarker, therefore assuming zeros are not true zeros but values below a limit of detection (i.e., too small to be observed). It has the advantage of providing an effect of covariates on the mean biomarker value and its evolution over time. However, it is limited when true zeros are observed. Depending on the proportion of zeros, a regression model for a binary outcome (i.e., zeros versus positives) could be more appropriate than assuming a continuous outcome. In the context of cancer clinical trials, there is an interest in the relationship between the occurrence of zero values and predictor variables, as long as the relationship between the occurrence of zeros, the distribution of positive values and the event of interest.

Firstly, we developed a new methodology for the joint analysis of the SLD and survival times (Chapter 3). We considered a conditional two-part regression model for the longitudinal biomarker which splits its distribution into a binary outcome (first part) represented by the positive versus zero values and a continuous outcome (second part) with the positive values only. A logistic mixed effects model is proposed to model the effect of covariates on the probability of a zero SLD value while a linear mixed effects model gives the effect of covariates on the log-transformed values of the positive measurements of the SLD. Survival times are modeled with a proportional hazards model for which we proposed three association structures with the biomarker. The conditional two-part joint model evaluates the effect of covariates on the probability of positive value of the biomarker, the expected value conditional on a positive value and the risk of death. We showed through simulation studies that assuming the true model is a two-part model, bias can arise in the evaluation of a treatment with standard methods. An application to advanced metastatic colorectal cancer data from the GERCOR study is performed where our new model finds a significant effect of treatment on the SLD values. We showed how the different association structures of the two-part joint model allows for an evaluation of this effect in terms of risk of death compared to the reference treatment.

Secondly, we proposed an alternative formulation of the two-part joint model in order to get the effect of covariates on the marginal mean of the biomarker. Indeed, a drawback of the two-part model is that by decomposing the distribution of the biomarker, the effect of treatment on the probability of positive value can be opposite to the effect of treatment among positive values, therefore making difficult clinical decision-making about the treatment efficacy. The marginal two-part joint model is an alternative formulation of the (conditional) two-part joint model, recently introduced in the literature to get a marginal effect of a covariate on a semicontinuous outcome. The probability of positive value is taken into account in the continuous part of the model in order to remove the condition on a positive value of the standard (i.e., conditional) two-part model. This alternative formulation is useful to facilitate the interpretation of covariates effect (e.g., treatment) on the mean value of the biomarker. We showed how both the conditional and the marginal formulations of the two-part joint model answers different clinical questions of interest. A simulation study assessed the good performance of the marginal two-part joint model in terms of estimation and coverage rates and how the variability of the mean biomarker value is reduced with the marginal model compared to the conditional formulation. An application to a randomized clinical trial of advanced head and neck cancer shows an effect of treatment on the odds of observing a disappearance of all target lesions, leading to a possible indirect effect of the

combined treatment on time to death.

Finally, we extended the modeling strategy initially proposed in a frequentist framework to the Bayesian framework. This work was motivated by the limitations in terms of model complexity often encountered within the frequentist framework. Indeed, the two-part joint model involves 3 regression models (binary, continuous and survival part), the model complexity is increased compared to standard joint models (i.e., using a single regression model for the biomarker), it can be challenging for complex models (i.e., large number of parameters and dimension of the random effects). We propose a Bayesian estimation of two-part joint models based on the Integrated Nested Laplace Approximation (INLA) algorithm to alleviate the computational burden and be able to fit more complex models. Our simulation studies show that the Bayesian estimations are associated to substantially reduced computation time and variability of the estimates, and improves the model convergence compared to the initially proposed frequentist estimation. We contrast the Bayesian and frequentist approaches in two randomized cancer clinical trials (GERCOR and PRIME studies), where INLA suggests a stronger association between the biomarker and the risk of event and was able to characterize subgroups of patients associated with different responses to treatment in the PRIME study where the frequentist approach had convergence issues. Our study suggests that the Bayesian approach using INLA algorithm enables broader applications of the two-part joint model to clinical applications.

In the following of this thesis, these three parts are presented with related articles. We give a theoretical background of the statistical methods required for the understanding of the rest of the manuscript in Chapter 2. This work is concluded by a general discussion with further perspectives in Chapter 6.

Chapter 2

Theoretical background

In this chapter, we provide some statistical background, necessary for the understanding of the thesis, with a focus on the analysis of the outcomes observed in cancer clinical trials.

2.1 Analysis of repeated measurements

Repeated measurements occur when there is a longitudinal follow-up of patients with repeated visits to measure the same marker. We therefore get multiple values of a marker associated with each individual at different time points. It is common in clinical trials to have repeated measurements.

Several types of longitudinal repeated measurements exist:

- Continuous values: The measure can take any value.
- Semicontinuous values: The measure can take any value in a half-bounded interval.
- Binary values: The measure is either a 0 or a 1.
- Counts: The measure is a positive integer.

Our biomarker of interest, the SLD, has a non-negative semicontinuous distribution, which is a mixture of a binary outcome ($SLD > 0$ vs. $SLD = 0$) and a continuous outcome (distribution of $SLD > 0$). There are several models to analyze count data but they are not discussed in this manuscript, as they are out of the scope of our application of the joint analysis of the SLD and the survival time. In order to analyze longitudinal data, it is important to take into account the correlation between the repeated measurements within an individual as well as measurement error.

Missing data

In a longitudinal analysis, each subject i is designed to be measured at visits j ($j = 1, \dots, n_i$), meaning that we expect to collect the full vector of measurements $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})^\top$. In practice we usually have missing data that is either intermittent (a subject misses a visit and comes back at the next one) or permanent (when a subject leaves the study). The missing data mechanisms can be classified into 3 categories:

- Missing completely at random (MCAR): The probability to have a missing data for an individual conditional on the covariates is independent from the observations and missing observations of the marker. Individuals with missing data and individuals with complete records should have the same characteristics under the MCAR assumption.
- Missing at random (MAR): There are systematic differences between the missing and observed values which can be entirely explained by observed variables. The probability to have a missing data for an individual is therefore related to some other measured covariates in the model, but not to the value of the variable with missing values itself.
- Missing non at random (MNAR): The probability to have a missing data for an individual is conditional on the covariates as well as the observed and unobserved values of the marker. It is not possible to verify that missing values are MNAR without knowing the missing values.

To illustrate these missing data mechanisms, assume we are interested in the tumor size of cancer patients. We have a sample of measurements but some of them are missing. In the situation where the data are MCAR, the probability of a missing measurement is the same for all patients. Now, if older patients have larger tumor sizes and have higher probability of drop-out, the observed sample will not reflect the true distribution of the biomarker from the original population but the distribution of the biomarker conditional on the age of the patients will be similar. The data are therefore MAR because missing values depend on an observed variable (i.e. age). Based on standard missing data theory (Rubin (1976)), likelihood-based methods ignoring the missingness mechanism provide unbiased estimates given that the data are missing at random (MAR) or completely at random (MCAR), because a model can predict the missing data to obtain unbiased estimates (Verbeke and Molenberghs (2000)). Finally, if patients with higher tumors size are less likely to produce measurements (e.g., because of the disability provoked by the large tumors themselves), the distribution of the observed tumors will differ from the distribution in the population of interest. When the missingness mechanism depends on some unobserved aspect of the data, i.e., the data are missing not at random (MNAR), likelihood-based methods may be biased for certain parameters (Fitzmaurice et al. (1995); Molenberghs and Verbeke (2001); Kurland and Heagerty (2005); Rouanet et al. (2019)). It is a fundamentally untestable assumption, because it concerns the unobserved values. Therefore, the assumption needs to be justified based on background knowledge and discussion with experts. In case missing data are the result of informative drop-out (e.g. death), the marker measurements and the drop-out mechanism have to be jointly modeled to obtain valid estimates.

Confounders

A confounder is a variable that affects the estimation of an association between an exposure and an outcome when included in a regression model. Confounders are the main obstacle to make causal inference with regression models because when it is missing in a regression model, the measured association between the exposure and the outcome can be biased. An example of this bias is given in figure 2.1.

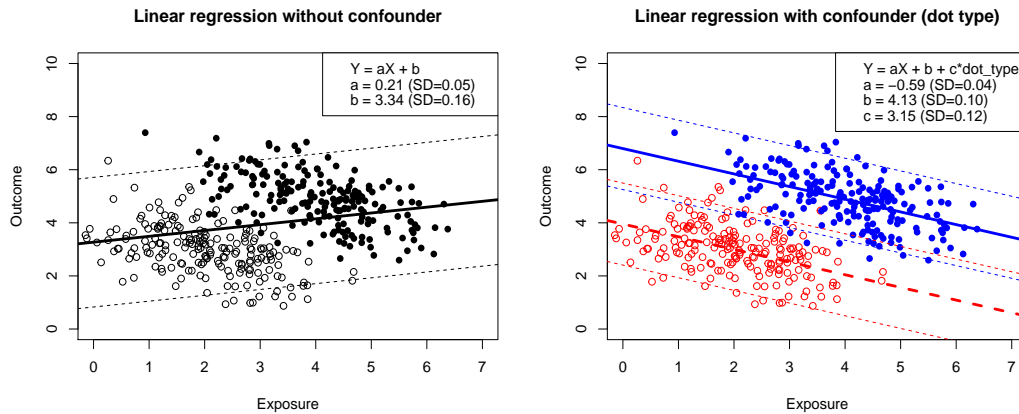


Figure 2.1: Illustration of Simpson's paradox. When the confounder (i.e., the dot type) is not included in the regression model, we conclude to a statistically significant positive trend while the true trend observed when the confounder is included in the model is negative.

2.1.1 Linear mixed effects model

Description

The linear mixed effects (LME) model (Harville (1977); Laird and Ware (1982)) is an extension of the simple linear regression model that allows the inclusion of both fixed and random effects to model a continuous marker. Fixed effects corresponds to the effect of observed variables, it is constant across individuals while a random effect corresponds to subject-specific effect of latent (i.e., unobserved) variables. The inclusion of random effects in a fixed effects model assists in controlling the unobserved heterogeneity. LME models are particularly useful for data with a hierarchical structure, for which the independence assumption is violated. The structure of the data can be decomposed into groups that share some common variability in the measurements. In the context of individual repeated measurements, each individual is a group and there is a variability intra-groups and a variability inter-groups.

Let Y_{ij} denote the biomarker measurement of subject i ($i = 1, \dots, n$) at time j ($j = 1, \dots, n_i$) and $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})^\top$ the vector of responses for subject i . We assume the observed biomarker value is noisy (e.g., due to the precision of the tools used to produce the measure), the true value of the biomarker Y_{ij}^* remains unobserved.

$$\begin{aligned} Y_{ij} &= Y_{ij}^* + \epsilon_{ij} \\ &= \mathbf{X}_{ij}^\top \boldsymbol{\beta} + \mathbf{Z}_{ij}^\top \mathbf{b}_i + \epsilon_{ij} \end{aligned}$$

The vectors of explanatory variables \mathbf{X}_{ij} and \mathbf{Z}_{ij} are associated with the fixed effects regression coefficients $\boldsymbol{\beta}$ and the random effects \mathbf{b}_i , respectively. The random effects are assumed to be Gaussian distributed and possibly correlated $\mathbf{b}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_b^2)$. The measurement error ϵ_{ij} follows a Gaussian distribution $\mathcal{N}(0, \sigma_\epsilon^2)$ and is assumed independent from the random effects. The repeated measures of Y are assumed independent conditional on the random effects. random effects account for unobserved confounders that affects the individual, such as genetic or environmental exposure. Sometimes, additional levels of hierarchy can explain some variability shared between individuals (e.g., geographical areas or shared genetic features). The model gives the marginal expectation of \mathbf{Y}_{ij} for the entire population sharing features \mathbf{X}_{ij} through the term $\mathbf{X}_{ij}^\top \boldsymbol{\beta}$ and the subject-specific expectation of \mathbf{Y}_{ij} by $\mathbf{X}_{ij}^\top \boldsymbol{\beta} + \mathbf{Z}_{ij}^\top \mathbf{b}_i$.

Estimation

The parameters of the LME model $(\boldsymbol{\beta}, \boldsymbol{\Sigma}_b, \sigma_\epsilon)$ can be estimated with maximum likelihood (ML), where the likelihood function corresponds to the joint probability distribution of the sample. The variance estimator of the parameters is obtained by the inverse of the Hessian matrix. Let \mathbf{X}_i and \mathbf{Z}_i denote the matrices of explanatory variables with row vectors \mathbf{X}_{ij}^\top and \mathbf{Z}_{ij}^\top , respectively. The model can be written in vector form as

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i \text{ with } \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma_\epsilon^2 \mathbf{I}_{n_i}).$$

The model can be given in a marginal formulation, assuming Y_{ij} is Gaussian distributed such that $\mathbf{Y}_i \sim \mathcal{N}(\mathbf{X}_i \boldsymbol{\beta}, \mathbf{V}_i = \mathbf{Z}_i \boldsymbol{\Sigma}_b \mathbf{Z}_i^\top + \sigma_\epsilon^2 \mathbf{I}_{n_i})$. Let $\boldsymbol{\Theta}$ denote the vector of parameters of the model, such that $\boldsymbol{\Theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\Sigma}_b, \sigma_\epsilon)$, the likelihood function can be expressed as

$$L_i(\boldsymbol{\Theta}) = \prod_{i=1}^n \prod_{j=1}^{n_i} p(Y_{ij}),$$

where $p(Y_{ij})$ is the Gaussian probability density function of the outcome. There is a closed-form expression for the log-likelihood, defined as follows

$$\log(L_i(\boldsymbol{\Theta})) = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^{n_i} \{n_i \log(2\pi) + \log |\mathbf{V}_i| + (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta})^\top \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta})\},$$

where $|\mathbf{V}_i|$ is the determinant of \mathbf{V}_i . It can be directly maximized with respect to $\boldsymbol{\Theta}$ by an iterative procedure to obtain the maximum likelihood estimate $\hat{\boldsymbol{\Theta}}$. The posterior distribution of the random effects $\mathbf{b}_i | \mathbf{Y}_i$ can be useful for prediction purpose, it has a multivariate normal distribution and the individual random effect b_i is usually estimated by taking the mean of this

posterior distribution, which has a closed-form expression in the context of linear mixed effects model. A drawback of the maximum likelihood approach is that it does not take into account the loss in the degrees of freedom resulting from estimating fixed effects, resulting in biased estimates of $\hat{\Sigma}_b$ and $\hat{\sigma}_\epsilon$. An alternative approach maximizes the restricted maximum likelihood, which takes into account the dimension of the orthogonal vectorial space of \mathbf{X}_i (Verbeke (1997)). However, maximum likelihood and restricted maximum likelihood techniques give similar estimates for large sample sizes and the ML approach is often used. The model gives an additive effect of covariates (e.g., treatment) on the marginal mean of the biomarker. The simple linear regression model is often limited for the analysis of clinical longitudinal biomarkers. Parametric functions or splines are often favoured to capture the non-linear evolution of the biomarker over time, which is useful to have a flexible fit for prediction purposes. The interpretation of covariates effect, when using such functions, is however more difficult. The outcome often requires a non-linear transformation to handle skewness and heteroscedasticity and to satisfy the hypothesis of a Gaussian distributed error term. However, when applying such transformation to the outcome, the effect of the covariates is given on the transformed scale. It is often difficult to transform back this effect on the natural scale because the non-linear transformation modifies the Gaussian distribution of the subject-specific random effects. It is however possible to use a non-linear link function for the biomarker to avoid such transformation for non-Gaussian outcomes.

2.1.2 Generalized linear mixed effects model

The linear mixed effects model is limited to Gaussian distributed outcomes but as described in Section 2.1, other types of distributions are encountered and require an appropriate methodology. The Generalized Linear Mixed effects Model (GLMM) is a flexible generalization of ordinary linear mixed effects model designed for response variables that have error distribution models other than a Gaussian distribution. The distribution of the response variable is assumed to belong to the exponential family including in particular the Gaussian, Bernoulli, Binomial and Poisson distributions. The linear model is related to the response variable through a link function $g(\cdot)$. Using the same notations as in previous section, the model is defined as

$$g(Y_{ij}) = \mathbf{X}_{ij}^\top \boldsymbol{\beta} + \mathbf{Z}_{ij}^\top \mathbf{b}_i + \epsilon_{ij}.$$

The corresponding likelihood function is

$$\begin{aligned} L_i(\boldsymbol{\Theta}) &= \prod_{i=1}^n \prod_{j=1}^{n_i} p(Y_{ij}), \\ &= \prod_{i=1}^n \prod_{j=1}^{n_i} \int_{b_i} p(Y_{ij}|b_i)p(b_i)db_i. \end{aligned}$$

As opposed to linear mixed effects model, the calculation of the log-likelihood has no analytic solution and the integral over the random effects \mathbf{b}_i is computed numerically in most cases (exceptions include, for example, the log-linear Poisson model with Gamma distributed random effects). The lack of analytical solution is due to the non-linear function $g(\cdot)$ that links the linear predictor to the outcome. Numerical computation of the integral over the random effects complicates the computation of the likelihood, making the maximum likelihood estimation of GLMMs

much more difficult than for the linear mixed model (see Section 2.4.3). In this manuscript, we will focus on the logarithm and logit link functions.

Logarithm link function

The logarithm function corrects for right skewness and heteroscedasticity. It is defined only for positive real numbers which is convenient for many applications that cannot have negative values. It is very common in biological data and fits particularly well the dynamics of cancer cells for which the unrestricted growth follows an exponential increase law (Koch (1966)). The model is defined as

$$\begin{aligned}\log(E[Y_{ij}]) &= \mathbf{X}_{ij}^\top \boldsymbol{\beta} + \mathbf{Z}_{ij}^\top \mathbf{b}_i, \\ \text{or } E[Y_{ij}] &= \exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta} + \mathbf{Z}_{ij}^\top \mathbf{b}_i).\end{aligned}$$

There is an important difference between the GLMM with a log link function and the LME with a log-transformed outcome. Indeed, the GLMM models the logarithm of the expected outcome $\log(E[Y_{ij}])$, which is different from the expected log-transformed outcome $E[\log(Y_{ij})]$. Therefore, with a GLMM, $\exp(\beta_k)$ represents the multiplicative effect on the mean biomarker value associated with a one unit increase in covariate k . With a LME using a log-transformed outcome, β_k gives the additive effect of covariate k on the log scale, which is less relevant for inference as we are usually interested in the effect of a covariate on the natural scale of the outcome.

Logit link function

The logit link function is useful to model a binary outcome ($Y_{ij} = 0$ or 1). It evaluates the effect of covariates on the probability p to observe $Y_{ij} = 1$. The logit function is defined by

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right).$$

Therefore the model has the form

$$\begin{aligned}\text{logit}(p) &= \mathbf{X}_{ij}^\top \boldsymbol{\beta} + \mathbf{Z}_{ij}^\top \mathbf{b}_i, \\ \text{or } p &= \frac{\exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta} + \mathbf{Z}_{ij}^\top \mathbf{b}_i)}{1 + \exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta} + \mathbf{Z}_{ij}^\top \mathbf{b}_i)}.\end{aligned}$$

The model does not include an error term because we model the mean of p and not each value of Y with the predicted mean plus an error term. With the logit link, $\exp(\beta_k)$ represents the subject-specific odds ratio to observe $Y_{ij} = 1$, associated with one-unit increase in the k th covariate. In the specific case of random intercept logistic models, the subject specific estimates can be converted to a population average coefficient through the following equation

$$\beta_{pa} = \frac{\beta_{ss}}{1 + 0.346\sigma_a^2},$$

where β_{pa} is a population average estimate, β_{ss} a subject specific estimate and σ_a^2 is the variance of the random intercept (Hu et al. (1998)). However, this formula only works for random intercept models, and cannot be used for models with additional random effects for which a Monte Carlo sampling procedure should be used instead to obtain marginal coefficients and their standard deviation.

2.1.3 Non-linear mixed effects models

The LME and the GLMM are both linear in the parameters (for fixed and random effects). In contrast, a non-linear mixed effects model includes also non-linear functions of the fixed and random effects. It is defined as

$$Y_{ij} = g(\mathbf{X}_i, \boldsymbol{\beta}, \mathbf{Z}_i, \mathbf{b}_i) + \epsilon_{ij},$$

where the function $g(\cdot)$ is any parametric function specified a priori. These models can be estimated using maximum likelihood techniques or Bayesian inference, they are however usually difficult to interpret. Mechanistic models have also been proposed, they use ordinary differential equations to reflect on biological knowledge about the mechanisms underlying the disease dynamics. They can be helpful to understand the effect of covariates on the dynamics of biomarkers and can be useful for prediction purposes but inference about a treatment is difficult because of the model specification. It is possible to estimate a mechanistic model as a linear mixed effect model when an analytical solution exists, otherwise a specific algorithm is required to solve the differential equations. Mechanistic models were also proposed in the context of joint modeling to evaluate the association of the dynamics of the biomarker with a time to event outcome (Guedj et al. (2011); Król et al. (2018)). Recently, a Bayesian estimation of a non-linear mechanistic joint model was proposed to study the repeated measurements of the SLD jointly with survival time in an immunotherapy trial in patients with advanced or metastatic bladder cancer (Keroui et al. (2020)). The mechanistic model based on ODE helps understanding the sources of variability to immunotherapy but is limited for inference about a treatment.

2.1.4 Modeling strategies for a semicontinuous outcome

A semicontinuous distribution is characterized by a continuous distribution with one or more point masses. It is usually a half-bounded interval where in most cases, the distribution has a lower bound at zero resulting in either a positive value or a zero. In this manuscript, we focus on zero-inflated nonnegative continuous outcomes, and therefore may omit the word “nonnegative”. Such distributions are common, examples include medical costs (Manning et al. (1981); Duan et al. (1983); Liu et al. (2010)), alcohol consumption (Liu et al. (2008, 2016); Han et al. (2019)), gene expression data (McDavid et al. (2013); Finak et al. (2015)) or microbiome compositional data (Chen and Li (2016); Chai et al. (2018)). In the context of cancer clinical trials, the SLD has a semicontinuous distribution because some patients have a complete shrinkage of their tumors after treatment initiation. They are usually characterized by an excess of zeros, right skewness and heteroscedasticity for positive continuous values. An illustration of the distribution of the SLD in a randomized clinical trial is presented in Figure 2.2.

Tobit model

The tobit regression model can accommodate a semicontinuous outcome by assuming that the continuous distribution of the outcome is censored. It is based on Tobin (1958), the idea is to modify the likelihood function so that it reflects the unequal sampling probability for each observation. Let Y_{ij} denote the biomarker value for subject i ($i = 1, \dots, n$), at visit j ($j = 1, \dots, n_i$).

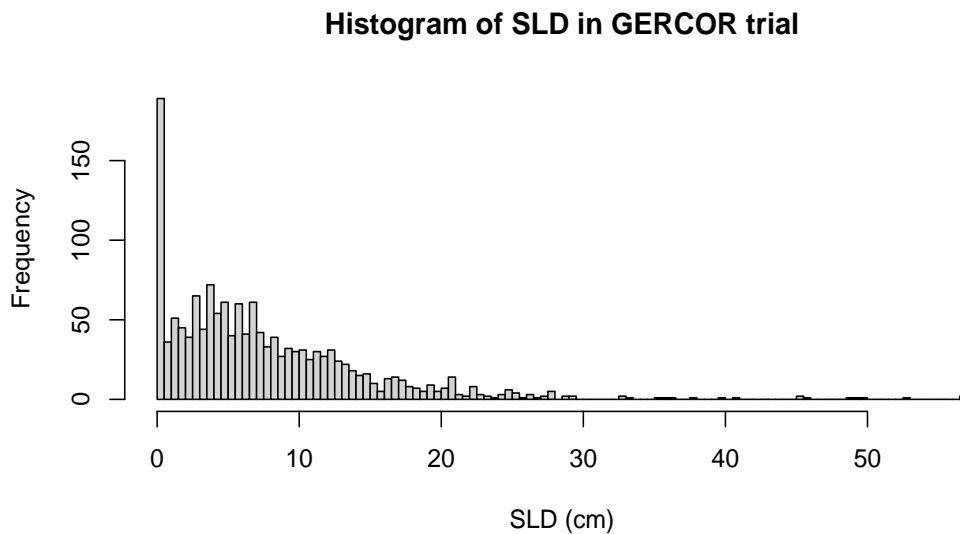


Figure 2.2: Histogram of the distribution of the SLD in the GERCOR study. The distribution is characterized by an excess of zeros and right skewness.

The model assumes the biomarker can be subject to left-censoring when it decreases below a limit of detection c .

$$Y_{ij}^* = \begin{cases} Y_{ij} & \text{if } Y_{ij} > c, \\ c & \text{otherwise.} \end{cases} \quad (2.1)$$

The resulting likelihood is based on the density function of the outcome when this one is observed or the corresponding cumulative distribution function for censored observations. The left-censoring (tobit) model has been applied in the context of HIV infection, where the outcome is composed of the longitudinal measurements of the viral load (Jacqmin-Gadda et al. (2000)). The semicontinuous distribution of the outcome is explained by the lower quantification limit of the viral load. An application of this model to the repeated measurements of the SLD in the context of cancer clinical trials was proposed in (Król et al. (2016)). The value of the SLD is assumed to have a quantification limit and the zero values are assumed to be censored values. When true zeros (i.e., not censored) are observed, the inference could be biased similarly as fitting a linear regression on a binary outcome because of the mixed discrete-continuous distribution of the outcome. Moreover, there is often an interest in what influences the probability of a zero value. For these reasons, two-part models were developed to account for true zeros.

Two-part model

The two-part model decomposes the semicontinuous outcome into a binary outcome (zero vs. positive values) and an outcome with positive continuous values. A GLMM with a probit or a logit link is used to fit the probability of observing a positive versus zero value. One of those models, a LME model on the log-transformed biomarker repeated measurements (Liu (2009)), a log-skew-normal distribution, a gamma generalized distribution (which includes the lognormal, gamma, inverse gamma, and Weibull distributions as special cases), links the continuously distributed values conditional on a positive outcome to the linear predictor (Smith et al. (2018)).

In the following, we describe the standard two-part model assuming a lognormal distribution for the positive values.

Let Y_{ij} denote the biomarker value for subject i ($i = 1, \dots, n$), at visit j ($j = 1, \dots, n_i$). The biomarker distribution is decomposed into a binary outcome $I[Y_{ij} > 0]$ and a positive-continuous outcome $Y_{ij}^+ = [Y_{ij}|Y_{ij} > 0]$. A GLMM with a logit link is assumed for the binary outcome. The distribution of the positive continuous values often requires a non linear transformation because it is not Gaussian. However, a GLMM can account for such transformation. We use a GLMM with a logarithm link for the positive continuous outcome in the following. The logarithm link in the continuous part is used to linearize the biomarker evolution over time and correct for right-skewness and heteroscedasticity. The two components are linked through correlated random effects. The two-part model is defined as follows:

$$\begin{cases} \text{Logit}(\text{Prob}(Y_{ij} > 0)) = \mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i & \text{(Binary part),} \\ \text{E}[Y_{ij}|Y_{ij} > 0] = \exp(\mathbf{X}_{Cij}^\top \boldsymbol{\beta} + \mathbf{Z}_{Cij}^\top \mathbf{b}_i) & \text{(Continuous part),} \end{cases}$$

where \mathbf{X}_{Bij} and \mathbf{Z}_{Bij} are vectors of covariates associated with the fixed effect parameters $\boldsymbol{\alpha}$ and the random effects \mathbf{a}_i for the binary part. Similarly, \mathbf{X}_{Cij} and \mathbf{Z}_{Cij} are vectors of covariates associated with the fixed effect parameters $\boldsymbol{\beta}$ and the random effects \mathbf{b}_i . We assume a normal and independently distributed error term in the continuous part $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$. The two vectors of random effects follow a multivariate normal distribution:

$$\begin{bmatrix} \mathbf{a}_i \\ \mathbf{b}_i \end{bmatrix} \sim MVN \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_a^2 & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ab} & \boldsymbol{\Sigma}_b^2 \end{bmatrix} \right). \quad (2.2)$$

The vectors of correlated subject-specific random effects \mathbf{a}_i and \mathbf{b}_i account for the correlation between repeated measurements within an individual and the correlation between the two components of the model. The overall mean with a conditional two-part model can be written as the product of expectations from the first and second parts of the model, as follows

$$\text{E}[Y_{ij}] = \text{Prob}(Y_{ij} > 0) \text{E}[Y_{ij}|Y_{ij} > 0].$$

The likelihood of a two-part model is defined as the product of the likelihood of the binary $L_i^B(\boldsymbol{\Theta})$ and continuous parts $L_i^C(\boldsymbol{\Theta})$, where $\boldsymbol{\Theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top, \boldsymbol{\Sigma}_b, \sigma_\epsilon)$ denotes the vector of parameters of the model. Introducing $U_{ij} = I[Y_{ij} > 0]$, the likelihood contribution from the binary part can be expressed as

$$\begin{aligned} L_i^B(\boldsymbol{\Theta}) &= \prod_{j=1}^{n_i} P(U_{ij}|\mathbf{a}_i), \\ &= \prod_{j=1}^{n_i} \text{Prob}(Y_{ij} > 0)^{U_{ij}} (1 - \text{Prob}(Y_{ij} > 0))^{(1-U_{ij})}, \\ &= \prod_{j=1}^{n_i} \left(\frac{\text{Prob}(Y_{ij} > 0)}{1 - \text{Prob}(Y_{ij} > 0)} \right)^{U_{ij}} (1 - \text{Prob}(Y_{ij} > 0)), \\ &= \prod_{j=1}^{n_i} \exp \left(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i \right)^{U_{ij}} \left(1 - \frac{\exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)}{1 + \exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)} \right). \end{aligned}$$

Assuming a lognormal distribution for positive values, the likelihood contribution from the continuous part is

$$L_i^C(\boldsymbol{\Theta}) = \prod_{j=1}^{n_i} \left\{ \frac{1}{Y_{ij} \sqrt{2\pi\sigma_\epsilon^2}} \exp\left(-\frac{(\log(Y_{ij}) - \mu_{ij})^2}{2\sigma_\epsilon^2}\right) \right\}^{U_{ij}},$$

where $\mu_{ij} = \mathbf{X}_{Cij}^\top \boldsymbol{\beta} + \mathbf{Z}_{Cij}^\top \mathbf{b}_i - \frac{\sigma_\epsilon^2}{2}$. This model is referred to as the conditional two-part model. An alternative marginal form of the model has been proposed, the difference is that the probability of positive value is taken into account in the continuous part in order to remove the condition on a positive value. Therefore, the continuous part of the marginal two-part model gives the effect of covariates on the unconditional mean biomarker value $E[Y_{ij}]$ instead of the conditional mean $E[Y_{ij}|Y_{ij} > 0]$ (Smith et al. (2014)). The difference in the likelihood is in the location parameter of the lognormal distribution for positive values that accounts for the probability of positive value, which is now

$$\begin{aligned} \mu_{ij} &= \mathbf{X}_{Cij}^\top \boldsymbol{\beta} + \mathbf{Z}_{Cij}^\top \mathbf{b}_i - \log(\text{Prob}(Y_{ij} > 0)) - \frac{\sigma_\epsilon^2}{2}, \\ &= \mathbf{X}_{Cij}^\top \boldsymbol{\beta} + \mathbf{Z}_{Cij}^\top \mathbf{b}_i + \mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i - \log(1 + \exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)) - \frac{\sigma_\epsilon^2}{2}. \end{aligned}$$

More details on the marginal two-part model and the difference in the interpretation compared to a conditional two-part model are given in Chapter 4.

2.2 Survival analysis

2.2.1 Outcome of interest

In survival analysis, the outcome is a time to event. Some patients will observe the terminal event while other patients do not. When the follow-up of the study is over, remaining patients are censored. Some patients leave the study early (i.e., before the end of follow-up), they are lost to follow-up and considered censored at their last visit. The event is usually death, but other events of interest (e.g., progression of cancer) can be analyzed with survival models. With the regression models used in survival analysis, multiple independent prognosis factors can be analyzed simultaneously and treatment differences can be assessed while adjusting for heterogeneity and imbalances in baseline characteristics. The shape of the distribution of survival time justifies the requirement for specific models, because survival times are always positive, they often have skewed shapes of distribution and thus, statistical methods that rely on normality are not directly applicable and may produce invalid results with survival data. However a suitable transformation of the event times, such as the logarithm of the square root can overcome this issue. The main difference in the analysis of survival times is the censoring event. It can be categorized as follows:

- Right censoring: The event time occurs after the last time point of observation.
- Left censoring: The event time occurs before the first time point of observation.
- Interval censoring: The event of interest is only known to occur between two time points.

Another way to classify censoring focuses on the relationship between the probability of a subject being censored and the failure process

- **Informative censoring:** It is similar to the MNAR missing data mechanism, the risk of event for individuals in the study is different from the risk of event from censored individuals.
- **Non-informative censoring:** It is similar to the MCAR missing data mechanism, the individual withdraws from the study for reasons not related to the study.

Censoring due to loss of follow-up should be non-informative to get unbiased estimate of survival curves (i.e., the censoring time is statistically independent from the failure time). In clinical trials, drop-out can occur when the disease progression leads a patient to die or withdraw from the trial to seek other treatment options, the drop-out is then informative for the estimation of the disease progression. If the disease progression correlates with a biomarker that is being monitored (e.g., SLD), modeling the disease progression separately from the drop-out process may be inefficient and produce biased estimates as explained in Section 2.1. Joint models can account for such informative drop-out but have limitations when the missing data mechanism is related to other reasons than the drop-out mechanism.

Let T^* denote the positive continuous response variable that represents the elapsed time between the beginning of the follow-up and the event of interest, usually referred to as survival time or event time. There are several ways to describe the distribution of survival times:

- **Survival function**

The survival function is the probability to survive at least until time t : $S(t) = P(T^* > t)$. It is a decreasing function starting at 1 at time 0 that converges towards 0 as t tends towards $+\infty$.

- **Cumulative distribution function**

The cumulative distribution function (cdf) represents the probability that death occurs before or at time t : $F(t) = P(T^* < t) = 1 - S(t)$.

- **Probability density function**

The probability density function (pdf) corresponds to the probability of dying in a very short time interval after t : $f(t) = \lim_{h \rightarrow 0^+} \frac{P(t \leq T^* < t+h)}{h}$.

We can therefore express the relationship between the pdf and the cdf: $F(t) = \int_0^t f(u) du$.

- **Hazard function**

The hazard function corresponds to the probability that death occurs in a small interval of time after t , conditionally on surviving until t , i.e., the instantaneous risk of event for individuals free from the event: $\lambda(t) = \lim_{h \rightarrow 0^+} \frac{P(t \leq T^* < t+h | T^* > t)}{h} = \frac{f(t)}{S(t)}$.

- **Cumulative hazard function**

Finally, the cumulative hazard function corresponds to the cumulative hazard up to time t (i.e., the total amount of cumulated risk): $\Lambda(t) = \int_0^t \lambda(u) du$.

We can deduce the relationship between the survival and hazard function: $S(t) = \exp(-\Lambda(t)) = \exp(-\int_0^t \lambda(u)du)$.

These quantities of interest can be estimated by non-parametric estimators along with regression models that describe these quantities as function of explanatory variables. A well-known non-parametric estimator of the survival function is the Kaplan-Meier estimator (with the Greenwood formula to estimate the variance). We can compare the survival functions of several groups of subjects with the log-rank test. Another common estimator is the Nelson-Aalen estimator for the cumulative hazard function and its variance. For a population of n patients, we can define the observations as couples (T_i, δ_i) , with $i = 1, \dots, n$ and the indicator variable $\delta_i = I_{T_i^* < C_i}$, equal to 1 if the survival time is observed and 0 in case of incomplete observation (i.e., censoring). The observed time T_i is $T_i = \min(T_i^*, C_i)$, where C_i denote the censoring time of individual i .

2.2.2 The proportional hazards model

The Cox proportional hazards model is the most commonly used statistical model to study the relationship between the survival time of patients and predictor variables. While the Kaplan-Meier curves and the log-rank test are limited to categorical variables, the Cox PH model can include both categorical and quantitative variables and study their effect on the risk of event simultaneously. The Cox PH model is usually described by its hazard function:

$$\lambda_i(t) = \lambda_0(t) \exp(\mathbf{X}_i^\top \boldsymbol{\gamma}),$$

where $\lambda_0(t)$ is the baseline hazard function, corresponding to the risk of event at time t if all the \mathbf{X}_i are equal to zero. This baseline hazard acts as a time-dependent intercept in the model, and the rest of the equation is a multiple linear regression of the logarithm of the hazard on the variables \mathbf{X}_i . We are usually interested in the hazard ratio $\exp(\gamma_k)$ of a covariate X_k comparing the risk of event for patients with $X_k = 1$ to the risk for patients with $X_k = 0$ in case of binary covariate (with a continuous covariate, $\exp(\gamma_k)$ gives the hazard ratio of a 1-unit increase in covariate X_k). A covariate associated to a hazard ratio > 1 increases the risk of event, it is a bad prognosis factor while a covariate associated to a hazard ratio < 1 is a good prognosis factor. The key assumption of the Cox PH model is the proportional hazards, meaning that the hazard of the event in any subgroup defined by the covariates is a constant multiple of the hazard in any other subgroup. However, when time-dependent covariates are included in the Cox PH model, the hazard ratio between two individuals can vary over time (Fisher and Lin (1999)). There are two types of contribution to the likelihood for survival times with a Cox PH model:

- Individual i is censored alive at T_i , his contribution to the likelihood is defined by the survival function $S(T_i)$.
- Individual i dies at time T_i , his contribution to the likelihood corresponds to the probability density function $f(T_i) = \lambda_i(T_i)S(T_i)$.

The likelihood contribution of individual i is therefore defined as

$$\begin{aligned} L_i(\cdot) &= \lambda_i(T_i)^{\delta_i} S(T_i), \\ &= \lambda_i(T_i)^{\delta_i} \exp\left(-\int_0^{T_i} \lambda_i(t) dt\right), \end{aligned}$$

where $\lambda_i(t) = \lambda_0(t) \exp\{\mathbf{X}_i^\top \boldsymbol{\gamma}\}$. An alternative approach avoids the specification of the baseline hazard: the semi-parametric proportional hazards model (Cox (1972)). The baseline hazard is left unspecified with this approach, and the Cox PH model can be estimated using the partial likelihood (Breslow (1972)). We are however often interested in the estimation of the baseline hazard functions. Moreover, in some situations (e.g., joint modeling), the baseline hazard function is approximated to facilitate likelihood inference.

2.2.3 Baseline hazard approximation

There are several ways to define the baseline hazard risk function $\lambda_0(t)$. In a parametric proportional hazards model, the baseline is defined as a function of parameters that are estimated with other parameters of the model, the most common choices are the exponential, Weibull and Gompertz distributions. Piecewise constant functions do not make any distribution assumption for the baseline hazard risk function but tend to lack flexibility because of the assumption of piecewise constant hazards (and therefore “brutal” jumps in the hazard). Additional flexible parametric methods (e.g., splines) have been proposed for the approximation of the baseline distribution function (Royston and Parmar (2002)). Spline functions should be used with caution as their flexibility relies on an appropriate choice of the number of knots. They might overfit the data and provide a rough (i.e., not smooth) hazard function, which is usually not suitable as the risk function in the population is usually smooth. The degrees of freedom of the spline can be chosen with post-estimation model selection criteria. It is sometimes preferable to opt for a penalized likelihood approach to obtain smooth non-parametric estimator of the hazard function (Rondeau et al. (2007)). The penalization of the second-order derivatives of the splines prevents rough changes in the hazards risk function. Choosing the smoothing parameter is the difficult part of the method and can be done by approximate cross-validation techniques. Alternatively, leaving the baseline hazard completely unspecified is possible but the estimation of the standard errors of the regression parameters is computationally intensive. It may raise convergence problems and these standard errors may be underestimated (Xu et al. (2020)).

2.2.4 Time-dependent covariates

It is possible to include time-dependent covariates in a survival model, they can be classified into two categories:

- Exogeneous (or external) covariates remain measurable and their distribution is unchanged after the occurrence of the event.
- Endogeneous (or internal) covariates’ distribution is affected by the event.

The proportional hazards model can handle exogeneous time-dependent covariates but the likelihood requires knowing the value of these covariates for all subjects at risk for each event time.

When covariates measurements does not coincide with event times in the sample, models are required to impute values at the times of events. However most biomarkers of interest in clinical research are endogeneous variables, their values are affected by a change in the risk of occurrence of the event. For example in a cancer clinical trial, if a treatment reduces both the risk of death and the SLD, adjusting a survival model on the SLD may severely bias the effect of treatment on the risk of death. It is often of interest to estimate the effect of the treatment on the biomarker. The biomarker is censored by the event and its value is only known at the specific time points at which it is measured. Separate models for the longitudinal and survival outcomes are prone to bias when the two outcomes are associated (Rubin (1976), Wang and Taylor (2001)). Alternatively, two-stage models can fit a longitudinal model and subsequently use the estimated longitudinal trajectory of the biomarker as a covariate in the survival model (Tsiatis et al. (1995)). They account for measurement error of the biomarker in the survival model but fail to account for informative drop-out in the longitudinal model. Ignoring informative censoring in a longitudinal analysis can lead to biased covariates' effect estimates (García-Hernandez et al. (2020)). In contrast, joint models offer the advantage of dealing with informative drop-out, measurement error and missing biomarker measurements not at random (MNAR) in the longitudinal and survival regression models (Ibrahim et al. (2010)).

2.3 Joint modeling for longitudinal data and a terminal event

When we observe repeated measurements of a biomarker and an event of interest, there is often an association between these two outcomes such that the risk of event depends on the longitudinal biomarker and the biomarker measurements are censored by the event.

The joint model is able to analyze event history data linked to a time-dependent endogeneous biomarker. It also improves efficiency of statistical inference by using both the longitudinal biomarker measurements and survival times simultaneously, taking into account the dependency and association between longitudinal data and time-to-event data. With joint models, the regression model for the longitudinal measurements allows for outcome dependent drop-out while the survival submodel provides inference on the distribution of time-to-event conditional on intermediate longitudinal measurements.

To summarize, joint models for a longitudinal biomarker and a terminal event are useful when we are interested in

- Study the biomarker's evolution when follow-up is censored by the terminal event, causing non-random drop-out.
- Study the risk of terminal event while accounting for the effect of an endogeneous time-dependent covariate measured with error.
- Explore the association between the biomarker and the risk of event.
- Predict the risk of event from the repeated measurements of the biomarker.

The standard joint model uses shared random effects to analyze a longitudinal process and an associated survival process, they are referred to as “shared random effects joint models”. The main alternative is the “latent class joint model”, which assumes that the population is divided in homogeneous groups of subjects with regards to both the marker trajectory and the event risk (Lin et al. (2000); Proust-Lima et al. (2014)). They are however out of the focus of this thesis and we describe only the shared random effects joint model.

2.3.1 Shared random effects joint model

The joint model connects separate models for different types of data in order to form one complex model. In this thesis, we focus on joint models for a longitudinal biomarker Y and the event time T (Faucett and Thomas (1996), Wulfsohn and Tsiatis (1997)). The most common joint model uses shared random effects to account for the correlation between the longitudinal and event history processes. To formulate the model, let’s consider the standard setting of a Gaussian longitudinal biomarker correlated to the time to a terminal event. The model is decomposed into two submodels:

$$\begin{cases} Y_{ij} = Y_{ij}^* + \epsilon_{ij} & \text{(Biomarker submodel),} \\ = \mathbf{X}_{ij}^\top \boldsymbol{\beta} + \mathbf{Z}_{ij}^\top \mathbf{b}_i + \epsilon_{ij}, & \\ \lambda_i(t) = \lambda_0(t) \exp(\mathbf{X}_i(t)^\top \boldsymbol{\gamma} + h(\cdot)^\top \boldsymbol{\varphi}) & \text{(Survival submodel),} \end{cases} \quad (2.3)$$

where \mathbf{X}_{ij} and \mathbf{Z}_{ij} are vectors of covariates that can be time-dependent and respectively associated with the fixed effects $\boldsymbol{\beta}$ and the individual random effects \mathbf{b}_i . We assume measurement error for the biomarker (e.g., the tool used to produce the measurement has a limited precision), therefore the true value of the biomarker Y_{ij}^* is not observed. The error term ϵ_{ij} is assumed to follow a Gaussian distribution with standard deviation σ_ϵ . The random effects are assumed to follow a multivariate normal distribution that takes into account their correlation $\mathbf{b}_i \sim \text{MVN}(0, \Sigma_b)$. The model assumes independence between the random effects and the error term. In the survival submodel, $\lambda_0(t)$ is the baseline hazard. It is rarely left unspecified to avoid untractable computation as discussed in Section 2.2.3. The vector of covariates (i.e., prognostic factors) $\mathbf{X}_i(t)$ are associated with the fixed effects $\boldsymbol{\gamma}$ and the multivariate function $h(\cdot)$ is associated with the vector of parameters $\boldsymbol{\varphi}$. This function is referred to as the association structure or the link function between the biomarker submodel and the survival submodel. In the context of shared random effects, the random effects or functions of the random effects are introduced in the survival model to account for the individual heterogeneity in the biomarker dynamics and evaluate their association with the risk of event.

The shared random effects joint model can refer to any association structure that involves the individual heterogeneity captured by the random effects. It contrasts with the latent class joint model, which assumes that the population is heterogeneous and therefore can be decomposed into subpopulations that share similar profiles for the evolution of the biomarker and the risk of event. The latent class joint model is well adapted for prediction purposes, because it makes no assumption about the structure of the population, while the shared random effects model is preferred to evaluate the relationship between the biomarker and the survival time. In this thesis, we focus on the shared random effects joint models and we will use the terminology “shared random effects” to define the association structure of the joint model where only the random

effects are shared between the biomarker submodel and the survival submodel, each random effect being associated independently to the risk of terminal event, such that $h(\cdot) = \mathbf{b}^\top \boldsymbol{\varphi}$. Another association structure often encountered is the “current level” (or current value) of the biomarker, which assumes the risk of event depends on the true unobserved value of the biomarker, given by the biomarker submodel such that $h(\cdot) = \mathbb{E}[Y_{ij}] \boldsymbol{\varphi}$. Beyond these two main association structures, any function of the random effects can be defined as the association structure but complex associations are often difficult to interpret.

Association structures

Any function of the random effects can define the association structure between the longitudinal and survival submodels. We review the most common association structures proposed in the literature for joint models:

- $h(\cdot) = \mathbf{b}^\top$.

The “shared random effects” (SRE) association structure refers to the joint model sharing only the random effects between the biomarker submodel and the survival submodel, each random effect being associated independently to the risk of terminal event.

- $h(\cdot) = \mathbf{Z}_i(t)^\top \mathbf{b}$.

An alternative formulation includes the covariates associated to the random effects, therefore the event risk at t is a function of the individual deviation of the marker at t , which is time-dependent.

- $h(\cdot) = Y_i^*(t) = \mathbf{X}_i(t)^\top \boldsymbol{\beta} + \mathbf{Z}_i(t)^\top \mathbf{b}_i$.

The “current level” (CL) association structure (or current value) assumes that the instantaneous risk of event at t depends on the value of the biomarker at t free of measurement error, given by the biomarker submodel.

- $h(\cdot) = Y_i^{*'}(t)$,

where $Y_i^{*'}(t)$ is the derivative of the function $Y_i^*(t)$ with respect to t at time t . It assumes the instantaneous risk of event at t depends on the slope at t .

- $h(\cdot)^\top = (Y_i^*(t), Y_i^{*'}(t))$.

This model is more flexible because it assumes that the instantaneous risk of event at t depends both on the true current value of the marker and on the slope at t .

In this thesis, we focus on the SRE association structure and the CL association structure. The first one is useful to account for the individual heterogeneity of the population in the survival model while the second can evaluate the association between the biomarker value and the risk of event.

Likelihood expression

With joint models, the baseline hazard function is usually approximated to facilitate likelihood inference. We can express the likelihood contribution of individual i by taking advantage of the conditional independence between the biomarker \mathbf{Y}_i and the event time T_i . Let

$\Theta = (\beta^\top, \gamma^\top, \Sigma_b, \lambda_0(t))$ be the vector of parameters of the model, where $\lambda_0(t)$ refers to the parameters associated to the estimation of the baseline hazard function. The likelihood contribution of individual i is defined as

$$\begin{aligned} L_i(\Theta) &= p(\mathbf{Y}_i, T_i, \delta_i), \\ &= \int_{b_i} p(\mathbf{Y}_i|b_i)p(T_i, \delta_i|b_i)p(b_i)db_i, \end{aligned}$$

where

$$p(\mathbf{Y}_i|b_i) = \prod_{j=1}^{n_i} p(Y_{ij}|b_i).$$

Computation of the integral over the random effects is the main difficulty since it has no closed form expression (See Section 2.4.3 for details on the integral over the random effects). Moreover, the univariate integral over time in the survival function has no analytical solution when the association structure is time-dependent and must be approximated too.

2.3.2 Extensions of the standard joint model

Several extensions of the standard joint model for a longitudinal biomarker and a terminal event have been proposed, for example joint models for multiple longitudinal biomarkers and a terminal event (Rizopoulos and Ghosh (2011)) or joint models for recurrent events and a terminal event (Liu et al. (2004); Rondeau et al. (2007)). In the context of cancer trials, clinical progression can be characterized by the occurrence of multiple events, for example the progressive disease status of non-target lesions or new lesions has been considered as a recurrent event in a joint analysis of the SLD and the survival time (Król et al. (2016, 2018)). Moreover, the linear mixed effects model for the biomarker can be replaced by a GLM or more complex non-linear regression models to characterize the evolution of the biomarker differently.

2.4 Computational aspects

The two main inferential approaches are the frequentist and the Bayesian approaches. A brief introduction is proposed in this section. Note that the optimization algorithms presented for the frequentist inference can also be used in Bayesian inference but specific algorithms are usually preferred. The frequentist approach most often relies on the maximum likelihood approach while in Bayesian inference, we want to calculate the estimator of the “maximum *a posteriori*”.

2.4.1 Frequentist inference

The frequentist inference provides an objective measure of uncertainty under a specified statistical model. The total likelihood is defined as

$$L(\Theta) = \prod_{i=1}^n L_i(\Theta).$$

For computational convenience, the maximization of the likelihood function $L(\Theta)$ is usually done using the log-likelihood function (natural logarithm of the likelihood). In simple cases such as

the linear mixed effects model, one can compute the maximum likelihood estimators analytically. In most non-linear models, there is no closed-form expression and the log-likelihood has to be maximized with an iterative procedure. We review the main optimization algorithms used to maximize the likelihood, which turns out to be a minimization problem when working on minus the log-likelihood.

The Nelder-Mead algorithm, also referred to as the “downhill simplex algorithm” has the advantage of not using the derivatives of the likelihood function, but it is not the most efficient. The gradient descent algorithm belong to the class of optimization algorithms known as conjugate gradient, which uses the first-order derivatives. However, the Newton-Raphson is much more efficient because it uses in addition the second derivative of the function to minimize. The first and second derivatives are often difficult to calculate analytically but there are efficient algorithms to compute them numerically. An advantage of the Newton-Raphson method is that it provides a direct estimate of the variance of the maximum likelihood estimators through the Fisher information matrix. Nonetheless, the Newton-Raphson algorithm may be unstable, in particular if the initial value is far from the maximum. Several algorithms were proposed based on Newton’s method (e.g., BGFS algorithm). The Levenberg-Marquardt algorithm combines the gradient descent algorithm with a modified Newton-Raphson algorithm (Gauss-Newton). The gradient descent is useful when parameters are far from their optimal value and the Newton algorithm intervenes when the parameters are close to their optimal value. This algorithm has a much more stable behavior than the Newton-Raphson algorithm in complex problems (Commenges and Jacqmin-Gadda (2015)). Note that an important point in the optimization procedure is to have a good stopping criterion, to prevent the algorithm to stop before convergence.

2.4.2 Bayesian inference

Under the Bayesian framework, the state of knowledge or ignorance about the set of parameters of the model Θ before the data is available is defined by a prior distribution $\pi(\Theta)$. It plays an important role in Bayesian analysis, prior distributions are often associated with the fear that the prior may dominate and distort the information in the observed data. In the context of scientific inference, we would usually like the data to “speak for themselves” and consequently conduct the analysis as if a state of relative ignorance existed a priori. Inference is based on the posterior distribution of Θ . Given the prior distribution $\pi(\Theta)$, the likelihood function $\pi(D|\Theta) = L(\Theta)$ and the data D , it is possible to calculate the posterior probability distribution $\pi(\Theta|D)$ of Θ given the data D using Bayes theorem:

$$\pi(\Theta|D) = \frac{\pi(D|\Theta)\pi(\Theta)}{\pi(D)},$$

where the marginal likelihood (i.e., distribution of the observed data marginalized over the parameters) $\pi(D) = \int_{\Theta} \pi(D|\Theta)\pi(\Theta)d\Theta$ acts as a normalizing constant. Complex models usually cannot be processed in closed form by a Bayesian analysis, therefore efficient simulation-based Monte Carlo techniques like the Gibbs sampling or Metropolis-Hastings algorithm are often used. In the context of joint modeling, Bayesian inference is particularly efficient for complex models, defined by a large number of random effects and/or outcomes but might lose efficiency compared to a frequentist alternative for simpler models. For joint models with a non-linear regression

model for the biomarker such as mechanistic models as discussed in Section (2.1.3), the Hamiltonian Monte Carlo algorithm is preferred over other MCMC techniques (Kerouhi et al. (2020)). Recently, the Integrated Nested Laplace Approximation (INLA) algorithm has been introduced as an alternative to MCMC techniques for latent Gaussian models (LGMs). Many statistical models for spatial statistics, time series, etc., can be formulated as LGMs. A key feature of INLA is to provide approximations of the posterior marginals needed for Bayesian inference very efficiently and that still remain very accurate compared to MCMC methods (Rue et al. (2017)). By formulating complex joint models as LGM's, **R-INLA** can be used to fit these models as developed recently (Van Niekerk, Bakka, and Rue (2019); Van Niekerk, Bakka, Rue, and Schenk (2019)). It improves drastically the applicability of the Bayesian estimation for joint models.

2.4.3 Integrals over the random effects

Several techniques can compute the integral over the random effects required in the likelihood function of most models presented above. The most frequently used are the Gaussian quadrature rules that approximate the integral by a sum of the integrand computed at predefined points and weighted according to the type of integral. When the random effects are Gaussian, the abscissas and weights of the Gauss-Hermite quadrature are used. The adaptive quadrature centers the quadrature points around the predicted values of the random effects at each iteration, which results in a more accurate approximation of the integral (Lesaffre and Spiessens (2001)). It is however limited to simple models as the computational burden increases sharply with the dimension of the random effects. Monte Carlo methods rely on random sampling to approximate the integral, it is particularly useful for higher-dimensional integrals as the computational burden is much less affected by the dimension of the random effects. Finally, Laplace approximation is an alternative analytical approximation based on Taylor expansions. It has been proposed for two-part models (Olsen and Schafer (2001)), showing greatly reduced computation times with consistent accuracy compared to other Monte Carlo and quadrature techniques (Liu et al. (2008)).

2.4.4 Available programs and packages

Most statistical softwares can fit joint models using the maximum likelihood method, the SAS macro **JMfit** (Zhang et al. (2016)) can fit a standard joint model for longitudinal and survival data. With STATA, this standard joint model can be fitted with the command **stjm** (Crowther (2013)), which allows to use splines or polynomials to model the biomarker over time and includes spline-based approach for the baseline hazard function. The recently introduced **merlin** STATA package (Crowther (2018)) provides a unified environment to fit various joint models with multiple outcomes of different kinds (e.g., terminal event, recurrent events and longitudinal biomarkers). With R, several packages provide functions for fitting joint models for a longitudinal biomarker and a terminal event. The R package **JM** (Rizopoulos (2010)) is among the most widely used in the frequentist framework. Other packages provide alternative methods for the estimation, for example the R package **frailtypack** (Rondeau et al. (2020)) uses penalized likelihood estimation on the hazard function to provide a smooth estimate of the baseline hazard function. Bayesian estimation of joint models is also proposed in many R packages, **JMbayes**

(Rizopoulos et al. (2016)) has been used in many biomedical researches (Lawrence Gould et al. (2015)) and **rstanarm** (Muth et al. (2018)) was introduced to provide an intuitive syntax, both using MCMC methods. Finally, **R-INLA** was recently used to fit joint models formulated as latent Gaussian models (Van Niekerk, Bakka, and Rue (2019); Van Niekerk, Bakka, Rue, and Schenk (2019)). In this thesis, we develop the conditional and marginal formulations of the two-part joint model in the **frailtypack** package, written in Fortran 90 within the R function *longiPenal*. This compiled language has the advantage of being highly efficient and useful for optimization problems but it is not adapted for fast operations on vectors or matrices, which are required for numerical integration. Finally, we developed the Bayesian estimation of the conditional two-part joint model using **R-INLA**.

Chapter 3

Two-part joint model for a longitudinal semicontinuous outcome and a terminal event with application to metastatic colorectal cancer data

3.1 Introduction

In this work we developed the two-part model in the context of joint modeling for a longitudinal semicontinuous biomarker and survival times. It was motivated by the GERCOR study, a randomized phase III clinical trial comparing two treatment strategies for metastatic colorectal cancer patients. The focus of this work was to propose a model using directly the tumor size instead of a tumor-based criteria while taking into account the excess of zero values of the biomarker. These zero values correspond to patients with a complete shrinkage of their target lesions. We evaluated the relationship of the tumor size with the event of death and compared the proposed two-part joint model with two alternative strategies. The first one is a standard joint model assuming the semicontinuous biomarker as continuous (i.e., ignoring the zero excess) and the second assumes the measurements of tumor size are subject to left-censoring due to the detectability limit of imaging machines, resulting in the observed zero values (assumed censored). With the new two-part joint model, we assume true zeros (i.e., not censored) can be observed. The two-part approach is preferred when there is interest in what influences the probability to observe a zero value of the biomarker because it evaluates the effect of covariates on both the probability to observe a positive value (i.e., not a zero) and the distribution of positive values. Moreover, it was also of interest to compare the joint effect of zero and positive values of the biomarker with their separate effects on overall survival.

The conditional two-part joint model has been proposed in the literature, using shared or cor-

related random effects to account for the correlation between the binary, continuous and survival part of the model (Liu (2009); Hatfield et al. (2011)). These formulations are however limited to random intercept models and the interpretation of the random effects when they are shared can be difficult as they capture different types of correlations (among repeated measurements and across submodels). Moreover, the association between the biomarker two-part model and the survival model is only based on shared or correlated random effects. In contrast, the model proposed in this chapter includes a vector of random effects specific to the binary and continuous parts and their correlation accounts for the relationship between the binary and continuous parts. They are separately shared in the survival model. We also propose a flexible estimation of the baseline hazard risk function unlike previous articles which considered piecewise constant functions or a Weibull parametric distribution. Moreover, we developed new association structures between the biomarker two-part model and the survival model. These new association structures are time-dependent and therefore require an additional integration step in the likelihood computation. We proposed a general formulation of the two-part joint model fitted using the maximum likelihood estimation obtained with the Levenberg-Marquardt algorithm. In a simulation study we evaluated the performances of the proposed estimation method in terms of bias and coverage probabilities and evaluated the consequences of model misspecification. The GERCOR phase III randomized clinical trial investigated two sequences of treatment:

- Arm A: folinic acid and irinotecan (FOLFIRI) followed by folinic acid and oxaliplatin (FOLFOX6)
- Arm B: FOLFOX6 followed by FOLFIRI

We were interested to compare the effect of treatment captured by both components of the biomarker (i.e., zero and positive values) and the relationship of this treatment effect on the biomarker with survival. Colorectal cancer is among the leading causes of cancer death, approximately half of all patients develop metastatic disease and palliative chemotherapy is often used to prolong survival. In this context, a complete response of target lesions (i.e., $SLD=0$) is not common and only a small subset of the patients will observe such response. In the GERCOR study, 12% of the repeated measurements of the SLD are zeros. We compared the clinical interpretation of the results and describe the strengths and limitations of the two-part joint model compared with alternative approaches. This work has been published in *Biostatistics* (Rustand et al. (2020)) and the two-part joint model is implemented in the function *longiPenal* of the R package **frailtypack**.

3.2 Article



Two-part joint model for a longitudinal semicontinuous marker and a terminal event with application to metastatic colorectal cancer data

DENIS RUSTAND*

Department of Biostatistics, Bordeaux Population Health Research Center, INSERM U1219, 146 Rue Léo Saignat, 33076 Bordeaux, France
denis@rustand.fr

LAURENT BRIOLLAIS

Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital and Dalla Lana School of Public Health (Biostatistics), University of Toronto, 600 University Ave., Ontario M5G 1X5, Canada

CHRISTOPHE TOURNIGAND

Hôpital Henri Mondor, 51 Avenue du Maréchal de Lattre de Tassigny, 94010 Créteil, France

VIRGINIE RONDEAU

Department of Biostatistics, Bordeaux Population Health Research Center, INSERM U1219, 146 Rue Léo Saignat, 33076 Bordeaux, France

SUMMARY

Joint models for a longitudinal biomarker and a terminal event have gained interests for evaluating cancer clinical trials because the tumor evolution reflects directly the state of the disease. A biomarker characterizing the tumor size evolution over time can be highly informative for assessing treatment options and could be taken into account in addition to the survival time. The biomarker often has a semicontinuous distribution, i.e., it is zero inflated and right skewed. An appropriate model is needed for the longitudinal biomarker as well as an association structure with the survival outcome. In this article, we propose a joint model for a longitudinal semicontinuous biomarker and a survival time. The semicontinuous nature of the longitudinal biomarker is specified by a two-part model, which splits its distribution into a binary outcome (first part) represented by the positive versus zero values and a continuous outcome (second part) with the positive values only. Survival times are modeled with a proportional hazards model for which we propose three association structures with the biomarker. Our simulation studies show some bias can arise in the parameter estimates when the semicontinuous nature of the biomarker is ignored, assuming the true model is a two-part model. An application to advanced metastatic colorectal cancer data from the GERCOR study is performed where our two-part model is compared to one-part joint models. Our results show that treatment arm B (FOLFOX6/FOLFIRI) is associated to higher SLD values over time and its positive association with the terminal event leads to an increased risk of death compared to treatment arm A (FOLFIRI/FOLFOX6).

*To whom correspondence should be addressed.

Keywords: Cancer (solid tumors); Joint model; Semicontinuous data; Two-part model; Zero inflation.

1. INTRODUCTION

In solid tumor cancer clinical trials, a biomarker of interest is the sum of the longest diameter (SLD) of target lesions as defined by the Response Evaluation Criteria in Solid Tumors (RECIST). It is often used to categorize the response of patients to treatment and help in the clinical decision-making (Litiere and others, 2017). A limited set of lesions is selected at baseline (target lesions), and those are measured at each follow-up visit of the patient after initiation of a new treatment. Other non-target lesions and new occurring lesions are evaluated qualitatively. Several updates of the RECIST criteria have been proposed over the past years but most of them require a limited number of target lesions to be measured. The SLD provides a longitudinal characterization of the tumor burden. A complete response of target lesions (CRTL) is observed when the SLD reaches a zero value, meaning that all the target lesions disappeared. A partial response (PR) is observed when a significant decline in the SLD is observed. The absence of response to the treatment is qualified as stable disease (SD) and a significant increase in the SLD, a non-target lesion progression or the appearance of one or more new lesions is considered as a progressive disease (PD). In most cancer clinical trials, only a subset of patients reach the CRTL state and interest often lies in the characterization of these complete responders. PR and SD patients are also of interest when the treatment aims at keeping the disease from progressing. Resistance to treatment is often observed, leading to progression of tumor size and an increased risk of death. Besides, some treatments (e.g., immune checkpoint inhibitors) can provide a complete response (CR) to a subset of patients and yield hyperprogressions to another part of the population (Champiat and others, 2017). The analysis of these complex responses requires appropriate methodology to inform clinical decisions.

Joint models have been proposed to fit a survival model jointly with the SLD. The biomarker is often characterized by an excess of zeros due to the subset of patients reaching the CRTL state. A biomarker distribution exhibiting inflated zeros and a continuous distribution of positive values is referred to as semicontinuous. The positive values are often right-skewed. Such distribution arises often in biomedical research when quantifying exposure or measuring symptoms of a disease. Zero-inflated Poisson models have been proposed to handle an excess of zeros with count data. The zero-inflated Poisson regression considers two zero generating processes, the Poisson distribution which generates natural zeros and a binary distribution that accounts for an excess of zeros referred to as structural zeros, see Lambert (1992). Hurdle models consider two data generating processes, a Bernoulli distribution for the zero versus positive counts and the conditional distribution of the positive counts, modeled by a truncated-at-zero count data model (Cragg, 1971; Mullahy, 1986). Duan and others (1983) proposed the two-part model as an extension of the hurdle model for semicontinuous outcomes. It was originally applied to cross-sectional medical cost data. The two-part model decomposes the biomarker distribution into a part with zero values and a part with positive continuous values. A probit or a logit model can be used for the binary outcome and a regression model fits the positive measurements. Olsen and Schafer (2001) and Tooze and others (2002) extended the model to longitudinal data. A generalized linear mixed effects model (GLMM) with a probit or a logit link is used to fit the probability of observing a positive versus zero value while either one of those models: a linear mixed effects (LME) model on the log-transformed biomarker repeated measurements (Liu, 2009), a log-skew-normal distribution, a gamma generalized distribution (which includes the lognormal, gamma, inverse gamma, and Weibull distributions as special cases), links the outcome to the linear predictor and take into account the zero inflation of the continuously distributed values conditional on a positive outcome (Smith and others, 2018). A review of methods to analyze semicontinuous data is proposed in Liu and others (2019).

Liu (2009) proposed a two-part joint model (TPJM) to analyze longitudinal medical cost data jointly with a survival time where the two components of the two-part model are linked through a shared subject-specific random intercept between the binary model and the conditional continuous model. Another subject-specific random intercept captures the residual individual variability of the continuous model. Both independent random effects are shared with a Cox proportional hazards (PH) model for a terminal event. This model therefore includes an association structure between its three components (binary, continuous, and survival) through those shared random effects but is restricted to a random intercept in the two-part model. Besides, the interpretation of the random intercept from the GLMM used to model the binary part is complicated since it captures both the correlation among the repeated measures over time and the correlation among the two components of the semicontinuous model. Dagne (2017) proposed a similar model but with a Bayesian inference approach via a Markov chain Monte Carlo algorithm. It assumes the independence between the two components of the two-part model and uses shared random effects between the two-part model and an accelerated failure time model for the survival outcome. The assumption of independence between the binary and continuous parts can lead to bias in the estimation of both regression coefficients and variance components in the continuous part, see Su and others (2009).

For the joint analysis of the SLD repeated measurements and a terminal event, Król and others (2016) proposed a model with a left-censoring of the biomarker distribution to take into account the excess of zeros in the regression model. It assumes that there is a limit of detection of the biomarker values below which we cannot observe the true value of the biomarker. Manning and others (1987) compared the left-censoring approach to a two-part model to fit semicontinuous outcomes and showed that when the true model is the left-censoring model, then the two-part model yields a good estimate in terms of mean behavior of the outcome.

In this article, we propose a TPJM for a longitudinal semicontinuous biomarker and a terminal event, with correlated random effects between the two components of the two-part model and a Cox PH model for the terminal event. The remainder of the article is structured as follows: in Section 2, we describe the TPJM and the estimation method. In Section 3, we present a simulation study to assess the performance of the TPJM as compared to competing approaches to treat the excess of zeros. An application to colorectal metastatic cancer data from the GERCOR study is proposed in Section 4 to illustrate the interest of our model. We conclude with a discussion in Section 5.

2. METHODS

2.1. Two-part model for the biomarker

Let Y_{ij} denote the biomarker value for subject i ($i = 1, \dots, n$), at visit j ($j = 1, \dots, n_i$). The biomarker distribution is decomposed into a binary outcome $I[Y_{ij} > 0]$ and a positive continuous outcome $Y_{ij}^+ = [Y_{ij} | Y_{ij} > 0]$. A logistic mixed effects model is assumed for the binary outcome and a LME model for the positive continuous outcome. A non-linear transformation $g(\cdot)$ is used to linearize the biomarker evolution over time and correct for right-skewness and heteroscedasticity. The two components are linked through correlated random effects. The two-part model for the biomarker is defined as follows:

$$\begin{cases} \text{Logit}(\text{Prob}(Y_{ij} > 0)) = \mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i & \text{(Binary part),} \\ E[g(Y_{ij}^+)] = \mathbf{X}_{Cij}^\top \boldsymbol{\beta} + \mathbf{Z}_{Cij}^\top \mathbf{b}_i & \text{(Continuous part),} \end{cases}$$

where \mathbf{X}_{Bij} and \mathbf{Z}_{Bij} are vectors of covariates associated with the fixed effect parameters $\boldsymbol{\alpha}$ and the random effects \mathbf{a}_i for the binary part. Similarly, \mathbf{X}_{Cij} and \mathbf{Z}_{Cij} are vectors of covariates associated with the fixed effect parameters $\boldsymbol{\beta}$ and the random effects \mathbf{b}_i . We assume a normal and independently distributed error term (ϵ_{ij}) in the continuous part and the two vectors of random effects follow a multivariate normal

distribution:

$$\begin{bmatrix} \mathbf{a}_i \\ \mathbf{b}_i \end{bmatrix} \sim \text{MVN} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \Sigma_a^2 & \Sigma_{ab} \\ \Sigma_{ab} & \Sigma_b^2 \end{bmatrix} \right).$$

The vectors of correlated subject-specific random effects \mathbf{a}_i and \mathbf{b}_i account for the correlation between repeated measurements within an individual and the correlation between the two components of the model. The logistic regression model includes covariates that represent the effect of an individual's characteristics on the probability of observing a positive versus zero SLD. The continuous part represents the expectation of the SLD given a positive SLD value.

2.2. The two-part joint model

The TPJM considers a two-part model to fit the biomarker evolution over time and a Cox PH model for the terminal event. It is defined as follows:

$$\begin{cases} \text{Logit}(\text{Prob}(Y_{ij} > 0)) = \mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i & \text{(Binary part),} \\ \text{E}[g(Y_{ij}^+)] = \mathbf{X}_{Cij}^\top \boldsymbol{\beta} + \mathbf{Z}_{Cij}^\top \mathbf{b}_i & \text{(Continuous part),} \\ \lambda_i(t) = \lambda_0(t) \exp(\mathbf{X}_i(t)^\top \boldsymbol{\gamma} + \mathbf{h}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{a}_i, \mathbf{b}_i)^\top \boldsymbol{\varphi}) & \text{(Survival part),} \end{cases}$$

where $X_i(t)$ corresponds to time-dependent or time-independent covariates, and $\boldsymbol{\gamma}$ is their effect on the risk of terminal event. The function $\mathbf{h}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{a}_i, \mathbf{b}_i)$ specifies the association between the terminal event and the longitudinal outcome, and $\boldsymbol{\varphi}$ is the corresponding vector of parameters. We propose three association structures, commonly used in joint models. (i) The first one is a ‘‘shared random effects association’’ with a parameter associated with each random effect from the two-part model. The hazard function for the terminal event is assumed to depend on some latent random effects:

$$\lambda_i(t) = \lambda_0(t) \exp(\mathbf{X}_i(t)^\top \boldsymbol{\gamma} + \mathbf{a}_i^\top \boldsymbol{\varphi}_1 + \mathbf{b}_i^\top \boldsymbol{\varphi}_2).$$

This association structure is useful to explore the association between an individual's deviation from the population mean evolution of the biomarker and the risk of terminal event, but it assumes that the association is constant over time and no correlation between the random effects, assumed independent in the Cox PH model.

(2) The second association structure (‘‘current probability of positive value + expected positive value’’) captures the biomarker evolution in the survival model using two parameters, which account for the effect of the current probability of positive value on the risk of event (binary part) and the effect of the linear predictor from the linear regression model on the risk of event (continuous part), respectively. They are both considered time-dependent effects in the Cox PH model. The binary model can capture the effect of the probability of having a CRTL on the risk of the terminal event while the continuous model accounts for other types of responses. This includes partial responders whose SLD has an initial decline following treatment initiation and then stabilizes without reaching a CRTL or patients who develop a resistance to treatment as assessed by a progression of the SLD. This association structure takes into account the correlation between the random intercept and the slope of the biomarker for each part of the model but assumes no correlation between the binary and continuous parts in the survival model,

$$\lambda_i(t) = \lambda_0(t) \exp(\mathbf{X}_i(t)^\top \boldsymbol{\gamma} + \text{Prob}(Y_{ij} > 0) \varphi_1 + \text{E}[g(Y_{ij}^+)] \varphi_2),$$

where $\text{Prob}(Y_{ij} > 0) = \exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i) (1 + \exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i))^{-1}$ and $\text{E}[g(Y_{ij}^+)] = \mathbf{X}_{Cij}^\top \boldsymbol{\beta} + \mathbf{Z}_{Cij}^\top \mathbf{b}_i$.

This second association structure allows to model the hazard of the terminal event conditional on a CRTL or a PR and also provides information on covariates that affect either response.

(3) The last association structure (“current value”), includes the subject-specific predictor of the expected longitudinal outcome from the two-part model into the Cox PH model, defined as

$$E[Y_{ij}] = \text{Prob}(Y_{ij} > 0)E[Y_{ij}|Y_{ij} > 0] + \text{Prob}(Y_{ij} = 0)E[Y_{ij}|Y_{ij} = 0].$$

The second term of the equation is 0 because $E[Y_{ij}|Y_{ij} = 0] = 0$. This association corresponds to the current value of the biomarker, commonly used in standard joint models. When a non-linear function $g(\cdot)$ is applied, the association corresponds to the expected value on the transformed scale

$$\lambda_i(t) = \lambda_0(t) \exp(\mathbf{X}_i(t)^\top \boldsymbol{\gamma} + \text{Prob}(Y_{ij} > 0)E[g(Y_{ij}^+)] \varphi_1).$$

The current value association is useful to explore the association between the expected biomarker value at the time of the terminal event and the risk of the terminal event. This association structure assumes that the expected current value of the longitudinal biomarker at time t is predictive of the risk of event at that particular time t . The non-linear transformation $g(\cdot)$ in the continuous part can handle the positive biomarker values that exhibit skewness.

2.3. Estimation method

We can derive the full likelihood of the TPJM, which combines the contributions from the binary part with a Bernoulli density, the continuous part with a Gaussian density, and the survival part which requires approximation of the baseline hazard

$$\begin{aligned} L_i(\Theta) &= \int_{a_i} \int_{b_i} \left[\frac{1}{(\sqrt{2\pi}\sigma_\epsilon^2)^{n_i}} \prod_{j=1}^{n_i} \exp\left(-\frac{(g(Y_{ij}^+) - \mathbf{X}_{Cij}^\top \boldsymbol{\beta} - \mathbf{Z}_{Cij}^\top \mathbf{b}_i)^2}{2\sigma_\epsilon^2}\right) \right. \\ &\quad \times (\exp[\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i])^{U_{ij}} + \left(1 - \frac{\exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)}{1 + \exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)}\right) \\ &\quad \left. \times \lambda_i(T_i|\Theta_i)^{\delta_i} \exp\left(-\int_0^{T_i} \lambda_i(t|\Theta_i) dt\right) p(\mathbf{a}_i, \mathbf{b}_i) \right] db_i da_i, \end{aligned}$$

with $U_{ij} = I(Y_{ij} > 0)$, $\Theta_i = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{a}_i, \mathbf{b}_i, \boldsymbol{\gamma}, \boldsymbol{\varphi})$.

The baseline hazard risk function is approximated by m cubic M-splines with Q knots. They are nonnegative functions that facilitate the calculation of the integrals and derivatives in the likelihood expression. We propose to penalize the log-likelihood in order to obtain smooth estimation of the baseline hazard function

$$pl(\Theta) = l(\Theta) - \kappa \int_0^\infty \lambda_0''(t)^2 dt,$$

where $l(\Theta) = \sum_{i=1}^n \log(L_i(\Theta))$ and κ a smoothing parameter chosen using an approximate cross-validation criterion from a separate Cox model. We propose to use the Levenberg-Marquardt algorithm to maximize this penalized log-likelihood (Marquardt, 1963). The integration over the random effects is performed by Monte Carlo integration with 5000 integration points in the real data application and

1000 integration points in the simulation studies. Standard errors are calculated from the inverse Hessian matrix of the penalized log-likelihood, which is directly available from our optimization algorithm. The TPJM with time-dependent covariates in the Cox PH model (“current probability of positive value + expected positive value” and “current value”) requires the current value from the two-part model (binary, continuous) between visit times in order to perform the numerical integration for the cumulative hazard function in the likelihood. This integral has no analytical solution and is approximated numerically with a Gauss-Kronrod quadrature with 15 points.

3. SIMULATION STUDY

3.1. Objectives

We conduct simulation studies to evaluate the performances of the TPJM in terms of efficiency and bias and compare it with alternative approaches. We compare the TPJM to two alternative approaches: a standard one-part joint model (OPJM), which considers the biomarker as a continuous variable modeled via a LME model and does not account for the excess of zeros and another OPJM, which considers that the biomarker distribution is left-censored. The censoring threshold is defined as the smallest positive value observed.

3.2. Methods

We propose four simulations scenarios, for each scenario 300 datasets were generated and for each dataset 300 individuals were included. The data simulated under the three first scenarios assumed that the TPJM is the true model while the last scenario considers it is the left-censoring OPJM. The current value association structure is assumed to generate the data with a random intercept and a fixed slope in the binary part and a random intercept plus a random slope in the continuous part of the biomarker. A binary covariate corresponding to the treatment effect (trt_i) is included in each submodel and generated from a Bernoulli distribution with $p = 0.5$ and with a time-interaction within each component of the two-part model. The model for data generation is given by

$$\begin{cases} \text{Logit}[\text{Prob}(Y_{ij} > 0)] = \alpha_0 + a_i + \alpha_1 \cdot \text{time}_j + \alpha_2 \cdot trt_i + \alpha_3 \cdot \text{time}_j \cdot trt_i, \\ Y_{ij}^+ = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) \cdot \text{time}_j + \beta_2 \cdot trt_i + \beta_3 \cdot \text{time}_j \cdot trt_i + \varepsilon_{ij}, \\ \lambda_i(t|Y_{ij}) = \lambda_0(t) \exp(\gamma \cdot trt_i + \varphi \cdot E[Y_{ij}]), \end{cases}$$

$$\begin{bmatrix} a_i \\ b_{0i} \\ b_{1i} \end{bmatrix} \sim \text{MVN} \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_a^2 & \rho \sigma_a \sigma_{b_0} & \rho \sigma_a \sigma_{b_1} \\ \rho \sigma_a \sigma_{b_0} & \sigma_{b_0}^2 & \rho \sigma_{b_0} \sigma_{b_1} \\ \rho \sigma_a \sigma_{b_1} & \rho \sigma_{b_0} \sigma_{b_1} & \sigma_{b_1}^2 \end{bmatrix} \right).$$

The longitudinal measurements are directly generated assuming a Gaussian distribution without any non-linear transformation. The first scenario mimicks several aspects of our real data with similar treatment effects in the three components and 15% of zero measurements overall. In the second scenario, the treatment is associated with higher odds of observing a zero value of the biomarker over time than in the real data. The difference between the first and the second scenario is that the sign of α_3 changes from positive to negative and the value of the linear slope, α_1 is reduced by one unit in order to get the same proportion of zero measurements. This approach is motivated by the complex patterns of responses observed in clinical trials, where a subset of patients could respond well to treatment until reaching a CR while other patients could experiment adverse treatment effect. The third scenario considers a larger proportion of zeros for the biomarker by changing both the linear slope α_1 and the effect of treatment over time α_3

Table 1. Comparison of OPJM and TPJM of a longitudinal semicontinuous biomarker and a terminal event with current value association. The true model is the TPJM, 300 datasets are generated with 15% ($SD = 2\%$) zeros in the biomarker distribution on average.

	Variable	Standard OPJM Est. [†] (SD [‡]) [CP [§]]	Left-censoring OPJM Est. (SD) [CP]	TPJM Est. (SD) [CP]
Binary part				
Intercept	$\alpha_0 = 6$			6.05 (0.59) [92%]
Time	$\alpha_1 = -4$			-3.98 (0.45) [94%]
Treatment	$\alpha_2 = -1$			-0.94 (0.52) [93%]
Time:treatment	$\alpha_3 = 1$			1.05 (0.57) [93%]
Continuous part				
Intercept	$\beta_0 = 4$	4.01 (0.07)	4.08 (0.07)	3.99 (0.18) [94%]
Time	$\beta_1 = -0.5$	-1.05 (0.17)	-1.25 (0.22)	-0.51 (0.12) [90%]
Treatment	$\beta_2 = 1$	0.91 (0.12)	0.87 (0.11)	1.01 (0.10) [91%]
Time:treatment	$\beta_3 = 1$	0.65 (0.33)	0.80 (0.34)	1.00 (0.17) [92%]
Residual S.E.	$\sigma_\epsilon = 0.5$	1.06 (0.05)	1.03 (0.10)	0.50 (0.02) [93%]
Survival part				
Treatment	$\gamma = 0.3$	0.32 (0.17) [96%]	0.35 (0.16) [92%]	0.34 (0.16) [95%]
Association	$\varphi = 0.3$	0.32 (0.07) [95%]	0.29 (0.06) [92%]	0.31 (0.07) [95%]
Random effects				
Intercept (continuous part)	$\sigma_{b_0} = 0.75$	0.74 (0.07)	0.68 (0.06)	0.71 (0.04)
Slope (continuous part)	$\sigma_{b_1} = 0.75$	1.18 (0.14)	1.48 (0.19)	0.47 (0.19)
Intercept (binary part)	$\sigma_a = 2$			2.13 (0.26)
	$corr_{b_0b_1} = -0.20$	0.18 (0.17)	0.15 (0.14)	-0.37 (0.31)
	$corr_{ab_0} = 0.20$			0.31 (0.15)
	$corr_{ab_1} = 0.70$			0.40 (0.39)
Convergence rate		99%	98%	98%

[†] Mean of parameter estimates.

[‡] Standard deviation from the mean.

[§] Coverage probability

in the binary part. This yields 35% of zeros with treated patients having a higher chance of CRTL but higher biomarker values in the continuous part over time. The last scenario was considered to evaluate the performance of the TPJM when the proportion of zeros result from a higher limit of detection imposed to the biomarker. These four scenarios entail several treatment effects and proportions of zeros for the biomarker. The performances of the analysis models are evaluated in terms of mean parameter estimate and coverage probabilities of the parameter estimates. We consider a maximum follow-up period of 4 years for each patient and a 80% death rate, as observed in our real data. The true value of parameters for data generation is given in the second column of Tables 1 to 4.

The survival times conditional on the biomarker time-dependent values are generated using the R package PermAlgo (Sylvestre and Abrahamowicz, 2008), which requires to specify the biomarker trajectory for the entire follow-up period among all patients, and then a permutation algorithm simulates the survival times based on these trajectories in order to get the correct effect of the time-dependent biomarker on survival times. The permutation algorithm also generates random censoring times, which makes the number of biomarker measurements variable among individuals (from 1 to 30 repeated measurements).

Table 2. Comparison of OPJM and TPJMs with opposite treatment effects between the odds of zeros (binary part) and the expected value among positives (continuous part). The true model is the TPJM, 300 datasets are generated with 14% ($SD = 2\%$) zeros in the biomarker distribution on average.

	Variable	Standard OPJM Est.† (SD‡) [CP§]	Left-censoring OPJM Est. (SD) [CP]	TPJM Est. (SD) [CP]
Binary part				
Intercept	$\alpha_0 = 6$			6.05 (0.60) [92%]
Time	$\alpha_1 = -3$			-3.01 (0.36) [94%]
Treatment	$\alpha_2 = -1$			-0.93 (0.52) [93%]
Time:treatment	$\alpha_3 = -1$			-0.92 (0.52) [94%]
Continuous part				
Intercept	$\beta_0 = 4$	3.99 (0.07)	4.02 (0.07)	3.99 (0.17) [92%]
Time	$\beta_1 = -0.5$	-0.76 (0.16)	-0.89 (0.19)	-0.52 (0.12) [86%]
Treatment	$\beta_2 = 1$	0.97 (0.12)	0.99 (0.12)	1.01 (0.11) [91%]
Time:treatment	$\beta_3 = 1$	-0.15 (0.34)	-0.31 (0.39)	1.02 (0.17) [92%]
Residual S.E.	$\sigma_\epsilon = 0.5$	1.06 (0.05)	1.08 (0.09)	0.50 (0.02) [95%]
Survival part				
Treatment	$\gamma = 0.3$	0.34 (0.16) [90%]	0.42 (0.16) [84%]	0.35 (0.16) [92%]
Association	$\varphi = 0.3$	0.33 (0.07) [92%]	0.27 (0.06) [88%]	0.30 (0.07) [95%]
Random effects				
Continuous intercept	$\sigma_{b_0} = 0.75$	0.74 (0.07)	0.68 (0.07)	0.71 (0.04)
Continuous slope	$\sigma_{b_1} = 0.75$	1.21 (0.14)	1.58 (0.20)	0.48 (0.19)
Binary intercept	$\sigma_a = 2$			2.13 (0.25)
	$corr_{b_0b_1} = -0.20$	0.16 (0.18)	0.13 (0.15)	-0.35 (0.31)
	$corr_{ab_0} = 0.20$			0.30 (0.18)
	$corr_{ab_1} = 0.70$			0.39 (0.43)
Convergence rate		99%	99%	96%

†Mean of parameter estimates.

‡Standard deviation from the mean.

§Coverage probability

The parameters from the longitudinal continuous model cannot be compared between the one-part and two-part models because the former does include the excess of zeros in the biomarker distribution. Therefore, the coverage probabilities are not provided for these parameters as their true value is unknown. The focus of our comparison is on the covariate effects affecting the survival part since all the models are measuring a direct effect of treatment (γ) and the biomarker effect (φ) (which itself includes an indirect treatment effect captured in the biomarker model) on the risk of terminal event.

3.3. Results

Results from the simulations are presented in Tables 1 to 4. The TPJM performs well across the first three scenarios in terms of bias. The coverage probabilities for the fixed regression coefficients are close to 95%, with a small empirical standard deviation reflecting the good precision of the estimations. The difference in the biomarker fit between the OPJM and TPJM concerns essentially the time-related parameters due to the zeros appearing during the follow-up. In the first scenario, the true value of the slope ($\beta_1 = -0.5$) and the treatment interaction with the slope ($\beta_3 = 1$) in the continuous part are biased downwards for the standard OPJM ($\hat{\beta}_1 = -1.05$, $SD = 0.17$) and ($\hat{\beta}_3 = 0.65$, $SD = 0.63$) and the left-censoring OPJM ($\hat{\beta}_1 = -1.25$, $SD = 0.22$) and ($\hat{\beta}_3 = 0.80$, $SD = 0.34$) due to the inclusion of zero values in the LME

Table 3. Comparison of OPJM and TPJM with an increased zero-rate. The true model is the TPJM, 300 datasets are generated with 35% (SD = 2%) zeros in the biomarker distribution on average.

	Variable	Standard OPJM Est. [†] (SD [‡]) [CP [§]]	Left-censoring OPJM Est. (SD) [CP]	TPJM Est. (SD) [CP]
Binary part				
Intercept	$\alpha_0 = 6$			5.92 (0.64) [93%]
Time	$\alpha_1 = -8$			-7.85 (0.87) [90%]
Treatment	$\alpha_2 = -1$			-0.83 (0.51) [94%]
Time:treatment	$\alpha_3 = -4$			-4.11 (1.06) [94%]
Continuous part				
Intercept	$\beta_0 = 4$	4.06 (0.08)	4.24 (0.08)	3.99 (0.15) [93%]
Time	$\beta_1 = -0.5$	-2.44 (0.26)	-3.10 (0.43)	-0.50 (0.15) [93%]
Treatment	$\beta_2 = 1$	0.78 (0.15)	1.09 (0.12)	1.01 (0.11) [91%]
Time:treatment	$\beta_3 = 1$	-1.85 (0.38)	-3.82 (0.69)	1.02 (0.23) [95%]
Residual S.E.	$\sigma_\epsilon = 0.5$	1.30 (0.04)	1.24 (0.10)	0.50 (0.02) [95%]
Survival part				
Treatment	$\gamma = 0.3$	0.31 (0.13) [96%]	0.46 (0.14) [80%]	0.34 (0.13) [94%]
Association	$\varphi = 0.3$	0.16 (0.09) [66%]	0.14 (0.05) [11%]	0.32 (0.07) [93%]
Random effects				
Intercept (continuous part)	$\sigma_{b_0} = 0.75$	0.94 (0.07)	0.58 (0.09)	0.71 (0.04)
Slope (continuous part)	$\sigma_{b_1} = 0.75$	1.41 (0.29)	2.65 (0.47)	0.44 (0.16)
Intercept (binary part)	$\sigma_a = 2$			2.06 (0.26)
	$corr_{b_0b_1} = -0.20$	-0.34 (0.16)	-0.32 (0.19)	-0.37 (0.32)
	$corr_{ab_0} = 0.20$			0.26 (0.25)
	$corr_{ab_1} = 0.70$			0.33 (0.46)
Convergence rate		99%	97%	98%

[†]Mean of parameter estimates.

[‡]Standard deviation from the mean.

[§]Coverage probability

model. In the second scenario, the negative treatment effect over time on the odds of zero values ($\alpha_3 = -1$) balances out the positive effect on the value among positives ($\beta_3 = 1$) and no treatment with time interaction was found by the standard OPJM ($\hat{\beta}_3 = -0.15$, SD = 0.34) nor by the left-censoring OPJM ($\hat{\beta}_3 = -0.31$, SD = 0.39). In the third scenario, with on average 35% of zero biomarker measurements, the interaction between treatment and time is found significantly negative with the standard OPJM ($\hat{\beta}_3 = -1.85$, SD = 0.38), where the true value is $\beta_3 = 1$. The left-censoring OPJM finds an even steeper negative slope because the censoring hypothesis allows for hypothetical negative values in the biomarker distribution ($\hat{\beta}_3 = -3.82$, SD = 0.69). The standard and left-censoring OPJM are unable to capture the positive treatment effect over time in the LME model under the third scenario because of the large excess of zeros. They conclude that the treatment is associated with an overall decrease of the biomarker value over time among the treated patients. However, the TPJM is able to recover this parameter value with good precision ($\hat{\beta}_3 = 1.02$, SD = 0.23, CP = 95%). With the OPJMs, the effect of treatment on the binary part is captured through the continuous model only and they fail to recover the adverse effect of treatment on the positive continuous measurements with an overall treatment effect on the slope estimated as significantly negative. The random intercept standard deviation from the continuous part is properly captured overall and the slope is systematically overestimated with one-part models, again due to the

Table 4. Comparison of OPJM and TPJM when the left-censoring OPJM is the true model, 300 datasets are generated with 13% ($SD = 2\%$) zeros in the biomarker distribution on average.

	Variable	Standard OPJM Est. [†] (SD [‡]) [CP [§]]	Left-censoring OPJM Est. (SD) [CP]	TPJM Est. (SD) [CP]
Binary part				
Intercept	α_0			3.87 (0.41)
Time	α_1			-1.34 (0.33)
Treatment	α_2			-0.93 (0.38)
Time:treatment	α_3			0.78 (0.44)
Continuous part				
Intercept	$\beta_0 = 2$	2.03 (0.07) [91%]	2.02 (0.08) [93%]	2.08 (0.14)
Time	$\beta_1 = -0.5$	-0.40 (0.12) [83%]	-0.53 (0.14) [92%]	-0.34 (0.17)
Treatment	$\beta_2 = -0.5$	-0.44 (0.10) [92%]	-0.49 (0.11) [94%]	-0.38 (0.13)
Time:treatment	$\beta_3 = 0.5$	0.41 (0.16) [89%]	0.49 (0.20) [94%]	0.38 (0.16)
Residual S.E.	$\sigma_\epsilon = 1$	0.90 (0.02) [0%]	1.00 (0.02) [97%]	0.89 (0.02)
Survival part				
Treatment	$\gamma = 0.3$	0.32 (0.14) [93%]	0.31 (0.14) [94%]	0.32 (0.14) [93%]
Association	$\varphi = 0.3$	0.38 (0.12) [89%]	0.32 (0.10) [95%]	0.38 (0.12) [88%]
Random effects				
Intercept (continuous part)	$\sigma_{b_0} = 0.75$	0.67 (0.04)	0.71 (0.05)	0.57 (0.05)
Slope (continuous part)	$\sigma_{b_1} = 0.75$	0.52 (0.11)	0.73 (0.11)	1.22 (0.19)
Intercept (binary part)	σ_a			0.99 (0.26)
	$corr_{b_0 b_1} = -0.20$	-0.26 (0.18)	-0.18 (0.12)	-0.08 (0.10)
	$corr_{ab_0}$			0.31 (0.28)
	$corr_{ab_1}$			0.62 (0.56)
Convergence rate		100%	100%	99%

[†]Mean of parameter estimates.

[‡]Standard deviation from the mean.

[§]Coverage probability.

inclusion of zero-values and it is slightly underestimated with the TPJM. One-part models cannot recover the slightly negative correlation between the intercept and slope in the continuous part for the two first scenarios, and the TPJM properly estimates the correlation structure of the random effects but with a high variability among models. In the last scenario, where the true model for data generation is a left-censoring OPJM, the standard OPJM provides slightly biased parameter estimations for the time trend ($\hat{\beta}_1 = -0.40$, $SD = 0.12$, $CP = 83\%$) and the treatment interaction with time ($\hat{\beta}_3 = 0.41$, $SD = 0.16$, $CP = 89\%$). The true values are $\beta_1 = -0.5$ and $\beta_3 = 0.5$. The TPJM finds slightly lower values for these two parameters ($\hat{\beta}_1 = -0.34$, $SD = 0.17$) and ($\hat{\beta}_3 = 0.38$, $SD = 0.16$) because the true value of these parameters are not known under the TPJM.

The parameters of interest, $\gamma = 0.30$ and $\varphi = 0.30$, respectively the treatment effect and the association between the biomarker and terminal event, are correctly estimated in the first and second scenarios by all three models, with the exception of the treatment effect in the second scenario for the left-censoring OPJM ($\hat{\gamma} = 0.42$, $SD = 0.16$, $CP = 84\%$) and a slightly decreased coverage probabilities for the association ($\hat{\varphi} = 0.27$, $SD = 0.06$, $CP = 88\%$). The association parameter is underestimated with the standard ($\hat{\varphi} = 0.16$, $SD = 0.09$) and left-censoring ($\hat{\varphi} = 0.14$, $SD = 0.05$) OPJMs under the third scenario. The coverage probabilities are low for this parameter (66% with the standard approach and 11% with the left-censoring

OPJM), and the direct treatment effect for the left-censoring OPJM is overestimated ($\hat{\gamma} = 0.46$). In the last scenario, the direct effect of treatment is unbiased for the three models and the effect of the current value of the biomarker tend to be overestimated with the standard OPJM ($\hat{\phi} = 0.38$, SD = 0.12, CP = 89%) and the TPJM ($\hat{\phi} = 0.38$, SD = 0.12, CP = 88%).

3.4. Discussion

The simulation studies show that the TPJM performs well regardless of the percentage of zero values. The sign of the treatment effect in the binary and continuous parts can be opposite without affecting the quality of the fit. We also show that an excess of true zeros can bias the treatment effect downwards when estimated from a one-part model for the biomarker. This bias affects both the biomarker and survival components of a joint model and therefore the excess of zeros should be considered when deciding a model for analysis.

4. APPLICATION TO METASTATIC COLORECTAL CANCER DATASET

4.1. GERCOR trial

Our methodological development was motivated by the analysis of the GERCOR study, a randomized clinical trial investigating two treatment strategies that included a total of 220 patients with metastatic colorectal cancer. The reference strategy (arm A) corresponds to FOLFIRI (irinotecan) followed by FOLFOX6 (oxaliplatin) while arm B involves the reverse sequence. Patients were randomly assigned from December 1997 to September 1999, and the date chose to assess overall survival was August 30, 2002. We refer the reader to the original report from the GERCOR study for more details ([Tournigand and others, 2004](#)). Complete data are available on 205 individuals for data analysis. Among them, 165 (80%) died during the follow-up. The median OS is similar in both arms of the trial (21.6 months). There are 1475 repeated measurements for the biomarker, 174 of which are zero values (12%). Our model uses death as the terminal event and the repeated SLD measurements (in centimeters) as the semicontinuous biomarker. Additional baseline covariates collected at the start of the study are also included. Figure S1 of the supplementary material available at Biostatistics online sketches the main difference between the OPJM and a TPJM in modeling the treatment effect. Figure S2 of the supplementary material available at Biostatistics online shows a few individuals' biomarker trajectories with respect to the treatment arm. Despite the randomization, the percentage of males was higher in treatment arm B and age >65 years with a percentage slightly higher in arm B. The variables age and sex were not found associated to any of the three components of the TPJM and thus not included in the final analysis.

4.2. Data analysis

4.2.1. Analysis models. We applied the TPJM and two competing models to the data: a standard OPJM, which assumes a simple continuous distribution for the biomarker modeled with a LME model and a left-censoring OPJM, for which the threshold was defined as the smallest positive value observed in the GERCOR trial (i.e., 0.5 mm). We propose to use the current value of the biomarker for the association between the survival and the longitudinal models, so that we can compare the survival function conditional on the biomarker across models. We also propose an alternative TPJM with a separate association for the probability of observing a positive biomarker value from the binary part and the current value of the biomarker for the continuous part. Therefore, we can evaluate whether the binary part is significantly associated with the risk of terminal event, independently from the positive biomarker values and subsequently whether these positive observations are associated with the risk of terminal event. We propose a log-transformation of the biomarker to handle its right-skewness with a 1 unit shift in

order to get a fair comparison between all the fitted models. We used a global backward selection procedure for each component of the model to select the covariates to include in the final joint model. We chose among the following clinical variables of interest measured at baseline: sex (yes/no); age (<60/60-69/ \geq 70); WHO performance status (0/1/2); primary site (colon/rectum/multiple); previous surgery (no surgery/curative/palliative); previous adjuvant chemotherapy (yes/no); previous adjuvant radiotherapy (yes/no); metastases (metachronous/synchronous); number of metastatic sites (1/ $>$ 1); liver metastatic site (yes/no); lung metastatic site (yes/no), patients characteristics are described in Table S1 of the supplementary material available at Biostatistics online. The baseline hazard function is approximated with cubic M-splines with five knots and the penalization term $\kappa = 0.08$, chosen by cross-validation from an univariate survival model.

We consider a random-intercept logistic model for the binary part of the biomarker and a LME model with both random intercept and random slope for the log-transformed continuous outcome. Parameter estimates for the binary component of the model are subject-specific estimates, $\exp(\alpha_k)$ represents the subject-specific odds ratio of observing a positive outcome associated with a one-unit increase in the k th covariate. Parameters in the continuous part gives the effect of covariates on the transformed outcome among the subset of individuals for whom a positive value of the biomarker is observed over time. The aim of the application is: (i) to evaluate the sensitivity of treatment effect estimation with respect to model assumptions. (ii) The impact of model assumptions on overall survival estimate and its confidence intervals. (iii) To assess the influence of the zero part (CRTL) and non-zero part (PR, SD, PD) on overall survival.

4.3. Results

4.3.1. Biomarker component (SLD). Results from the application are presented in Table 5, and details of covariates effect are available in Table S3 of the supplementary material available at Biostatistics online. Despite the fact that the trial was randomized, we found a treatment effect at baseline in both the binary part ($\hat{\alpha}_2 = -1.34$, SD = 0.72) and the continuous part ($\hat{\beta}_2 = -0.24$, SD = 0.08) of the TPJM. We tried fitting the TPJM with B-splines in order to allow for more flexibility in the biomarker evolution over time but the treatment effect at baseline remained significant. An ANOVA test confirmed a slightly significant difference in means at baseline ($p = 0.04$). The parameter estimates in the binary and continuous parts are nearly identical for the TPJM with current value association and the TPJM with separate association. Therefore, we only describe results from the current value association model. The binary part provides information on the probability of observing a positive SLD. The baseline value is positive ($\hat{\alpha}_0 = 5.44$, SE = 0.75) which corresponds to a probability of positive value at baseline close to 1. This is expected since the eligibility criteria for inclusion was at least one measurable lesion > 2 cm. The treatment arm A is associated to a significant decrease in the odds of positive SLD over time ($\hat{\alpha}_1 = -2.22$, SE = 0.39), but this effect decreases with treatment arm B as the interaction term was found positive but not significant ($\hat{\alpha}_3 = 0.43$, SE = 0.45). Treatment arm A is therefore associated to a slightly higher probability of CRTL over time. In the continuous part (i.e., positive SLD), the treatment arm A is associated to a significant decrease of the SLD over time ($\hat{\beta}_1 = -0.31$, SE = 0.06) and the treatment arm B is associated to a SD as the interaction term cancels out the slope parameter ($\hat{\beta}_3 = 0.30$, SE = 0.09). In the OPJMs, the treatment arm A is associated to a stronger decrease of the SLD over time compared to the TPJM ($\hat{\beta}_1 = -0.40$, SE = 0.08 for the standard OPJM and $\hat{\beta}_1 = -0.42$, SE = 0.10 for the left-censoring OPJM), likely because these two models include zeros in the continuous biomarker distribution. Under the standard OPJM, treatment arm B is associated to a similar trend of the SLD over time ($\hat{\beta}_3 = 0.29$, SE = 0.11) to the TPJM but to a reduced value under the left-censoring OPJM ($\hat{\beta}_3 = 0.11$, SE = 0.15). The random intercept in the binary part of the TPJM captures some variability at the individual level for the probability of observing a positive SLD ($\hat{\sigma}_a = 2.40$). The random intercept in the continuous part corresponds to the individual variability of the SLD at baseline, $\hat{\sigma}_{b_0} = 0.59 - 0.62$ across all four models. The random slope accounts for individual

Table 5. Application to advanced colorectal cancer data. The binary model is adjusted on the WHO performance status (0/1/2), lung metastatic site (yes/no), previous adjuvant radiotherapy (yes/no), the continuous model is adjusted on the WHO performance status, metastases (metachronous/synchronous), previous surgery (no surgery/curative/palliative), previous adjuvant radiotherapy, and the survival model is adjusted on WHO performance status, metastases, previous surgery, and previous adjuvant radiotherapy.

Variable (association)	Standard OPJM	Left-censoring OPJM	TPJM	TPJM
	(current value)	(current value)	(current value)	(current probability, current value ⁽⁺⁾)
	Est. (SE)	Est. (SE)	Est. (SE)	Est. (SE)
Binary part (SLD > 0 versus SLD = 0)				
Intercept			5.44**** (0.75)	5.41**** (0.74)
Time (years)			-2.22**** (0.39)	-2.29**** (0.36)
Treatment (B/A)			-1.34* (0.72)	-1.31* (0.72)
Time:treatment (B/A)			0.43 (0.45)	0.51 (0.42)
Continuous part (SLD>0)				
Intercept	2.06**** (0.14)	2.10**** (0.14)	2.10**** (0.14)	2.10**** (0.13)
Time (years)	-0.40**** (0.08)	-0.42**** (0.10)	-0.31**** (0.06)	-0.31**** (0.06)
Treatment (B/A)	-0.28*** (0.09)	-0.27*** (0.10)	-0.24*** (0.08)	-0.24*** (0.08)
Time:treatment (B/A)	0.29** (0.11)	0.11 (0.15)	0.30**** (0.09)	0.29**** (0.09)
Residual S.E.	0.43 (0.01)	0.45 (0.01)	0.31 (0.01)	0.31 (0.01)
Death risk				
Treatment (B/A)	0.20 (0.17)	0.26 (0.17)	0.25 (0.17)	0.31* (0.18)
	[HR = 1.22]	[HR = 1.30]	[HR = 1.28]	[HR = 1.36]
Association				
$E[g(Y_{ij})]$	0.62**** (0.10)	0.46**** (0.08)	0.81**** (0.12)	
$P(Y_{ij} > 0)$				1.80*** (0.61)
$E[g(Y_{ij}^+)]$				0.45**** (0.16)
Random effects				
Intercept (continuous part, σ_{b_0})	0.62	0.60	0.59	0.59
Slope (continuous part, σ_{b_1})	0.61	0.94	1.51	1.58
Intercept (binary part, σ_a)			2.40	2.33
$corr_{b_0b_1}$	-0.05	-0.09	-0.08	-0.08
$corr_{ab_0}$			0.17	0.18
$corr_{ab_1}$			0.78	0.76

**** $p < 0.001$, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

deviation from the mean slope and is higher under the TPJM ($\hat{\sigma}_{b_1} = 1.51$) than under the left-censoring OPJM ($\hat{\sigma}_{b_1} = 0.94$) or the standard OPJM ($\hat{\sigma}_{b_1} = 0.61$). It means that the TPJM captures more variability over time from unobserved covariates. Finally, the positive correlation between the binary intercept and the continuous slope is high in the TPJM, i.e., patients who are more likely to observe a positive SLD value also tend to have steeper increase in the SLD over time.

4.3.2. *Survival component.* Unlike our simulation results (first scenario), we do observe some differences here between the OPJMs and TPJMs for the effect of the current SLD value on the risk of death ($\hat{\varphi} = 0.62$, SE = 0.10 for the standard OPJM, $\hat{\varphi} = 0.46$, SE = 0.08 for the left-censoring OPJM, $\hat{\varphi} = 0.81$,

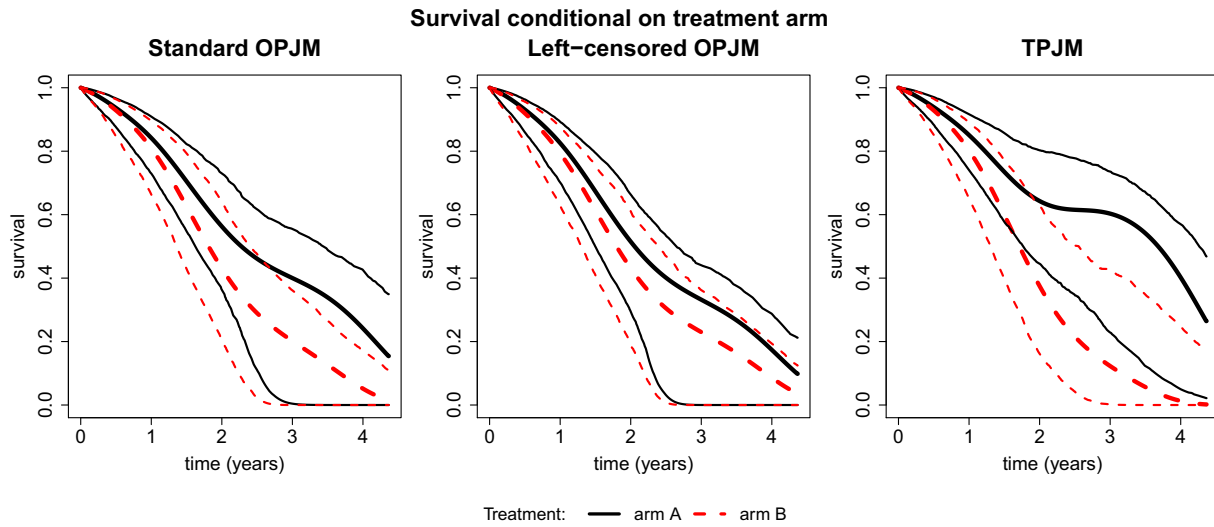


Fig. 1. Survival curves from the application (current value association structure) according to treatment arm. Estimations are based on models with current value association structure and 95% confidence intervals are provided using Monte Carlo method with 1000 curves.

SE = 0.12 for the TPJM). The difference observed is similar to that of the third simulation scenario with an increased zero-rate, where the association parameter was found biased downwards for the standard and left-censoring OPJM. The hazard ratio for the direct effect of treatment arm B versus A on the risk of death ranges from 1.22 to 1.36 across the four models, which corresponds to an increase in the risk of death from 22% to 36%. The TPJM with a separate association structure shows that the association with the biomarker and the risk of death depends significantly on both the probability of positive SLD ($\hat{\varphi}_1 = 1.80$, SE = 0.61) and the expected positive SLD value ($\hat{\varphi}_2 = 0.45$, SE = 0.16). The hazard ratio for the direct effect of treatment arm B on the risk of death is slightly higher under this model ($\exp(\hat{\gamma}) = 1.36$, SE = 0.18) compared to the other three. We can estimate from our fitted models the survival curves conditional on the treatment arm (Figure 1) and conditional on the biomarker (Figure 2). The TPJM yields a stronger discrimination between the two treatment lines although not significant, in agreement with the GERCOR initial trial. The second plot illustrates the influence of the biomarker trajectory on the overall survival curve. Again the TPJM provides a stronger separation of the different profiles of response to treatment, especially when a CRTL response is observed. The results of the models fitted with the shared random effects association (“shared random effects association”) are presented in Table S2 of the supplementary material available at Biostatistics online. The fit is similar to the models with current value association, except that the fixed treatment effect on the SLD is not shared in the survival model. This leads to an increased hazard ratio for the direct effect of treatment (arm B versus arm A) on the risk of death with the TPJM ($\exp(\hat{\gamma}) = 1.30$) compared to the standard ($\exp(\hat{\gamma}) = 1.15$) and left-censored ($\exp(\hat{\gamma}) = 1.13$) OPJM. The association between the SLD and risk of death captured by the shared random effects is significant with the standard and left-censoring OPJM but not significant with the TPJM. This suggests that that some of the direct treatment effect on death is most likely captured by the random effects in the two OPJMs. Our final conclusion about treatment effect from all our models is that overall, the treatment arm B is associated to a higher value of the SLD over time, and the positive association parameter leads to an increased risk of death compared to treatment arm A.

4.3.3. *Conclusion.* The different forms of the association structure allow to answer different questions of interest. With the shared random-effects association, we can evaluate the effect on the risk of death of an

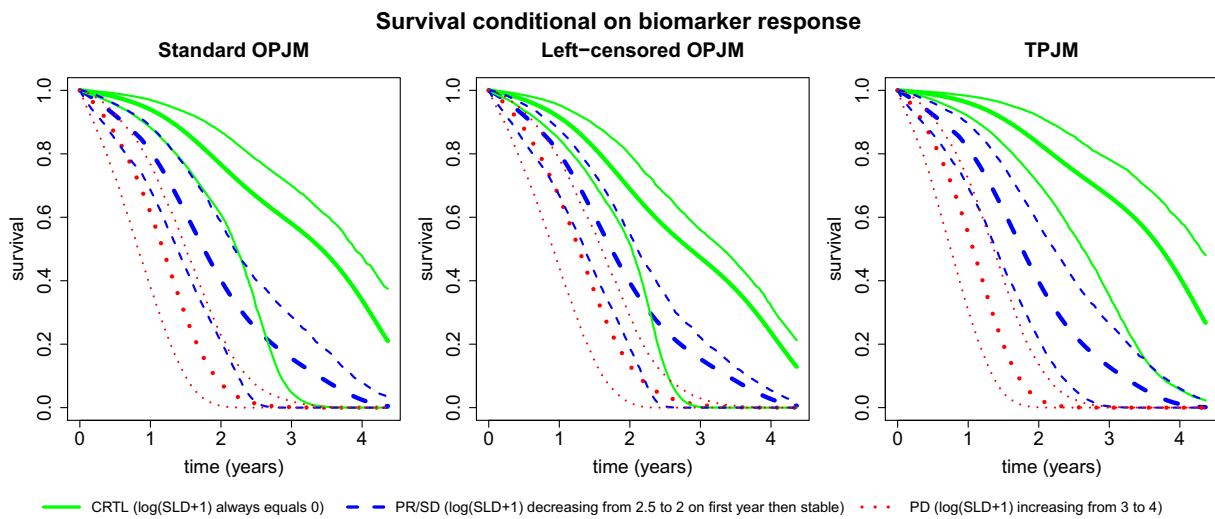


Fig. 2. Survival curves from the application according to biomarker response. Estimations are based on models with current value association structure and 95% confidence intervals are provided using Monte Carlo method with 1000 curves. We propose three patterns of response in order to illustrate the association between the biomarker and the terminal event. The CRTL corresponds to a zero value for the biomarker, the PR/SD corresponds to an individual with an initial exponential decline of the SLD from 11 cm to 6 cm and then a stabilization for the rest of the follow-up. The PD is defined as an exponential progression from a 19 cm SLD to 54 cm (maximum observed in the application is 57 cm).

individual deviation from the population mean distribution of the biomarker. For instance, let's assume a clinician is interested in the top 15% patients who had the largest SLD increase during follow-up compared to the population average. Their random effect b_{1i} should be higher than 1 standard deviation, that is from Table S2 of the supplementary material available at Biostatistics online, $b_{1i} > 1.46$. Conditional on $b_{1i} > 1.46$, the mean values of the random effects can be derived analytically (Aitken, 1935) or by sampling from a conditional multivariate normal distribution with correlation matrix given in Table S2 of the supplementary material available at Biostatistics online (last 3 rows). These (conditional) means are 2.81, -0.10 , and 2.23 for a , b_0 , and b_1 , respectively. Therefore, these top 15% individuals increase their chance to have the terminal event (i.e., to die) measured by an hazard ratio of $HR = \exp(0.21 \cdot 2.81 + 0.30 \cdot (-0.10) + 0.15 \cdot 2.23) = \exp(0.89) = 2.44$, compared to a patient who has an average longitudinal SLD profile. The second association ("current probability of positive value + expected positive value" and "current value") is helpful to clinicians interested to assess the effect of a CR versus a PR on the risk of terminal event. For instance, from Table 5, a patient whose expected log SLD value is 1 unit increase and probability of positive SLD is 50% at follow-up time t , has an increase in the risk of terminal event (i.e., to die) measured by an $HR = \exp(0.5 \cdot 1.80 + 1 \cdot 0.45) = \exp(1.35) = 3.86$, compared to a patient who had CR. The hazard ratio of having a PR versus CR is given by $HR = \exp(0.5 \cdot 1.80) = \exp(0.90) = 2.46$. Finally, the current value association measures the association of the expected value of the SLD (on the log scale) at follow-up time t on the risk of terminal event. For instance, from Table 5, a patient with an expected log SLD of 1 unit increase at follow-up time t , has an increased risk of terminal event (i.e., to die) measured by an $HR = \exp(1 \cdot 0.81) = 2.25$, compared to a patient with an expected SLD of 0.

5. DISCUSSION

In this article, we proposed a new TPJM for longitudinal semicontinuous biomarker data and a terminal event, which allows to account for excess of zeros in the biomarker distribution and joint inference on the

biomarker and survival outcomes. From a clinical standpoint, the TPJM is of particular interest because it can account for various clinical responses to treatment (e.g., CRTL, PR, SD, and PD) and has some flexibility in specifying the association structure between the biomarker and risk of event, as outlined in our application. Our simulation studies showed that some bias can arise when estimating time and time by treatment parameters, particularly when a larger proportion of zeros is observed for the biomarker. This is in line with previous results from [Smith and others \(2017\)](#) who showed through simulation studies that a negative treatment effect in the binary part can bias negatively the overall treatment effect captured by a one-part part model and can lead to misleading results. The real data application illustrates how the zeros and positive values of SLD could impact the overall survival. The results are consistent with the simulation studies, showing that one-part models can fail to explain some informative variability in the biomarker evolution over time, resulting in a reduced discrimination of the risk of death between treatment arms. The left-censored approach does not assume an excess of zeros in the biomarker distribution but instead a limit of detection, i.e., below this limit the biomarker values cannot be assessed. The choice between a one-part or a two-part approach should be guided by the research question, the two-part approach models the probability of a zero value as well as the conditional mean of positive values while the one-part approach models the marginal mean of the biomarker. The data structure can also help deciding which model to use.

A well-known computational challenge with joint models is the lack of analytical solution for the integrals over the random effects, which require some numerical approximation in the likelihood function. We initially proposed a standard Gauss--Hermite quadrature method for the numerical approximation of these integrals but our simulation studies showed that a Monte Carlo method gave more accurate estimations for better computational times for our model with three correlated random effects. There are several limitations to our approach including the interpretation of the LME model used to fit the positive continuous biomarker values. If a log link function was used in the likelihood instead of a log-transformation of the outcome, $\exp(\beta_k)$ would represent the multiplicative increase in the SLD associated with a one unit increase in the covariate X_k conditional on observing a positive value (instead of the additive effect on the log scale). Alternatively, marginalized two-part models have also been proposed in order to obtain interpretable covariate effects on the marginal mean ([Smith and others, 2014](#)) but have not been developed for joint models. Another limitation is that in the binary part, regression coefficients are subject-specific because of the non-linear link function from the GLMM which makes the distribution of the Gaussian random effects non normal when applying a back transformation to get either the odds or the probability of positive values of the biomarker. Although marginal odds and probability can be obtained using Monte Carlo method. Finally, our real data application includes patients who switched treatment during follow-up. This information was ignored from our analyses for simplicity, but it could have been a relevant information to include the time to treatment switch in the model. Non-target lesions and new lesions occurring during follow-up are also not included in the model and could be informative. [Król and others \(2018\)](#) applied a joint model to the GERCOR dataset, taking into account non-target lesions progressions and new lesions as recurrent events and observed a positive relationship between the risk of recurrent event and the risk of death.

Our application of the TPJM differs from usual applications of two-part models (e.g., medical costs and alcohol consumption) because the subset of complete responders to treatment for whom the disease disappeared, resulting in a zero biomarker value during follow-up, are quite informative when assessing treatment effect. Being able to characterize and predict patients with a CR is of primary clinical importance for personalized treatment options. For instance in 2017, the Food and Drug Administration (FDA) approved for the first time a cancer treatment (pembrolizumab) for any solid tumor based on patients tumor biomarker status (DNA/RNA/protein features) rather than on tumor histology. The TPJM allows to estimate covariate effects on the probability of CRTL and on the expected value among positive values of the sum of the longest diameter of target lesions. This could help characterize which part of the

population will get a beneficial or deleterious effect from a given treatment strategy. We also assessed the robustness of the TPJM when the excess of zeros of the biomarker was taken into account by a left-censoring mechanism. In our future developments, we plan to extend the model to a marginal TPJM in order to get an effect of covariates on the marginal mean of the biomarker. The advantage of the current conditional versus marginal TPJM is that the former models specifically the zeros and the positive values, allowing to evaluate better the covariate effect on the binary and on the positive continuous outcomes whereas the marginal model provides covariate effects on the binary outcome and on the marginal mean of the biomarker, accounting for the excess of zeros. The application of the TPJM is not limited to the joint analysis of tumor size and survival in the context of cancer clinical trials. Two-part models are now also becoming very popular in various fields such as microbiome analysis, to account for the excess of zeros of count data generated from high-throughput sequencing technologies (Chen and Li, 2016; Chai and others, 2018).

6. SOFTWARE

All of our model developments are implemented in the `longiPenal` function of the freely available R package *frailtypack* (Król and others, 2017). This package can be used to fit a variety of joint frailty models or other frailty models for recurrent or clustered time-to-event data with several different options for the baseline risk functions. The standard version of the function uses OpenMP to parallel computations within a multi-core node. An MPI version of the function was also developed to use parallel computing between nodes when using the function on a server. The package can be downloaded from the Comprehensive R Archive Network accessible via <http://cran.r-project.org/package=frailtypack>.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at <http://biostatistics.oxfordjournals.org>.

ACKNOWLEDGMENTS

Computer time for this study was provided by the computing facilities MCI (Mésocentre de Calcul Intensif Aquitain) of the Université de Bordeaux and of the Université de Pau et des Pays de l'Adour. The authors acknowledge the insightful and constructive comments made by associate editor and two reviewers. These comments have greatly helped to sharpen the original submission.

Conflict of Interest: None declared.

FUNDING

This project has received funding from the EHESP, the National Cancer Institute (INCA EVALUATES OPE 2017-0680), the Canadian Institute of Health Research (CIHR) project grants and NSERC discovery grants.

REFERENCES

- AITKEN, A. C. (1935). Note on selection from a multivariate normal population. *Proceedings of the Edinburgh Mathematical Society* **4**, 106--110.
- CHAI, H. JIANG, H. LIN, L. AND LIU, L. (2018). A marginalized two-part beta regression model for microbiome compositional data. *PLoS Computational Biology* **14**, 1--16.
- CHAMPIAT, S., DERCLE, L., AMMARI, S., MASSARD, C., HOLLEBECQUE, A., POSTEL-VINAY, S. AND FERTÉ, C. (2017). Hyperprogressive disease is a new pattern of progression in cancer patients treated by anti-PD-1/PD-L1. *Clinical Cancer Research* **23**, 1920--1928.

- CHEN, E. Z. AND LI, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics* **32**, 2611--2617.
- CRAGG, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica* **39**, 829.
- DAGNE, G. A. (2017). Joint two-part Tobit models for longitudinal and time-to-event data. *Statistics in Medicine* **36**, 4214--4229.
- DUAN, N., MANNING, W. G., MORRIS, C. N. AND NEWHOUSE, J. P. (1983). A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics* **1**, 115.
- KRÓL, A., FERRER, L., PIGNON, J.-P., PROUST-LIMA, C., DUCREUX, M., BOUCHÉ, O. AND RONDEAU, V. (2016). 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**, 907--916.
- KRÓL, A., TOURNIGAND, C., MICHIELS, S. AND RONDEAU, V. (2018). Multivariate joint frailty model for the analysis of nonlinear tumor kinetics and dynamic predictions of death. *Statistics in Medicine* **37**, 2148--2161.
- KRÓL, A., MAUGUEN, A., MAZROUI, Y., LAURENT, A., MICHIELS, S. AND RONDEAU, V. (2017) Tutorial in joint modeling and prediction: a statistical software for correlated longitudinal outcomes, recurrent events and a terminal event. *Journal of Statistical Software* **81**, 1--52.
- LAMBERT, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1.
- LITIÈRE, S., COLLETTE, S., DE VRIES, E. G. E., SEYMOUR, L. AND BOGAERTS, J. (2017). RECIST—learning from the past to build the future. *Nature Reviews Clinical Oncology* **14**, 187--192.
- LIU, L. (2009). Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Statistics in Medicine* **28**, 972--986.
- LIU, L., SHIH, T., STRAWDERMAN, R. L., ZHANG, D., JOHNSON, B. A. AND CHAI, H. (2019). Statistical analysis of zero-inflated nonnegative continuous data: a review. *Statistical Science* **34**, 253--279.
- MANNING, W. G., DUAN, N. AND ROGERS, W. H. (1987). Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics*, **35**, 59--82.
- MARQUARDT, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics* **11**, 431--441.
- MULLAHY, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics* **33**, 341--365.
- OLSEN, M. K. AND SCHAFER, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association* **96**, 730--745.
- SMITH, V. A., PREISSER, J. S., NEELON, B. AND MACIEJEWSKI, M. L. (2014). A marginalized two-part model for semicontinuous data. *Statistics in Medicine* **33**, 4891--4903.
- SMITH, V. A., NEELON, B., MACIEJEWSKI, M. L. AND PREISSER, J. S. (2017). Two parts are better than one: modeling marginal means of semicontinuous data. *Health Services and Outcomes Research Methodology* **17**, 198--218.
- SMITH, V. A., MACIEJEWSKI, M. L. AND OLSEN, M. K. (2018). Modeling semicontinuous longitudinal expenditures: a practical guide. *Health Services Research* **53**, 3125--3147.
- SU, L., TOM, B. D. M. AND FAREWELL, V. T. (2009). Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics* **10**, 374--389.
- SYLVESTRE, M.-P. AND ABRAHAMOWICZ, M. (2008). Comparison of algorithms to generate event times conditional on time-dependent covariates. *Statistics in Medicine* **27**, 2618--2634.

Joint modeling of a semicontinuous biomarker and a terminal event

19

- TOOZE, J. A., GRUNWALD, G. K., AND JONES, R. H. (2002). Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research* **11**, 341--355.
- TOURNIGAND, C., ANDRÉ, T., ACHILLE, E., LLEDO, G., FLESH, M., MERY-MIGNARD, D. AND DE GRAMONT, A. (2004). FOLFIRI followed by FOLFOX6 or the reverse sequence in advanced colorectal cancer: a randomized GERCOR study. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **22**, 229--237.

[Received October 3, 2019; revised January 19, 2020; accepted for publication February 19, 2020]

Biostatistics (2020), **0**, **0**, pp. 1–6
doi:10.1093/biostatistics/output

**Supplementary material to “Two-part joint
model for a longitudinal semicontinuous marker
and a terminal event with application to
metastatic colorectal cancer data”**

DENIS RUSTAND*

*Biostatistic Team, Bordeaux Population Health Center, ISPED, Centre INSERM U1219,
Bordeaux, France*

denis@rustand.fr

LAURENT BRIOLLAIS

*Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital and Dalla Lana School of
Public Health, University of Toronto, ON, Canada*

CHRISTOPHE TOURNIGAND

Hôpital Henri Mondor, Creteil, France

VIRGINIE RONDEAU

*Biostatistic Team, Bordeaux Population Health Center, ISPED, Centre INSERM U1219,
Bordeaux, France*

[Received August 1, 2019; revised October 1, 2019; accepted for publication November 1, 2019]

*To whom correspondence should be addressed.

Table S1. *Patients characteristics*

Treatment	arm A	arm B
Sample size	n=101	n=104
Covariates		
Sex		
Male	55	77
Female	46	27
Age		
< 60	35	32
60 – 69	48	43
≥ 70	18	29
WHO performance status		
0	45	49
1	40	48
2	16	7
Primary site		
Colon	65	76
Rectum	36	26
Multiple	0	2
Previous surgery		
No	15	8
Curative	35	32
Palliative	51	64
Previous chemotherapy		
Yes	20	25
No	81	79
Previous radiotherapy		
No	78	86
Yes	23	18
Metastases		
Synchronous	73	79
Metachronous	28	25
Number of metastatic sites 1		
2+	83	83
Liver metastases		
No	74	79
Yes	27	25
Lung metastases		
No	74	79
Yes	27	25
Biomarker		
Number of repeated measurements	748	727
% of positive biomarker values	92.5%	83.8%
Baseline mean value (log scale)	2.37 (SD=0.68)	2.18 (SD=0.63)
Mean value	1.92 (SD=0.86)	1.66 (SD=0.97)
Mean value among positives	2.08 (SD=0.69)	1.98 (SD=0.69)

Joint modeling of a semicontinuous biomarker and a terminal event

3

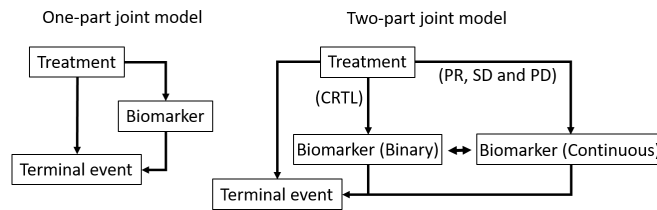


Fig. S1. Diagrams describing the decomposition of treatment effect with a standard one-part joint model (left) and a two-part joint model (right). Treatment effect is decomposed into a direct effect on the risk of terminal event and an indirect effect on the risk of terminal event, through the biomarker. The TPJM decomposes the biomarker distribution into a binary outcome corresponding to the probability of CRTL and a continuous outcome with the positive biomarker measurements.

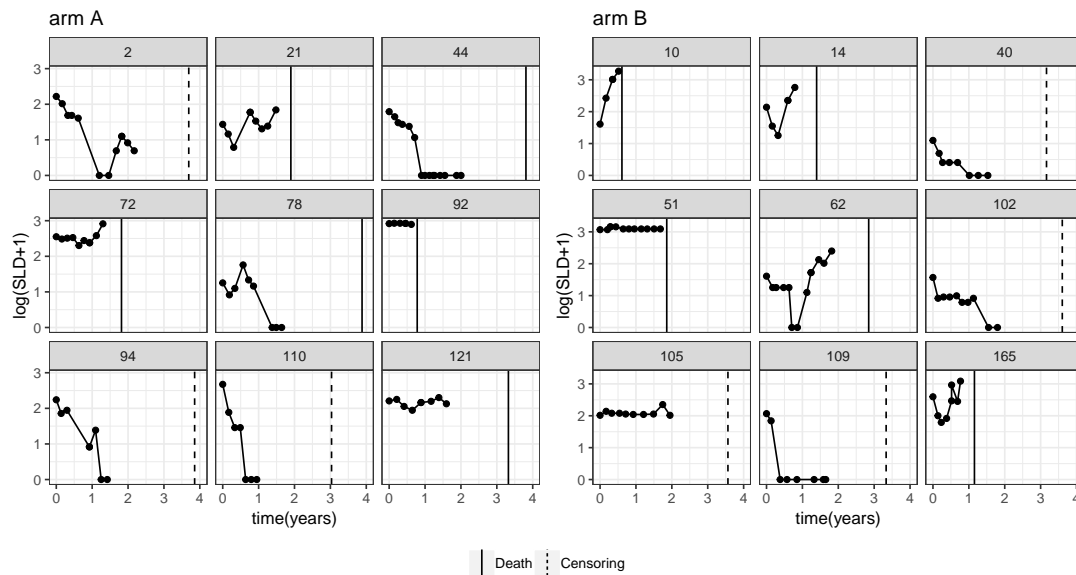


Fig. S2. Individual trajectories of the biomarker for a subset of patients allocated to arm A (FOLFIRI followed by FOLFOX6) and arm B (FOLFOX6 followed by FOLFIRI).

Table S2. Application to advanced colorectal cancer data: shared random-effects association. The binary model is adjusted on the WHO performance status (0/1/2), lung metastatic site (yes/no), previous adjuvant radiotherapy (yes/no), the continuous model is adjusted on the WHO performance status, metastases (metachronous/synchronous), previous surgery (no surgery/curative/palliative), previous adjuvant radiotherapy and the survival model is adjusted on WHO performance status, metastases, previous surgery and previous adjuvant radiotherapy.

Variable	Joint standard Est. (SE)	Joint left-censored Est. (SE)	Two-part joint Est. (SE)
Binary part (SLD>0 versus SLD=0)			
intercept			6.89*** (0.78)
time (year)			-2.13*** (0.42)
treatment (B/A)			-1.40• (0.75)
time:treatment (B/A)			0.27 (0.49)
Continuous part (SLD>0)			
intercept	2.05*** (0.10)	2.06*** (0.10)	2.10*** (0.08)
time (years)	-0.37* (0.17)	-0.45* (0.19)	-0.35*** (0.08)
treatment (B/A)	-0.49*** (0.14)	-0.43** (0.14)	-0.29** (0.11)
time:treatment (B/A)	0.14 (0.22)	-0.03 (0.27)	0.34** (0.10)
residual S.E.	0.84 (0.02)	0.80 (0.02)	0.40 (0.01)
Death risk			
treatment (B/A)	0.14 (0.22) [HR=1.15]	0.12 (0.23) [HR=1.13]	0.27 (0.20) [HR=1.30]
Association			
continuous intercept	0.54*** (0.13)	0.51*** (0.13)	0.30 (0.30)
continuous slope	0.71*** (0.17)	0.58*** (0.12)	0.15 (0.59)
binary intercept			0.21• (0.12)
random-effects			
continuous intercept (σ_{b_0})	0.86	0.86	0.78
continuous slope (σ_{b_1})	1.19	1.75	1.46
binary intercept (σ_a)			2.56
$corr_{b_0b_1}$	0.05	-0.09	-0.09
$corr_{ab_0}$			0.19
$corr_{ab_1}$			0.72

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ • $p < 0.1$

Joint modeling of a semicontinuous biomarker and a terminal event

5

Table S3. *Application to advanced colorectal cancer data: covariates effects*

Variable (association)	Joint standard (current value) Est. (SE)	Joint left-censored (current value) Est. (SE)	Two-part joint (current value) Est. (SE)	Two-part joint (current probability, current value ⁽⁺⁾) Est. (SE)
Binary part (SLD>0 versus SLD=0)				
WHO performance status (1)			1.77** (0.60)	1.72** (0.60)
WHO performance status (2)			1.72• (0.98)	1.69• (0.98)
prev. radio			0.33 (0.69)	0.35 (0.71)
lung metastatic site			2.37** (0.72)	2.29** (0.76)
Continuous part (SLD>0)				
WHO performance status (1)	0.47*** (0.10)	0.46*** (0.10)	0.37*** (0.09)	0.36*** (0.09)
WHO performance status (2)	0.52*** (0.15)	0.55*** (0.15)	0.42** (0.13)	0.41** (0.14)
surgery (curative)	-0.51* (0.22)	-0.51* (0.20)	-0.35* (0.16)	-0.34* (0.16)
surgery (palliative)	-0.10 (0.15)	-0.10 (0.14)	0.01 (0.13)	0.02 (0.13)
prev. radio	-0.19 (0.14)	-0.21 (0.13)	-0.22• (0.12)	-0.21• (0.11)
metastases (metachronous)	0.42* (0.20)	0.41* (0.17)	0.29* (0.13)	0.28* (0.13)
Death risk				
WHO performance status (1)	0.46* (0.18)	0.46** (0.17)	0.25 (0.17)	0.38* (0.18)
WHO performance status (2)	1.22*** (0.28)	1.28*** (0.28)	1.09*** (0.28)	1.24*** (0.28)
surgery (curative)	-0.65• (0.38)	-0.69• (0.38)	-0.61 (0.38)	-0.72• (0.39)
surgery (palliative)	-0.53* (0.26)	-0.57* (0.26)	-0.50• (0.26)	-0.58* (0.27)
metastases (metachronous)	0.76* (0.32)	0.74* (0.32)	0.76* (0.32)	0.73* (0.32)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ • $p < 0.1$

Table S4. *Summary of a simulated dataset (github.com/DenisRustand/TPJM_sim). The design is the same as for scenario 1 of the simulation studies but with a reduced sample size (150) and Monte-Carlo integration points (500) to save computation time.*

Survival dataset	id	deathTimes	d	trt
Min.	1.00	0.0020	0.00	0.0000
1st Qu.	38.25	0.2105	1.00	0.0000
Median	75.50	0.3930	1.00	0.0000
Mean	75.50	0.6127	0.82	0.4733
3rd Qu.	112.75	0.7890	1.00	1.0000
Max.	150.00	2.8490	1.00	1.0000
Longitudinal dataset	id	timej	trtY	Y
Min.	1.0	0.0000	0.0000	0.000
1st Qu.	39.0	0.0700	0.0000	2.949
Median	80.0	0.2800	0.0000	4.096
Mean	79.8	0.4541	0.4052	3.665
3rd Qu.	124.0	0.7000	1.0000	4.930
Max.	150.0	2.8000	1.0000	7.511

Table S5. Application of the TPJM to the simulated dataset from Table S4, also available in the examples of the longiPenal function of the R package frailtypack.

Variable	Two-part joint Est. (SE)
Binary part (SLD>0 versus SLD=0)	
intercept	5.71 (0.60)
time (years)	-3.87 (0.46)
treatment	-1.27 (0.61)
time:treatment	1.44 (0.66)
Continuous part (SLD>0)	
intercept	4.06 (0.09)
time (years)	-0.48 (0.13)
treatment	1.02 (0.14)
time:treatment	1.11 (0.21)
residual S.E.	0.48 (0.01)
Survival part	
treatment	0.35 (0.21)
association	0.38 (0.10)
random-effects covariance matrix	
continuous intercept ($\sigma_{b_0}^2$)	$\begin{bmatrix} 0.66 & & \\ -0.20 & 0.16 & \\ 0.55 & 0.30 & 3.36 \end{bmatrix}$
continuous slope ($\sigma_{b_1}^2$)	
binary intercept (σ_a^2)	

3.3 Additional remarks

3.3.1 On the left-censoring

With the left-censoring model, we remove the constraint of no negative measurements with a nonlinear transformation and assume the value of the tumor size can decrease infinitely (usually with a decreasing speed over time) towards zero. From a clinical point of view, the tumor size is defined by an accumulation of cancerous cells. As a consequence, there is a limit in the tumor size where the tumor will only be composed of a single cell. An additional decrease of the tumor size implies the disappearance of the last cell and therefore observing a true zero. For example, if we focus on the results of the application of the left-censoring model on the GERCOR study, the mean baseline value of the biomarker is $\simeq 2$ on the log scale ($\simeq 7.4\text{cm}$) and the slope is $\simeq -0.5$. Using these values it takes about 20 years of decrease to reach a size less than a single cell ($\simeq 0.001\text{cm}$) for the mean biomarker value estimated with the left-censoring model. This sounds reasonable because we are in the context of metastatic cancer and only a few patients reach a zero value of the biomarker. In the simulation studies, we evaluated the left-censoring model when the two-part model was specified for the data generation, using a baseline value on the log scale of 4 and a decrease by year of -3 , resulting in 35% zeros on average. This simulation scenario is justified by the increasing efficacy of treatments, leading to increased rates of zero measurements of the biomarker. Recent clinical trials can have up to 80% of complete responders, i.e., SLD=0 (Mangal et al. (2018)). With our simulation scenario with 35% zeros, the mean predicted value of the biomarker with the left censoring model reaches the limit of the size of a single cell after only 4 years of follow-up. Moreover this is the estimated mean value and since we assume Gaussian individual-specific random effects around this mean, some individuals reach the limit before 4 years of follow-up. This confirms our conclusion that the left-censoring model is not appropriate in situations where true zeros of the biomarker can be observed and should be used only in case of a small zero rate and when there is no interest in the process driving the occurrence of zero values.

3.3.2 On the numerical integration

The standard joint model for a longitudinal biomarker and a terminal event developed in **frailty-pack** uses a multivariate non-adaptive Gaussian quadrature or a pseudo-adaptive Gaussian quadrature rule for the numerical approximation of the integral over the random effects. It is fast and efficient for the standard joint model because only one regression model includes random effects for the biomarker, resulting in a low dimension integral. In order to compute integrals in multiple dimensions, the quadrature rules requires the function evaluations to grow exponentially as the number of dimensions increases (“curse of dimensionality”). The computational burden for this numerical approximation therefore depends on the dimension of the random effects. With a two-part joint model, the biomarker distribution is decomposed into two regression models with correlated random effects to characterize the biomarker’s distribution. In this context, a method to overcome this limitation is the Monte-Carlo method that computes the integral by taking draws from the corresponding distribution. The Monte Carlo method relies on the law of large numbers, it is time consuming for low dimensional integrals compared to the

quadrature rules but is much less affected by an increase in the dimension of the integration. In this context, we developed a Monte Carlo method for the two-part joint model in **frailtypack**.

3.3.3 Erratum in the likelihood function

The likelihood function presented in the article contains 2 errors, the correct likelihood should be

$$\begin{aligned}
 L_i(\Theta) = & \int_{\mathbf{a}_i} \int_{\mathbf{b}_i} \left[\frac{1}{\left(\sqrt{2\pi\sigma_\epsilon^2}\right)^{n_i}} \prod_{j=1}^{n_i} \exp\left(-\frac{(g(Y_{ij}^+) - \mathbf{X}_{Cij}^\top \boldsymbol{\beta} - \mathbf{Z}_{Cij}^\top \mathbf{b}_i)^2)^{U_{ij}}}{2\sigma_\epsilon^2}\right) \right. \\
 & \times \left(\exp\left[\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i\right] \right)^{U_{ij}} \times \left(1 - \frac{\exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)}{1 + \exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)} \right) \\
 & \left. \times \lambda_i(T_i | \Theta_i)^{\delta_i} \exp\left(-\int_0^{T_i} \lambda_i(t | \Theta_i) dt\right) p(\mathbf{a}_i, \mathbf{b}_i) \right] d\mathbf{b}_i d\mathbf{a}_i.
 \end{aligned}$$

Chapter 4

A marginal two-part joint model for a longitudinal biomarker and a terminal event with application to head and neck cancer data

4.1 Introduction

The methodological development of the two-part joint model proposed in Chapter 3 was limited to the (standard) conditional form of the two-part model. It decomposes the biomarker into a binary and a continuous part for the positive vs. zero values and the positive continuous observations. These two parts are independent conditional on the random effects and evaluate the effect of covariates on the probability of positive value and the expected conditional positive values. The focus of Chapter 3 was the comparison of the two-part model with alternative strategies such as the left-censoring model, exploring the properties of the model under various data generation scenarios. The marginal two-part model was recently proposed as an alternative formulation of the conditional two-part model. A reformulation of the likelihood allows to evaluate the effect of covariates on the marginal mean value of the biomarker instead of the mean across positive values only. The marginal two-part model is a mix between the left-censoring approach that provides an effect of covariates on the mean biomarker value and the conditional two-part model that accounts for and characterize the zero excess with a specific submodel. In this Chapter, we describe the differences between the conditional and marginal formulations of the two-part model for the SLD, in the context of joint modeling with survival. A simulation study evaluates the behavior of each formulation along with the left-censoring joint model using alternatively each model as the true model for the data generation. We show how the marginal formulation is robust regardless of the simulation scenario while the other formulations perform poorly when

the model is misspecified. We illustrate how the new marginal two-part joint model facilitates the clinical interpretation of the results of a phase III randomized clinical trial for patients with metastatic and/or recurrent squamous cell carcinoma of the head and neck, the SPECTRUM (Study of Panitumumab Efficacy in Patients With Recurrent and/or Metastatic Head and Neck Cancer) study. The purpose of this study was to determine the treatment effect of panitumumab in combination with chemotherapy versus chemotherapy alone as first line therapy. We show how the conditional form of the TPJM remains relevant because the marginal two-part joint model does not inform about the expected value of the biomarker conditional on a positive value. The model choice depends therefore on the clinical question of interest. In case of high zero rate, the mean biomarker value could be driven mostly by the zero vs. positive values rather than by the distribution of positive values and the two formulations could produce very different results. In the context of metastatic cancer, we often observe a small zero rate but there is a great interest in these zeros as they represent patients with a complete response of their target lesions to treatment. The conditional two-part joint model can be unstable when the zero rate is small and we investigate how the marginal formulation for the two-part model provides a more stable model because the association between the binary and continuous process is accounted for in the likelihood in addition to the correlation between the random effects.

4.2 Article

A marginal two-part joint model for a longitudinal biomarker and a terminal event with application to advanced head and neck cancers

Denis Rustand*

Biostatistic Team, Bordeaux Population Health Center, INSERM U1219,
146 rue Léo Saignat, 33076 Bordeaux, France

**email:* denis@rustand.fr

and

Laurent Briollais

Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital and
Dalla Lana School of Public Health (Biostatistics), University of Toronto,
600 University Ave., Ontario M5G 1X5, Canada

and

Virginie Rondeau

Biostatistic Team, Bordeaux Population Health Center, INSERM U1219,
146 rue Léo Saignat, 33076 Bordeaux, France

SUMMARY: The Sum of the Longest Diameter of the target lesions (SLD) is a longitudinal biomarker used to assess tumor response in cancer clinical trials, which can inform about early treatment effect. This biomarker is semicontinuous, often characterized by an excess of zeros and right skewness. Conditional two-part joint models were introduced to account for the excess of zeros in the longitudinal biomarker distribution and link it to a time-to-event outcome. A limitation of the conditional two-part model is that it does not provide an overall estimate of covariate effects, such as treatment, on the biomarker, which is often of clinical relevance. As an alternative, we propose in this paper, a marginal two-part joint model (M-TPJM) for the repeated measurements of the SLD and a terminal event, where the covariates affect the overall mean of the biomarker. Our simulation studies assessed the good performance of the marginal model in terms of estimation and coverage rates. Our application of the M-TPJM to a randomized clinical trial of advanced head and neck cancer shows that the combination of panitumumab in addition with chemotherapy increases the odds of observing a disappearance of all target lesions compared to chemotherapy alone, leading to a

possible indirect effect of the combined treatment on time to death.

KEY WORDS: conditional two-part; joint model; left-censored GLM; marginal two-part; randomized clinical trial; semicontinuous; solid tumors.

1 Introduction

In solid tumor cancer clinical trials, there is an increased interest in the joint analysis of the time to death and the Sum of the Longest Diameter of the target lesions (SLD), defined according to the Response Evaluation Criteria in Solid Tumours (RECIST). This biomarker reflects the tumor burden and its evolution over time. It is important to account for the association between the longitudinal outcome and the risk of terminal event because the former is censored by the terminal event, and the latter is highly affected by the value and the evolution of the biomarker over time. The SLD distribution is often characterized by an excess of zeros and right skewness. Patients whose treatment removes all visible signs of the disease generates zero values for the SLD. This excess of zeros is therefore highly informative of treatment efficiency.

A conditional two-part joint model (C-TPJM) has been introduced to fit the SLD evolution over time jointly with the risk of terminal event, while taking into account the semicontinuous distribution of the biomarker. When an excess of true zeros is observed, the model was shown superior to standard approaches such as left-censoring the biomarker's distribution (i.e. assuming zero values are censored values, too small to be observed) to compare clinical treatment strategies. Indeed, the left-censoring one-part joint model (OPJM) fails to explain some informative variability in the biomarker evolution over time, resulting in a reduced discrimination of the risk of death between treatment arms (Rustand et al. (2020)). The conditional two-part model decomposes the distribution of the outcome into a binary part corresponding to zero versus positive values and a continuous part with positive values only, both outcomes being modelled by a mixed effects regression model. The binary and continuous parts are linked through correlated random effects. The model yields covariate effects, such as treatment effect, on the probability of observing a positive versus zero SLD in the binary part and on the expected value of the biomarker conditional on observing a

2

positive value (i.e., zeros excluded) in the continuous part. On the other hand, this model cannot provide treatment effect on the marginal mean of the biomarker, which is often of clinical interest. For the terminal event, the hazard function can be expressed conditionally on observing a zero SLD value (which is indicative of a complete response to treatment) and on the expected value of the biomarker among positive values (which is indicative of a partial response).

The marginal two-part model (Smith et al. (2017)) is an alternative approach, useful when the interest lies in the population-average effects of covariates, such as treatment effect, on the biomarker. This model accounts for the zero values in the continuous part of the model and provides covariates effects on the marginal mean of the biomarker. In addition, a binary part, similar to the conditional two-part model, accounts for the excess of zeros and can assess covariate effects on the probability of observing a positive biomarker value vs. a zero value. The conditional and marginal two-part models can address different clinical questions. When the interest is in the expectation of the biomarker among positive values, the conditional model is more appropriate while the marginal two-part model may lead to arbitrary heterogeneity and provides less interpretable estimates on the conditional mean of the biomarker among positive values (Smith et al. (2014)). The left-censoring one-part model provides similar covariates effects on the marginal mean of the biomarker as the marginal two-part model, but does not account for the excess of zero values. The OPJM rather considers an excess of values under a certain threshold or limit of detection. The marginal two-part model combines the advantages of the conditional two-part model and the left-censoring one-part model by allowing a direct interpretation of covariate effect on the population mean value of the biomarker while also accounting for the excess of zeros. In the application section of this article, we illustrate the differences between these modelling

strategies. The marginal two-part model has not been yet proposed in the context of joint models.

In this paper, we propose a marginal two-part joint model (M-TPJM) for a longitudinal semicontinuous outcome and a terminal event. We compare the new model with the C-TPJM and the left-censoring OPJM through simulation studies and provide a detailed interpretation of these models. The remainder of the article is structured as follows: in Section 2, we describe the M-TPJM and its estimation method. In Section 3, we present a simulation study to assess the performance of the model and compare it to competing approaches that treat the excess of zeros differently. An application to a randomized clinical trial comparing a combination of chemotherapy and panitumumab (anti-EGFR monoclonal antibody) to chemotherapy alone, in patients with metastatic and/or recurrent squamous-cell carcinoma of the head and neck, is proposed in Section 4 and we conclude with a discussion in Section 5.

2 Model

2.1 Left-censoring one-part model for the biomarker

Let Y_{ij} denote the biomarker value for subject i ($i = 1, \dots, n$), at visit j ($j = 1, \dots, n_i$). The model assumes the biomarker can be subject to left-censoring when it decreases below a limit of detection c .

$$Y_{ij}^* = \begin{cases} Y_{ij} & \text{if } Y_{ij} > c \\ c & \text{otherwise} \end{cases} \quad (1)$$

The Y_{ij}^* has the same distribution as the Y_{ij} when $Y_{ij} > c$. For the observations $Y_{ij} = c$, all we know is $P(Y_{ij}^* = c) = P(Y_{ij} < c)$, see (Tobin (1958)).

2.2 Conditional two-part model for the biomarker

The biomarker distribution is decomposed into a binary outcome $I[Y_{ij} > 0]$ and a positive-continuous outcome $Y_{ij}|Y_{ij} > 0$. A GLMM with a logit link is assumed for the binary outcome and a GLMM with a logarithm link is specified for the positive continuous outcome. The logarithm link in the continuous part is used to linearize the biomarker evolution over time and correct for right-skewness and heteroscedasticity. The two components are linked through correlated random effects. The two-part model for the biomarker is defined as follows:

$$\begin{cases} \text{Logit}(\text{Prob}(Y_{ij} > 0)) = \mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i & \text{(Binary part),} \\ \text{E}[Y_{ij}|Y_{ij} > 0] = \exp(\mathbf{X}_{Cij}^\top \boldsymbol{\beta} + \mathbf{Z}_{Cij}^\top \mathbf{b}_i) & \text{(Continuous part),} \end{cases} \quad (2)$$

where \mathbf{X}_{Bij} and \mathbf{Z}_{Bij} are vectors of covariates associated with the fixed effect parameters $\boldsymbol{\alpha}$ and the random effects \mathbf{a}_i for the binary part. Similarly, \mathbf{X}_{Cij} and \mathbf{Z}_{Cij} are vectors of covariates associated with the fixed effect parameters $\boldsymbol{\beta}$ and the random effects \mathbf{b}_i . We assume a normal and independently distributed error term in the continuous part $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon)$.

The two vectors of random effects follow a multivariate normal distribution:

$$\begin{bmatrix} \mathbf{a}_i \\ \mathbf{b}_i \end{bmatrix} \sim MVN \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \Sigma_a^2 & \Sigma_{ab} \\ \Sigma_{ab} & \Sigma_b^2 \end{bmatrix} \right). \quad (3)$$

The vectors of correlated subject-specific random effects \mathbf{a}_i and \mathbf{b}_i account for the correlation between repeated measurements within an individual and the correlation between the two components of the model. The logistic regression model includes covariates that represent the effect of an individual's characteristics on the probability of observing a positive versus zero biomarker value. The continuous part represents the log of the expected value of the biomarker given a positive biomarker value. This model differs from the conditional two-part model proposed in Rustand et al. (2020) for which the continuous part represented the expected value of the log-transformed longitudinal outcome, resulting in an additive effect of covariates on the transformed scale of the biomarker. Using now a logarithm link facilitates the interpretation of a covariate k , where $\exp(\beta_k)$ represents the multiplicative effect on the

(natural scale) biomarker value at a given time point, conditional on a positive value at that time point, associated with a one-unit increase in the covariate X_k (Smith et al. (2018)).

2.3 Marginal two-part model for the biomarker

In the context of two-part models, the term marginal refers to the biomarker distribution including both zeros and positive values (to contrast with the conditional form) and it does not refer to the “marginal/subject-specific” usage. In the M-TPJM, the binary part is similar to the one used in the conditional model, but the continuous part models the covariate effects on the marginal mean of the biomarker. The model is defined as follows:

$$\begin{cases} \text{Logit}(\text{Prob}(Y_{ij} > 0)) = \mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i & \text{(Binary part),} \\ \text{E}[Y_{ij}] = \exp(\mathbf{X}_{Cij}^\top \boldsymbol{\beta} + \mathbf{Z}_{Cij}^\top \mathbf{b}_i) & \text{(Continuous part),} \end{cases} \quad (4)$$

The marginal two-part model gives the effect of covariates on the marginal mean of the biomarker instead of the mean conditional on observing a positive value of the biomarker by including both the zeros and positive values in the continuous part. The correlated random effects capture some correlation due to potentially unobserved process driving the probability of positive value and the marginal mean value, i.e., lower values of the biomarker are more likely correlated with the probability of observing a zero. Another induced correlation is that the expression of the overall mean also depends on the probability of observing a positive value (see Equation 7). With the conditional two-part model, the association between the binary and continuous part is only captured through the correlation structure of the random effects.

2.4 Marginal two-part joint model

The proposed model considers a marginal two-part model for the biomarker evolution over time and a Cox proportional hazards model for the terminal event.

$$\left\{ \begin{array}{ll} \text{Logit}(\text{Prob}(Y_{ij} > 0)) = \mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i & \text{(Binary part),} \\ \text{E}[Y_{ij}] = \exp(\mathbf{X}_{Cij}^\top \boldsymbol{\beta} + \mathbf{Z}_{Cij}^\top \mathbf{b}_i) & \text{(Continuous part),} \\ \lambda_i(t) = \lambda_0(t) \exp(\mathbf{X}_i^\top \boldsymbol{\gamma} + \mathbf{h}(\cdot)^\top \boldsymbol{\varphi}) & \text{(Survival part),} \end{array} \right. \quad (5)$$

The two vectors of random effects \mathbf{a}_i and \mathbf{b}_i follow a multivariate normal distribution as defined by Equation 3. The function $h(\cdot)$ corresponds to the association function between the biomarker and the risk of event, and $\boldsymbol{\varphi}$ the corresponding vector of association parameters. In the C-TPJM, the continuous part is given by the second line of Equation 2. The log-normal distribution assumed for the continuous part cannot be used with the standard OPJM because of the presence of zeros. Left-censoring is then required, assuming the values observed below a censoring threshold (i.e., the zeros) are positive but too small to be measured. The left-censoring OPJM only includes the continuous part and the survival part of Equation 5.

2.5 Association structures

We propose two possible association structures. In the “shared random effects” (SRE) association structure between the biomarker and the survival model, we have

$$\lambda_i(t) = \lambda_0(t) \exp(\mathbf{X}_i^\top \boldsymbol{\gamma} + \mathbf{a}_i^\top \boldsymbol{\varphi}_a + \mathbf{b}_i^\top \boldsymbol{\varphi}_b)$$

and in the “current level” (CL) association (also referred to as “current value”) we have

$$\lambda_i(t) = \lambda_0(t) \exp(\mathbf{X}_i^\top \boldsymbol{\gamma} + \text{E}[Y_{ij}] \boldsymbol{\varphi})$$

The SRE association is useful to explore the association between an individual’s deviation from the population mean evolution of the biomarker and the risk of terminal event but does not take into account the covariance between the two vectors of random effects in the survival model. The difference in the SRE association structure between the M-TPJM and the C-

TPJM is that the individual heterogeneity captured in the continuous part by the random effects is conditional on observing a positive value of the biomarker with the C-TPJM, while for the marginal model it corresponds to the entire population. The biomarker model takes into account informative censoring by the terminal event through the shared random effects while the survival part gives the hazard ratio of the covariates conditional on the random effects, assuming proportional hazards.

In the CL association, φ quantifies the strength of the association between the true unobserved value of the longitudinal biomarker and the risk of event, the interpretation depends on the model for the biomarker. With a C-TPJM, the “current value” of the biomarker is given by

$$E[Y_{ij}] = \text{Prob}(Y_{ij} > 0)E[Y_{ij}|Y_{ij} > 0], \quad (6)$$

which is a combination of two non-linear regressions resulting in a difficult interpretation of the association of the evolution over time of the biomarker value and its effect on the risk of terminal event. The M-TPJM directly models the mean value of the biomarker $E[Y_{ij}]$, which facilitates the interpretation of covariates effect on the biomarker mean value. Because $E[Y_{ij}]$ is directly obtained from the M-TPJM, the variance of the estimated value of the biomarker is reduced under the M-TPJM, as illustrated in Figure 2. In terms of covariate effects, the CL association can be thought of as modelling both the direct effect of covariates on the risk of terminal event and an indirect effect through the biomarker, which in turn is linked to the terminal event through the association structure, although the joint model does not provide a formal mediation analysis.

2.6 Interpretation of treatment effect

[Figure 1 about here.]

A diagram describing the decomposition of treatment effect with the C-TPJM and the M-TPJM is given in Figure 1. With the SRE association, the survival model gives the

8

hazard ratio of treated vs. untreated patients and the biomarker model gives the effect of treatment on the probability of observing a positive value of the biomarker in the binary part. In the continuous part of the C-TPJM, the effect of treatment (β_{trt}) corresponds to the multiplicative effect on the mean value of the biomarker conditional on observing a positive value. With the M-TPJM, the effect of treatment in the continuous part (β_{trt}) can be interpreted as the multiplicative effect of treatment on the marginal mean biomarker value.

With the CL association, the treatment effect estimated by the biomarker model affects also the survival part. This is a flexible approach that allows the treatment effect to vary over time in the biomarker model, resulting in a non-proportional effect on the survival model. We recommend to use graphical representations to get a clear idea of the time-dependent effect of treatment on survival time with the CL association structure.

We can get an approximation of the marginal effect of treatment on the biomarker with the C-TPJM, but this effect is conditional on the random effects and the value of other covariates included in the model. Moreover, the delta method or resampling techniques must be employed to get a confidence interval and a Wald test on this marginal effect of treatment (See Web Appendix B for more details). We can compute the subject-specific (i.e. conditional on the random effects) time-dependent overall treatment effect (direct + indirect effect) with the CL association. It corresponds to the treatment effect for the average patient, with random effects equal to zero. Moreover, it is possible to compute the average treatment effect in the population (i.e. marginal) from the subject-specific one using Monte-Carlo simulations, as discussed in van Oudenhoven et al. (2020).

2.7 Estimation procedure

The full likelihood of the M-TPJM is given by

$$\begin{aligned}
 L_i(\cdot) &= \int_{a_i} \int_{b_i} \prod_{j=1}^{n_i} \text{Prob}(Y_{ij} > 0)^{U_{ij}} (1 - \text{Prob}(Y_{ij} > 0))^{(1-U_{ij})} \\
 &\quad \times \left\{ \frac{1}{\left(\sqrt{2\pi\sigma_\epsilon^2}\right)} Y_{ij}^{-1} \exp\left(-\frac{(\log(Y_{ij}) - \mu_{ij})^2}{2\sigma_\epsilon^2}\right) \right\}^{U_{ij}} \\
 &\quad \times \lambda_i(T_i|\Theta)^{\delta_i} \exp\left(-\int_0^{T_i} \lambda_i(t|\Theta) dt\right) p(\mathbf{a}_i, \mathbf{b}_i) db_i da_i
 \end{aligned}$$

With $U_{ij} = I[Y_{ij} > 0]$ and $\Theta = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{a}_i, \mathbf{b}_i, \boldsymbol{\gamma}, \boldsymbol{\varphi})$. Details on the construction of the log-likelihood is given in Web Appendix A.

With a M-TPJM, the marginal mean of Y_{ij} is

$$E[Y_{ij}] = \text{Prob}(Y_{ij} > 0) \exp(\mu_{ij} + \sigma_\epsilon^2/2)$$

The positive values of the biomarker are assumed to take a log-normal density: $\log \mathcal{N}(\mu_{ij}, \sigma_\epsilon)$. Using the parameterization from Equation 5, we can derive the corresponding location parameter of the log-normal distribution as

$$\mu_{ij} = \mathbf{X}_{Cij}^\top \boldsymbol{\beta} + \mathbf{Z}_{Cij}^\top \mathbf{b}_i - \log(\text{Prob}(Y_{ij} > 0)) - \sigma_\epsilon^2/2 \quad (7)$$

With a C-TPJM, the likelihood contributions from the binary part and the continuous part are only linked through the random effects correlation structure. The location parameter of the log-normal distribution for the positive values is therefore

$$\mu_{ij} = \mathbf{X}_{Cij}^\top \boldsymbol{\beta} + \mathbf{Z}_{Cij}^\top \mathbf{b}_i - \sigma_\epsilon^2/2, \quad (8)$$

For the left-censoring OPJM, the positive values are assumed to have a log-normal density and the zeros (i.e. censored values), a cumulative log-normal distribution corresponding to the density of probability of a value observed below the threshold (the censoring threshold is chosen as the smallest positive value observed when not provided by investigators).

The baseline hazard in the survival part of the model is approximated with m cubic M-

10

splines with Q knots. A penalization term ensure that the baseline hazard is smooth

$$pl(\Theta) = l(\Theta) - \kappa \int_0^\infty \lambda_0''(t)^2 dt,$$

where $l(\Theta) = \sum_{i=1}^n \log(L_i(\Theta))$ and κ a smoothing parameter chosen using an approximate cross-validation criterion from a separate Cox model.

We propose the Levenberg-Marquardt algorithm (Marquardt (1963)) to maximize the log-likelihood. The integration over the random effects has no analytical solution, therefore we approximate their value with a Monte-Carlo method. The number of points for the Monte-Carlo integration methods defines the tradeoff between the precision of the approximation of the random effects distribution and the computation time.

The approximated likelihood cross-validation (LCV) criterion (Commenges et al. (2007)) for evaluation of the goodness-of-fit of the models can be used as a model choice criterion. It corresponds to the Akaike information criterion (AIC) in the case of the penalized maximum likelihood estimation. The LCV requires the outcome to be the same and while the overall mean is the same between the left-censoring OPJM and the MTPJM, the latter does include the contribution of the binary component into the likelihood, making them not comparable according to this criterion. However, the LCV can compare the goodness-of-fit between the C-TPJM and the M-TPJM as well as the SRE and CL association structures for each type of joint model. The left-censoring OPJM, the C-TPJM and the M-TPJM are estimated with the function *longiPenal* of the R package *frailtypack*, available on the comprehensive R archive network (CRAN).

3 Simulation study

3.1 Simulation study design

We conducted simulation studies to compare the left-censoring OPJM, the conditional TPJM and the marginal TPJM in terms of bias and coverage probabilities. We propose three scenarios where the true model for data generation is either the M-TPJM (scenario 1), the

C-TPJM (scenario 2) or the left-censoring OPJM (scenario 3). The parameters used for the data generation are based on the results from the real data application. For each scenario, 300 datasets are generated with 400 individuals each. We focus on the CL association structure for these simulations since it is a more challenging joint model to estimate (the survival model requires an additional integration step in the optimization procedure). Moreover, the CL association structure provides a slightly better fit compared to the SRE association structure in our application. Besides, simulation studies for a conditional TPJM with SRE association have been proposed in Liu (2009). For the data generation assuming the M-TPJM as the true model, we first generate the zero values from a bernoulli distribution, then the longitudinal biomarker measurements assuming a log-normal distribution for the positive biomarker values, using the location parameter of Equation 7. The longitudinal measurements are generated for the entire follow-up and then we use the R package *PermAlgo* to generate random death times that depends on the time-dependent biomarker value and random censoring times (Sylvestre and Abrahamowicz (2008)). The data generation for the C-TPJM is similar, except that the location parameter does not include the linear predictor from the binary part (Equation 8). The observed value of the biomarker with the C-TPJM is therefore defined by Equation 6, which is non-linear on the log scale. For the one-part model, we generate the longitudinal measurements and then the zero excess with a censoring threshold chosen as the first decile of the distribution. The number of repeated measurements of the biomarker per individual varies between 1 and 16, with a median of 2. The percentage of patients who die during the 4 years follow-up is 80% following the real data death rate. Therefore, most of the biomarker observations are in the early follow-up (the sample size decreases over time as censoring and death occurs). A binary covariate generated from a Bernoulli distribution with $p = 0.5$ corresponding to the treatment effect is included in each

12

submodel of the joint model, with a time-interaction for each submodel of the two-part models.

We use the same parameters as in the application to decide the number of knots of the baseline hazard approximation (5 knots). The penalization term was chosen by cross-validation from an univariate survival model. We use 1000 Monte-Carlo integration points for the numerical approximation of the integral over the random effects distribution.

The parameters of the binary part can be compared between the C-TPJM and the M-TPJM as they both give the effect of covariates on the probability to observe a positive value. The parameters of the continuous part can be compared between the left-censoring OPJM and the M-TPJM as they both give the effect of covariates on the marginal mean of the biomarker. The continuous part of the C-TPJM cannot be directly compared to the continuous part of the other two models. The direct effect of treatment on the risk of death and the association between the biomarker and the risk of death in the survival part can be compared between the three models.

3.2 Results

Results from the simulation study are presented in Tables 1 - 3.

3.2.1 Scenario 1 - True model: M-TPJM

[Table 1 about here.]

The M-TPJM recovers the true parameters value with good accuracy, coverage probabilities are close to 95% (Table 1). Fixed effects parameters for the continuous part of the left-censoring OPJM are biased, with an intercept value $\hat{\beta}_0 = 1.69$ (SD=0.06, CP=6%) where the true value is $\beta_0 = 1.5$. The model is not able to handle properly the excess of zero values. We illustrate this systematic bias with a plot of the estimated mean trajectory of the biomarker compared to the true trajectory under the three scenarios (Figure 2). The time by treatment interaction effect under the left-censoring OPJM is negative ($\hat{\beta}_3 = -0.13$,

SD=0.17, CP=22%) where the true value is positive ($\beta_3 = 0.3$). The binary part of the C-TPJM gives unbiased results. For the survival part, both the left-censoring OPJM and the C-TPJM are able to capture the direct treatment effect on the risk of death ($\hat{\gamma} = -0.16$, SD=0.13, CP=92% for the left-censoring OPJM and $\hat{\gamma} = -0.16$, SD=0.12, CP=91% for the C-TPJM), with a mean value slightly lower than the true value ($\gamma = -0.2$) compared to the M-TPJM ($\hat{\gamma} = -0.18$, SD=0.12, CP=92%). The association between the biomarker and the survival is also unbiased for the left-censoring OPJM ($\hat{\varphi} = 0.09$, SD=0.02, CP=94%) and the C-TPJM ($\hat{\varphi} = 0.08$, SD=0.02, CP=95%) where the true value is $\varphi = 0.08$. The standard deviations of the random effects are properly estimated with the two TPJMs but not with the left-censoring OPJM, the random intercept ($\sigma_{b_0} = 0.6$) is biased downwards ($\hat{\sigma}_{b_0} = 0.45$, SD=0.06) and the random slope ($\sigma_{b_1} = 0.3$) is biased upwards ($\hat{\sigma}_{b_1} = 0.69$, SD=0.12). The correlation between the random intercept and slope in the continuous part ($corr_{b_0b_1} = 0.2$) is biased with the C-TPJM ($c\hat{o}r_{b_0b_1} = -0.20$, SD=0.17) as well as the correlation between the intercept from the binary part and the slope in the continuous part ($corr_{ab_1} = 0.5$), finding almost no correlation ($c\hat{o}r_{ab_1} = 0.07$, SD=0.30).

3.2.2 Scenario 2 - True model: C-TPJM

[Table 2 about here.]

The parameter estimates in the binary part ($\alpha_0 = 6, \alpha_1 = -3, \alpha_2 = 1, \alpha_3 = -2$) are biased with the M-TPJM ($\hat{\alpha}_0 = 5.46$, SD=0.57, CP=69% ; $\hat{\alpha}_1 = -2.34$, SD=0.38, CP=39% ; $\hat{\alpha}_2 = 0.66$, SD=0.74, CP=89% ; $\hat{\alpha}_3 = -1.45$, SD=0.62, CP=69%) while the C-TPJM is unbiased with similar variability (Table 2). This could be due to the correlation between the binary part and the continuous part in the M-TPJM (Equation 7), while they are simulated independent conditional on the random effects. As displayed in Figure 2, the mean behaviour of the biomarker is not linear on the log scale with the C-TPJM as opposed to the left-censoring OPJM and the M-TPJM. In particular, the mean value of the biomarker converges towards zero at the end of the follow-up because the probability of positive value

14

decreases over time in the binary part. The M-TPJM is not able to capture this trend in the late follow-up (i.e. where there are less observations because some patients got censored or died during follow-up). The left-censoring OPJM seems severely biased for this simulation scenario, especially for the time by treatment interaction effect on the marginal mean of the biomarker ($\hat{\beta}_3 = -0.24$, $SD=0.16$). As observed in the first scenario, the direct effect of treatment on the risk of event and the association parameter is properly recovered for the three models except the left-censoring OPJM with a slightly higher estimate and standard error for the association ($\hat{\varphi} = 0.10$, $SD=0.03$) than the true value recovered by the M-TPJM and the C-TPJM ($\hat{\varphi} = 0.08$, $SD=0.02$) while coverage probabilities are close to 95% with the three models. The standard deviations of the random effects and their correlation is properly captured with the C-TPJM and similarly with the M-TPJM while the left-censoring OPJM exhibits a similar bias as in scenario 1.

3.2.3 Scenario 3 - True model: Left-censoring OPJM

[Table 3 about here.]

The convergence rate of the C-TPJM (73%) is low, this is due to the data generating mechanism that gives unstable parameter estimates in the binary part (the probability of positive value at baseline is close to 1, corresponding to a linear predictor that converges towards $+\infty$ with the logit link function). Fixing the intercept to a reasonable value of 6.0 solves this issue while not changing the parameters estimates. As expected, the M-TPJM gives unbiased values for the continuous part (Table 3), although we notice slightly lower coverage probabilities for the fixed slope effect ($\hat{\beta}_1 = -0.55$, $SD=0.06$, $CP=83\%$), the interaction of the slope and treatment ($\hat{\beta}_3 = 0.35$, $SD=0.08$, $CP=86\%$) and the error term ($\hat{\sigma}_\epsilon = 0.30$, $SD=0.01$, $CP=75\%$). In the survival part, all the models are again unbiased, with similar precision and coverage. The random intercept and slope are properly estimated but their correlation is slightly biased upwards with the M-TPJM ($c\hat{o}r r_{b_0 b_1} = 0.44$, $SD=0.18$, true value is $c\hat{o}r r_{b_0 b_1} = 0.2$) while the C-TPJM finds no correlation between the intercept

and slope ($\hat{corr}_{b_0b_1} = -0.01$, $SD=0.21$). We also notice a strong correlation between the random intercept from the binary and continuous parts for both the C-TPJM ($c\hat{orr}_{ab_0} = 0.93$, $SD=0.04$) and the M-TPJM ($c\hat{orr}_{ab_0} = 0.94$, $SD=0.04$), this is due to the data generating mechanism where censored values are the 10% smallest observed values, thus inducing a strong correlation between the mean value of the biomarker and the probability of positive value.

[Figure 2 about here.]

3.2.4 Conclusion

To conclude, the left-censoring OPJM gives biased estimates of the mean biomarker value and evolution over time when the true model is either the C-TPJM or the M-TPJM. The M-TPJM provides an accurate estimate of the biomarker trajectory under scenarios 1 and 3 but not under scenario 2, as expected (see Figure 2). In our scenarios, the association between the biomarker and the survival was driven largely by early follow-up (where censorship rate is low), thus the parameter quantifying this association was not affected by the bias of the mean biomarker value observed in the late follow-up (see Figure 2). The assumption of independence between the binary and continuous parts conditionally on the random effects with the C-TPJM can result in unstable parameter estimations and convergence issues when this assumption does not hold, as observed in scenario 3. Finally, the C-TPJM is not able to recover the correlation between the random effects when it is not the true model while the M-TPJM gives a good approximation of this correlation structure in all three scenarios.

4 Application to metastatic head and neck cancer data

4.1 Data

The study consists of a phase 3 randomised clinical trial (RCT) of chemotherapy with or without panitumumab in patients with metastatic and/or recurrent squamous cell carcinoma of the head and neck (SCCHN). The objective of the study is to compare the treatment

16

effect of panitumumab in combination with chemotherapy versus chemotherapy alone as first line therapy for metastatic and/or recurrent SCCHN. This dataset is freely available on ProjectDataSphere.org (Project Data Sphere is an initiative to provide access to individual patient data from RCTs across numerous cancer types from industry and academia).

Between May 15, 2007 and March 10, 2009, 657 patients were randomly assigned (327 to the panitumumab group and 330 to the control group). The data for analysis includes a subset of 449 patients (i.e., 137 patients excluded from the publicly available dataset and out of them, 71 had no biomarker measurements). The median overall survival (OS) is 0.61 for the control group (arm A) and 0.81 for the panitumumab group (arm B), 370 patients (82%) died during follow-up. There are 1913 repeated measurements of the SLD, 161 of which are zero values (8%). The number of individual repeated measurements for this biomarker varies between 1 and 29 with a median of 4. The main conclusion of the trial was that the addition of panitumumab to chemotherapy did not improve the OS but it improved the progression-free survival (PFS) and had an acceptable toxicity (Vermorken et al. (2013)). However a better PFS does not always lead to improved OS (Prasad et al. (2015)).

We chose 5 knots for the splines approximating the baseline hazard function based on an univariate survival model. The penalization term, found with cross-validation, is $\kappa = 0.02$. We use 2000 Monte-Carlo integration points for the numerical approximation of the integral over the random effects.

4.2 Results from the M-TPJM

In this RCT, there is no zero value at baseline as all patients have at least one measurable lesion at inclusion. For that reason, the estimation of the intercept in the binary part of the two-part models is unstable and led to convergence issues. We therefore decided to fix the intercept value at 8.0, which corresponds to a baseline mean probability of zero value of

3×10^{-4} . The results of the M-TPJM with the CL and the SRE association are presented in Table 4.

4.2.1 Binary part

The treatment effect at baseline is positive and slightly significantly different from zero with the CL association ($\hat{\alpha}_{trt} = -1.00$, SE= 0.44) and the SRE association ($\hat{\alpha}_{trt} = -0.94$, SE= 0.51). In a RCT, a significant treatment effect at baseline could result from a bias in randomization (i.e., since we only used a subset of individuals) or a lack of flexibility in the function describing the evolution of the outcome over time. The slope effect of time is negative and highly significant with the CL ($\hat{\alpha}_{time} = -3.67$, SE= 0.37) and the SRE associations ($\hat{\alpha}_{time} = -3.38$, SE= 0.47). This means that the probability of observing a zero value (i.e., complete remission of the measured tumors) increases over time for the reference treatment (arm A). The time by treatment interaction effect on the probability of observing a positive value is significantly negative for both the CL ($\hat{\alpha}_{time \cdot trt} = -2.02$, SE= 0.49) and SRE associations ($\hat{\alpha}_{time \cdot trt} = -1.54$, SE= 0.57), meaning that the patients receiving treatment arm B are associated with higher odds of zero value over time compared to patients receiving treatment arm A.

4.2.2 Continuous part

The marginal mean value of the SLD at baseline is found similar between the two treatment arms ($\beta_0 \simeq 1.4$). The slope effect of time is negative and significant (CL: $\hat{\beta}_{time} = -0.68$, SE= 0.08 and SRE: $\hat{\beta}_{time} = -0.61$, SE= 0.07). This effect can be interpreted as a multiplicative time effect on the marginal mean of the biomarker given by $\exp(-0.68) = 0.51$ for the CL association (respectively $\exp(-0.61) = 0.54$ for the SRE association). This corresponds to a reduction of 49% of the SLD value per year for patients receiving the reference treatment (arm A) with the CL association model (respectively a reduction of 46% per year with the SRE association model). The two M-TPJMs do not find a significant treatment effect at

18

baseline nor time by treatment interaction, therefore patients receiving treatment arm B have a similar decreasing trend of SLD over time than those in arm A.

4.2.3 *Survival part*

The interpretation of the covariate effects in the survival part depends on the association structure specified for the TPJM. With the SRE association, the parameters are interpreted in terms of effect on the risk of death adjusted for some individual heterogeneity of the population (captured by the random effects of the biomarker model). With the CL association, the effect of covariates is decomposed into a direct effect on the time to death and an indirect effect through their association with the biomarker, whose current value affects the terminal event. The direct treatment effect is not significantly different from zero neither with the SRE association structure ($\hat{\gamma} = -0.07$, SE=0.11) nor with the CL association ($\hat{\gamma} = -0.05$, SE=0.11). In terms of indirect effect, the treatment and treatment by time interaction are not significant with either the CL or SRE associations. For the SRE association, the association between the individual heterogeneity at baseline (random intercept from the continuous part) and the risk of death is positive and slightly significant ($\hat{\varphi}_{b_0} = 0.42$, SE= 0.19), indicating that the baseline value of the SLD is predictive of the risk of death. However, for the CL association, the current value of the biomarker is positively and very significantly associated with the terminal event ($\hat{\varphi} = 0.08$, SE= 0.01), indicating that the risk of death increases with the value of the SLD where the probability of a positive SLD value is decreasing with time and at a higher rate for treatment arm B vs. A. The M-TPJM with CL association suggests therefore a possible indirect effect of the biomarker only through its binary component, that is the probability of a zero value, interpreted as a complete remission of the tumor. However, this model did not conclude to an indirect effect of treatment on the overall biomarker trajectory, i.e., when accounting for the continuous part in addition to the zero part of the biomarker. Moreover, Web Figure 1 shows no difference in the survival curves according to treatment arm. Another example of such graphical representation of the total treatment

effect (i.e. direct + indirect effect) is proposed in Rustand et al. (2020). The other advantage of the M-TPJM with the CL association, is that it allows to quantify the effect of one unit increase in the biomarker on the risk of terminal event. For instance, the hazard ratio of a patient with a 1 cm increase in the SLD value is associated with an increased risk of death of 8% ($\exp(0.08) = 1.08$). The M-TPJM with the SRE association can be helpful to characterize how individuals who deviate by a certain amount from the mean SLD trajectory (e.g., 1 standard deviation of the baseline biomarker value) have an increased risk of terminal event compared to a patient with an average SLD profile.

4.3 Comparison of the M-TPJM with the left-censoring OPJM

The difference in the treatment of the zero values lead to a steeper fixed slope effect on the mean biomarker value for the left-censoring OPJM (CL: $\hat{\beta}_{time} = -0.87$, SE= 0.10, SRE: $\hat{\beta}_{time} = -0.86$, SE= 0.10) compared to the M-TPJM (CL: $\hat{\beta}_{time} = -0.68$, SE= 0.08, SRE: $\hat{\beta}_{time} = -0.61$, SE= 0.07). The residual error is higher with the left-censoring OPJM ($\hat{\sigma}_{\varepsilon} = 0.41$, SE= 0.01) than with the M-TPJM ($\hat{\sigma}_{\varepsilon} = 0.30$, SE= 0.01), indicating a better fit of the latter. Our results are therefore in line with our simulations when OPJM is not the true model. Overall, the standard errors of the parameter estimates in the biomarker model are lower under the M-TPJM than the OPJM. Nonetheless, the SRE association shows a stronger relationship between the random intercept and slope of the mean biomarker value with the left-censoring OPJM ($\hat{\varphi}_{b_0} = 0.48$, SE=0.11, $\hat{\varphi}_{b_1} = 0.16$, SE=0.06) compared to the M-TPJM ($\hat{\varphi}_a = 0.00$, SE=0.07, $\hat{\varphi}_{b_0} = 0.42$, SE=0.19, $\hat{\varphi}_{b_1} = 0.30$, SE=0.23), likely due to the better fit of the mean biomarker value with the M-TPJM. In particular, the random slope standard deviation is found higher with the left-censoring OPJM (CL: $\hat{\sigma}_{b_1} = 1.51$, SRE= $\hat{\sigma}_{b_1} = 1.44$) compared to the M-TPJM (CL: $\hat{\sigma}_{b_1} = 0.99$, SRE= $\hat{\sigma}_{b_1} = 1.01$).

4.4 Comparison of the M-TPJM with the C-TPJM

The LCV criterion indicates that the M-TPJM fits better the data than the C-TPJM for each association structure. As proposed in Commenges et al. (2007), the comparison of the LCV value can be classified according to the order of the difference. A difference of order 10^{-1} , 10^{-2} , 10^{-3} and 10^{-4} may be qualified as ‘large’, ‘moderate’, ‘small’ and ‘negligible’, respectively. The difference in the LCV value between the M-TPJM (CL: LCV=1.0072, SRE: LCV=1.0082) and the C-TPJM (CL: LCV=1.0525, SRE: LCV=1.0524) is moderate in favour of the M-TPJM and the difference between the CL and the SRE association structures is small in favour of the CL with the left-censoring OPJM (CL: LCV=1.9790, SRE: LCV=1.9803) and the M-TPJM where it is negligible with the C-TPJM. We plotted the mean biomarker trajectory estimated by the left-censoring OPJM, the C-TPJM and the M-TPJM in Web Figure 2. As observed in the simulations when the M-TPJM is the true model, the C-TPJM tends to over-estimate the probability of zeros over time.

In line with the simulation results when M-TPJM is the true model, the variability of the parameter estimates in the binary part is lower under the M-TPJM than the C-TPJM. Treatment arm B (chemotherapy + panitumumab) vs. arm A (chemotherapy alone) is associated with a more significant reduction in the probability of positive value over time with the M-TPJM (CL: $\hat{\alpha}_{time\cdot trt} = -2.02$, SE=0.49, SRE: $\hat{\alpha}_{time\cdot trt} = -1.54$, SE=0.57) compared to the C-TPJM with the CL association structure ($\hat{\alpha}_{time\cdot trt} = -1.84$, SE=0.71) and the SRE association ($\hat{\alpha}_{time\cdot trt} = -1.83$, SE=1.31). In the continuous part, the effect of treatment is not found significantly different from zero under either the C-TPJM or the M-TPJM but its interpretation is different under these 2 models, as illustrated in Figure 1. The M-TPJM finds no treatment effect on the overall mean biomarker value (CL: $\hat{\beta}_{time\cdot trt} = -0.12$, SE=0.12, SRE: $\hat{\beta}_{time\cdot trt} = -0.13$, SE=0.11) where the C-TPJM finds no treatment effect on the biomarker value conditional on a positive value (CL: $\hat{\beta}_{time\cdot trt} = -0.10$, SE=0.09, SRE:

$\hat{\beta}_{time:trt} = 0.07, SE=0.09$). As found in our simulation results, the direct effect of treatment and association parameter are very similar between the M-TPJM and C-TPJM.

To conclude, as observed in our simulation results, the C-TPJM can lead to biased estimates and incorrect statistical inference when it is not the true model (the best model in the application). In term of treatment effect, this affects mostly the inference about the indirect effect of the treatment on the terminal event.

[Table 4 about here.]

5 Discussion

We proposed a marginal two-part joint model for a longitudinal semicontinuous biomarker and a terminal event that allows to obtain directly the population average effect of covariates, such as treatment effect, on the marginal mean of the biomarker. This M-TPJM is as an alternative to the conditional two-part joint model. While the mean biomarker value at baseline and over time is directly estimated from the M-TPJM, it is obtained from the mixing distributions of the zero and non-zero components in the C-TPJM, which imposes a non-linear curve for the mean biomarker value over time, not always justified. The population average effect of covariates under the C-TPJM is also not directly available. We also proposed two association structures to link the biomarker to the risk of terminal event. The first one, the current value association, allows to explore time-dependent effect of covariates on survival through the biomarker (indirect effect), as well as direct effect of covariates on the terminal event. The model can evaluate survival conditionally on a specific pattern of clinical responses. The second one consists of sharing only the random effects from the two-part model, which evaluates the relationship between the risk of terminal event and the individual deviation from the population mean of the biomarker, including baseline odds of a positive value, baseline value and slope for the whole trajectory. The M-TPJM could therefore be relevant in many clinical applications.

Our simulation studies shows marked differences across the three models applied: the left-censoring OPJM, the C-TPJM and the M-TPJM. The left-censoring OPJM was severely biased in the estimation of treatment effect on the biomarker when true zero values (i.e. not censored) were present. The C-TPJM can account for excess of zeros but led to biased estimates and wrong inference about treatment effect on the marginal mean value of the biomarker whenever it was not the true model. In addition, the C-TPJM could have convergence issues due to the assumed independence between the probability of zero and the expected value among positive conditional on the random effects. The M-TPJM provided an accurate inference about the biomarker and covariate effects on the biomarker in most situations, unless the distribution of the biomarker over time is not linear on the log scale.

The differences observed across models did impact the inference about the indirect effect of treatment on the terminal event but to a lesser extent, the direct association of treatment on the terminal event and the association between the biomarker and the terminal event. This could be the consequence of the heavy censoring present in the simulated data (which mimicked the real data) and the fact that the estimated mean value of the biomarker was relatively close across models during the early follow-up (see Figure 2). In other situations with lower censoring rate or higher proportion of zeros, it is not excluded that the direct treatment effect and association parameter(s) be also affected by the model assumptions.

Our application to a cancer clinical trial assessing two treatment arms for squamous cell carcinoma of the head and neck illustrates the interest of the M-TPJM. We recall that the original trial concluded that the addition of panitumumab to chemotherapy did not improve OS but led to better progression-free survival (Vermorken et al. (2013)). In contrast, the M-TPJM concluded to a possible indirect effect of the combined treatment vs. single treatment on the risk of death, where this indirect effect was mainly explained by a higher probability of observing a zero biomarker value for the combined treatment, that is higher odds of

observing a disappearance of all target lesions, compared to chemotherapy alone. Based on LCV criterion, the M-TPJM fitted the data better than the C-TPJM. In line with our simulation results when the M-TPJM was the true model, the C-TPJM could lead to bias in the treatment effect on the biomarker and false inference about the indirect effect of treatment on the terminal effect. In particular, the C-TPJM found an attenuated indirect effect compared to the M-TPJM in terms of observing a disappearance of all target lesions over time. Finally, the M-TPJM did not conclude to a treatment effect on the overall mean of the biomarker either at baseline or during the follow-up.

This work has several limitations. For instance, in the cancer clinical trial application, the SLD measures the longest diameter of target lesions, which can be subject to important measurement error. It could be more appropriate to use instead a more accurate measurement such as the total volume of the tumors, although it is usually unavailable in clinical trials as it is not part of the RECIST criteria. In this work, the M-TPJM was not developed specifically to capture a non-linear mean biomarker trajectory on the log scale. The inclusion of time-dependent covariates and interactions could account for such trajectories. Besides, an extension of the marginal two-part model has been proposed using a generalized Gamma distribution (that includes the logarithm as a specific case) to link the outcome to the linear predictor in the continuous part and allows more flexibility in the biomarker trajectory but has not yet been developed for joint models (Smith and Preisser (2017); Smith et al. (2017)).

Beyond the application to solid tumor cancer data, we propose a tool that can be applied to several other situations that include a longitudinal semicontinuous biomarker and survival times (covariates measuring symptoms of a disease or quantifying exposure are often semicontinuous). To our knowledge, this is the first software (*frailtypack*, Król et al. (2017)) available that proposes to fit a marginal TPJM.

ACKNOWLEDGEMENTS

This publication is based on research using information obtained from www.projectdatasphere.org, which is maintained by Project Data Sphere, LLC. Neither Project Data Sphere, LLC nor the owner(s) of any information from the web site have contributed to, approved or are in any way responsible for the contents of this publication.

REFERENCES

- Commenges, D., Joly, P., Gégout-Petit, A., and Liqueur, B. (2007). Choice between semi-parametric estimators of Markov and non-Markov multi-state models from coarsened observations. *Scandinavian Journal of Statistics* **34**, 33–52.
- Król, A., Mauguen, A., Mazroui, Y., Laurent, A., Michiels, S., and Rondeau, V. (2017). Tutorial in joint modeling and prediction: A statistical software for correlated longitudinal outcomes, recurrent events and a terminal event. *Journal of Statistical Software* **81**, 1–52.
- Liu, L. (2009). Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Statistics in Medicine* **28**, 972–986.
- Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics* **11**, 431–441.
- Prasad, V., Kim, C., Burotto, M., and Vandross, A. (2015). The Strength of Association Between Surrogate End Points and Survival in Oncology: A Systematic Review of Trial-Level Meta-analyses. *JAMA Internal Medicine* **175**, 1389–1398.
- Rustand, D., Briollais, L., Tournigand, C., and Rondeau, V. (2020). Two-part joint model for a longitudinal semicontinuous marker and a terminal event with application to metastatic colorectal cancer data. *Biostatistics* kxaa012.
- Smith, V. A., Maciejewski, M. L., and Olsen, M. K. (2018). Modeling semicontinuous longitudinal expenditures: A practical guide. *Health Services Research* **53**, 3125–3147.

- Smith, V. A., Neelon, B., Maciejewski, M. L., and Preisser, J. S. (2017). Two parts are better than one: modeling marginal means of semicontinuous data. *Health Services and Outcomes Research Methodology* **17**, 198–218.
- Smith, V. A., Neelon, B., Preisser, J. S., and Maciejewski, M. L. (2017). A marginalized two-part model for longitudinal semicontinuous data. *Statistical Methods in Medical Research* **26**, 1949–1968.
- Smith, V. A. and Preisser, J. S. (2017). Direct and flexible marginal inference for semicontinuous data. *Statistical Methods in Medical Research* **26**, 2962–2965.
- Smith, V. A., Preisser, J. S., Neelon, B., and Maciejewski, M. L. (2014). A marginalized two-part model for semicontinuous data. *Statistics in Medicine* **33**, 4891–4903.
- Sylvestre, M.-P. and Abrahamowicz, M. (2008). Comparison of algorithms to generate event time conditional on time-dependent covariates. *Statistics in medicine* **27**, 2618–34.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* **26**, 24–36.
- van Oudenhoven, F. M., Swinkels, S. H., Ibrahim, J. G., and Rizopoulos, D. (2020). A marginal estimate for the overall treatment effect on a survival outcome within the joint modeling framework. *Statistics in Medicine* sim.8713.
- Vermorken, J. B., Stöhlmacher-Williams, J., Davidenko, I., Licitra, L., Winqvist, E., Vilanueva, C., Foa, P., Rottey, S., Skladowski, K., Tahara, M., et al. (2013). Cisplatin and fluorouracil with or without panitumumab in patients with recurrent or metastatic squamous-cell carcinoma of the head and neck (spectrum): an open-label phase 3 randomised trial. *The lancet oncology* **14**, 697–710.

SUPPORTING INFORMATION

Web Appendices and Figures referenced in Sections (2.6, 2.7, 4.2 and 4.4) are available with this paper in the Supporting Information.

26

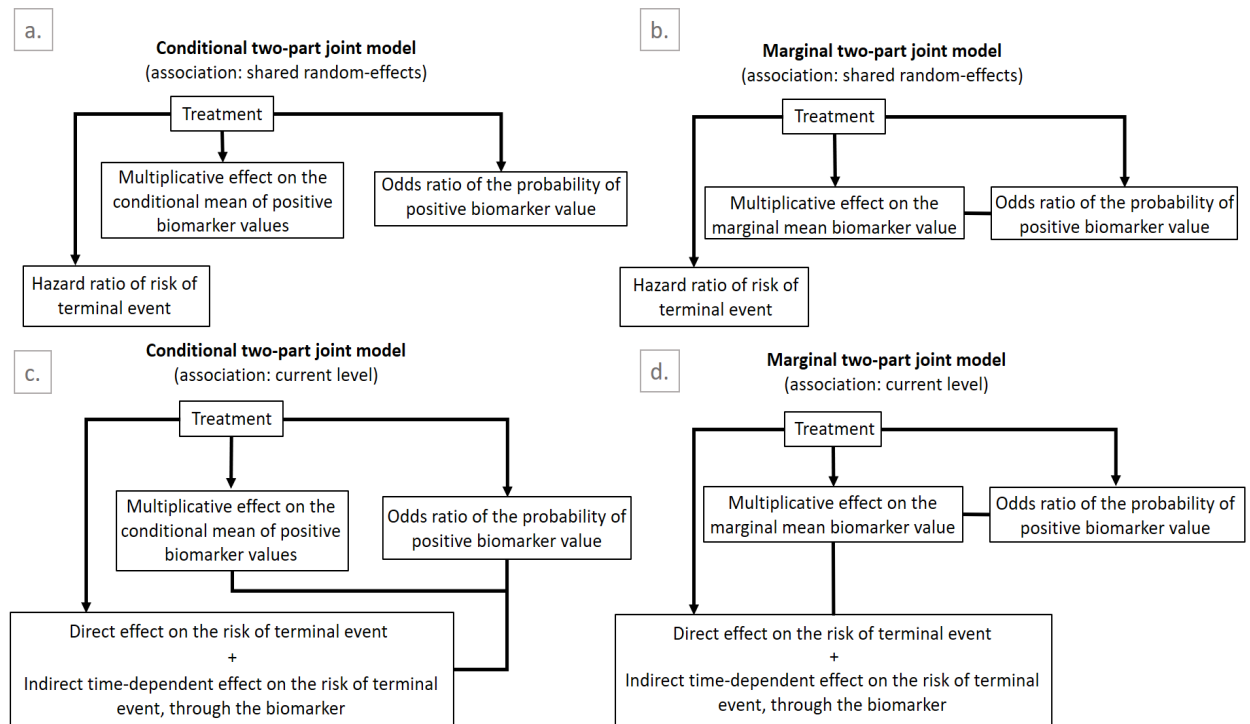


Figure 1. Diagrams describing the treatment effect decomposition with the C-TPJM (left) and the M-TPJM (right) for the shared random effects (up) and the current level (down) association structures. The marginal model includes zero values to give the effect of treatment on the marginal mean biomarker value. The current level association structure shares the treatment effect captured in the biomarker model with the survival model and therefore provides a decomposition into a direct and an indirect effect of treatment on the risk of event.

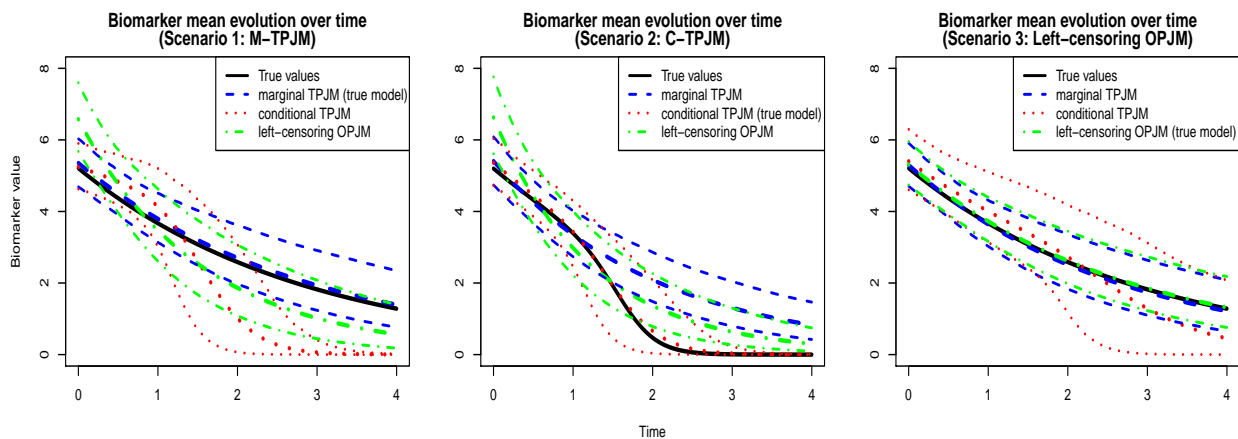


Figure 2. Mean biomarker trajectory captured in the simulation studies (Tables 1-3) for the M-TPJM, the C-TPJM and the left-censoring OPJM compared to the true trajectory. (This figure appears in color in the electronic version of this article.)

Table 1

Summary of the results of simulations scenario 1 (true model : marginal TPJM) ; 300 datasets with 400 individuals each and 1000 integration points, 10.17% zeros on average ($SD=1.33$). The true value of the parameters estimated in the continuous part of the C-TPJM are unknown, therefore coverage probabilities are not provided for these parameters.

Variable (appli)	Left-censoring OPJM Est.* (SD) [†] [CP] [‡]	conditional TPJM Est. (SD) [CP]	marginal TPJM Est. (SD) [CP]	
Binary part				
intercept	$\alpha_0 = 6$	6.09 (0.64) [96%]	6.11 (0.57) [94%]	
time	$\alpha_1 = -3$	-3.04 (0.44) [96%]	-3.04 (0.35) [93%]	
treatment	$\alpha_2 = 1$	0.93 (0.75) [96%]	0.94 (0.68) [96%]	
time:treatment	$\alpha_3 = -2$	-1.93 (0.65) [94%]	-1.95 (0.53) [94%]	
Continuous part				
intercept	$\beta_0 = 1.5$	1.69 (0.06) [06%]	1.52 (0.05)	1.53 (0.05) [90%]
time	$\beta_1 = -0.5$	-0.58 (0.10) [93%]	-0.35 (0.03)	-0.50 (0.06) [93%]
treatment	$\beta_2 = 0.3$	0.38 (0.08) [77%]	0.28 (0.07)	0.30 (0.07) [93%]
time:treatment	$\beta_3 = 0.3$	-0.13 (0.17) [22%]	0.42 (0.09)	0.30 (0.08) [95%]
residual S.E.	$\sigma_\epsilon = 0.3$	0.64 (0.07) [00%]	0.32 (0.01)	0.30 (0.01) [92%]
Survival part				
treatment	$\gamma = -0.2$	-0.16 (0.13) [92%]	-0.16 (0.12) [91%]	-0.18 (0.12) [92%]
association	$\varphi = 0.08$	0.09 (0.02) [94%]	0.08 (0.02) [95%]	0.08 (0.02) [95%]
Random effects				
intercept (binary part)	$\sigma_a = 1.4$		1.33 (0.28)	1.37 (0.28)
intercept (continuous part)	$\sigma_{b_0} = 0.6$	0.45 (0.06)	0.62 (0.03)	0.61 (0.03)
slope (continuous part)	$\sigma_{b_1} = 0.3$	0.69 (0.12)	0.33 (0.07)	0.28 (0.08)
	$cor_{ab_0} = 0.5$		0.51 (0.17)	0.56 (0.16)
	$cor_{ab_1} = 0.5$		0.07 (0.30)	0.45 (0.30)
	$cor_{b_0b_1} = 0.2$	0.20 (0.23)	-0.20 (0.17)	0.27 (0.24)
Convergence rate	100%	100%	100%	100%

* Mean of parameter estimates; [†] Standard deviation from the mean; [‡] Coverage probability

Table 2

Summary of the results of simulations scenario 2 (true model : conditional TPJM) ; 300 datasets with 400 individuals each and 1000 integration points, 10.53% zeros on average ($SD=1.36$). The true value of the parameters estimated in the continuous part of the left-censoring OPJM and the M-TPJM are unknown, therefore coverage probabilities are not provided for these parameters.

Variable (appli)	Left-censoring OPJM	conditional TPJM	marginal TPJM
	Est.* (SD [†]) [CP [‡]]	Est. (SD) [CP]	Est. (SD) [CP]
Binary part			
intercept	$\alpha_0 = 6$	6.13 (0.64) [96%]	5.46 (0.57) [69%]
time	$\alpha_1 = -3$	-3.07 (0.45) [95%]	-2.34 (0.38) [39%]
treatment	$\alpha_2 = 1$	1.03 (0.85) [96%]	0.66 (0.74) [89%]
time:treatment	$\alpha_3 = -2$	-2.04 (0.72) [95%]	-1.45 (0.62) [69%]
Continuous part			
intercept	$\beta_0 = 1.5$	1.68 (0.07)	1.53 (0.05) [90%]
time	$\beta_1 = -0.5$	-0.68 (0.10)	-0.50 (0.06) [90%]
treatment	$\beta_2 = 0.3$	0.41 (0.08)	0.30 (0.07) [91%]
time:treatment	$\beta_3 = 0.3$	-0.24 (0.16)	0.30 (0.08) [94%]
residual S.E.	$\sigma_\epsilon = 0.3$	0.63 (0.08)	0.30 (0.01) [94%]
Survival part			
treatment	$\gamma = -0.2$	-0.21 (0.13) [95%]	-0.20 (0.12) [95%]
association	$\varphi = 0.08$	0.10 (0.03) [92%]	0.08 (0.02) [94%]
Random effects			
intercept (binary part)	$\sigma_a = 1.4$		1.35 (0.29)
intercept (continuous part)	$\sigma_{b_0} = 0.6$	0.47 (0.07)	0.61 (0.03)
slope (continuous part)	$\sigma_{b_1} = 0.3$	0.80 (0.16)	0.29 (0.05)
	$cor_{ab_0} = 0.5$		0.53 (0.16)
	$cor_{ab_1} = 0.5$		0.51 (0.25)
	$cor_{b_0b_1} = 0.2$	0.18 (0.19)	0.20 (0.19)
Convergence rate	100%	100%	100%

* Mean of parameter estimates; [†] Standard deviation from the mean; [‡] Coverage probability

Table 3

Summary of the results of simulations scenario 3 (true model : Left-censoring OPJM) ; 300 datasets with 400 individuals each and 1000 integration points, 10.04% zeros on average (SD=0.02). The true value of the parameters estimated in the continuous part of the C-TPJM are unknown, therefore coverage probabilities are not provided for these parameters.

Variable (appli)	Left-censoring OPJM	conditional TPJM	conditional TPJM	marginal TPJM	
	Est.* (SD) [†] [CP] [‡]	Est. (SD) [CP]	Est. (SD) [CP]	Est. (SD) [CP]	
Binary part					
intercept	α_0	7.89 (0.78)	6.00 (fixed)	5.95 (0.81)	
time	α_1	-3.50 (0.55)	-2.73 (0.34)	-2.47 (0.47)	
treatment	α_2	2.52 (0.90)	2.96 (0.83)	1.60 (0.62)	
time:treatment	α_3	1.18 (0.76)	0.63 (0.63)	0.72 (0.49)	
Continuous part					
intercept	$\beta_0 = 1.5$	1.52 (0.05) [93%]	1.54 (0.04)	1.50 (0.05)	1.51 (0.05) [96%]
time	$\beta_1 = -0.5$	-0.51 (0.05) [94%]	-0.42 (0.05)	-0.44 (0.05)	-0.55 (0.06) [83%]
treatment	$\beta_2 = 0.3$	0.30 (0.06) [94%]	0.29 (0.06)	0.34 (0.06)	0.32 (0.07) [93%]
time:treatment	$\beta_3 = 0.3$	0.31 (0.08) [92%]	0.23 (0.07)	0.24 (0.07)	0.35 (0.08) [86%]
residual S.E.	$\sigma_\epsilon = 0.3$	0.30 (0.01) [93%]	0.29 (0.01)	0.29 (0.01)	0.30 (0.01) [75%]
Survival part					
treatment	$\gamma = -0.2$	-0.21 (0.13) [95%]	-0.21 (0.13) [95%]	-0.20 (0.13) [94%]	-0.21 (0.13) [95%]
association	$\varphi = 0.08$	0.08 (0.02) [92%]	0.08 (0.02) [93%]	0.08 (0.02) [93%]	0.08 (0.02) [93%]
Random effects					
intercept (binary part)	σ_a		4.53 (0.44)	3.62 (0.27)	2.82 (0.42)
intercept (continuous part)	$\sigma_{b_0} = 0.6$	0.60 (0.03)	0.59 (0.03)	0.59 (0.03)	0.62 (0.03)
slope (continuous part)	$\sigma_{b_1} = 0.3$	0.30 (0.05)	0.21 (0.05)	0.22 (0.05)	0.29 (0.06)
	cor_{ab_0}		0.93 (0.04)	0.93 (0.04)	0.94 (0.04)
	cor_{ab_1}		0.33 (0.21)	0.33 (0.21)	0.69 (0.13)
	$cor_{b_0b_1} = 0.2$	0.22 (0.18)	-0.01 (0.21)	-0.02 (0.21)	0.44 (0.18)
Convergence rate		100%	73%	100%	100%

* Mean of parameter estimates; [†] Standard deviation from the mean; [‡] Coverage probability

Table 4

Model Association	left-censoring OPJM		conditional TPJM		marginal TPJM	
	Current level Est. (SE)	Shared random effects Est. (SE)	Current level Est. (SE)	Shared random effects Est. (SE)	Current level Est. (SE)	Shared random effects Est. (SE)
Binary part (SLD>0 versus SLD=0)						
intercept		8.00 (fixed)	8.00 (fixed)	8.00 (fixed)	8.00 (fixed)	8.00 (fixed)
time (year)		-3.41*** (0.78)	-3.27* (0.95)	-3.67*** (0.37)	-3.38*** (0.47)	-3.38*** (0.47)
treatment (B/A)		-0.95 (0.74)	-0.40 (0.89)	-1.00* (0.44)	-0.94 ^o (0.51)	-0.94 ^o (0.51)
age (> 65 years)		-1.14 (0.81)	-1.77* (0.79)	1.58*** (0.37)	-0.69 (0.46)	-0.69 (0.46)
sex (M/F)		2.74*** (0.83)	2.58** (0.84)	2.46*** (0.47)	2.06** (0.68)	2.06** (0.68)
time:treatment (B/A)		-1.84** (0.71)	-1.83 (1.31)	-2.02*** (0.49)	-1.54** (0.57)	-1.54** (0.57)
Continuous part						
intercept	1.46*** (0.10)	1.46*** (0.10)	1.43*** (0.09)	1.37*** (0.08)	1.37*** (0.09)	1.37*** (0.09)
time (years)	-0.87*** (0.10)	-0.86*** (0.10)	-0.62*** (0.07)	-0.68*** (0.08)	-0.61*** (0.07)	-0.61*** (0.07)
treatment (B/A)	-0.05 (0.06)	-0.04 (0.06)	-0.04 (0.06)	-0.03 (0.06)	-0.01 (0.06)	-0.01 (0.06)
age (> 65 years)	0.11 (0.08)	0.11 (0.08)	0.09 (0.08)	0.10 (0.07)	0.03 (0.08)	0.03 (0.08)
sex (M/F)	0.32*** (0.09)	0.32*** (0.09)	0.30*** (0.10)	0.33*** (0.09)	0.38*** (0.09)	0.38*** (0.09)
time:treatment (B/A)	-0.06 (0.14)	-0.05 (0.14)	-0.10 (0.09)	-0.12 (0.12)	-0.13 (0.11)	-0.13 (0.11)
residual S.E.	0.41 (0.01)	0.41 (0.01)	0.31 (0.01)	0.31 (0.01)	0.30 (0.01)	0.30 (0.01)
Death risk						
treatment (B/A)	-0.06 (0.11)	-0.04 (0.11)	-0.05 (0.11)	-0.07 (0.11)	-0.05 (0.11)	-0.07 (0.11)
age (> 65 years)	0.18 (0.14)	0.29* (0.14)	0.20 (0.14)	0.24 (0.14)	0.18 (0.14)	0.23 (0.14)
sex (M/F)	0.14 (0.17)	0.28 (0.17)	0.14 (0.17)	0.30 (0.17)	0.14 (0.17)	0.29 (0.17)
Association						
$\varphi(E[Y_{ij}])$	0.08*** (0.01)	0.08*** (0.01)	0.08*** (0.01)	0.08*** (0.01)	0.08*** (0.01)	0.08*** (0.01)
φ_a (Binary part intercept)			0.04 (0.03)			0.00 (0.07)
φ_{b_0} (Continuous part intercept)		0.48*** (0.11)	0.39*** (0.12)			0.42* (0.19)
φ_{b_1} (Continuous part slope)		0.16* (0.06)	0.06 (0.16)			0.30 (0.23)
Random effects						
intercept (binary part, σ_a)			4.65	4.65	4.47	4.08
intercept (continuous part, σ_{b_0})	0.59	0.59	0.62	0.62	0.61	0.63
slope (continuous part, σ_{b_1})	1.51	1.44	0.72	0.69	0.99	1.01
$corr^{ab_0}$			0.12	0.37	0.56	0.59
$corr^{ab_1}$			0.52	0.36	0.66	0.80
$corr^{b_0b_1}$	0.26	0.26	-0.06	0.13	0.25	0.31
LCV	1.9790	1.9803	1.0525	1.0524	1.0072	1.0082

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ^o $p < 0.1$

Supporting Information for “A marginal two-part joint model for a longitudinal biomarker and a terminal event with application to head and neck cancer data”

Denis Rustand*

Biostatistic Team, Bordeaux Population Health Center, INSERM U1219,
146 rue Léo Saignat, 33076 Bordeaux, France

**email:* denis@rustand.fr

and

Laurent Briollais

Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital and
Dalla Lana School of Public Health (Biostatistics), University of Toronto,
600 University Ave., Ontario M5G 1X5, Canada

and

Virginie Rondeau

Biostatistic Team, Bordeaux Population Health Center, INSERM U1219,
146 rue Léo Saignat, 33076 Bordeaux, France

Web Appendix A Details on the likelihood of the model

The full likelihood of the model can be expressed as

$$L_i(\cdot) = \int_{\mathbf{a}_i} \int_{\mathbf{b}_i} L_i^B(\cdot) L_i^C(\cdot) L_i^S(\cdot) p(\mathbf{a}_i, \mathbf{b}_i) d\mathbf{b}_i d\mathbf{a}_i$$

Where $L_i^B(\cdot)$, $L_i^C(\cdot)$ and $L_i^S(\cdot)$ corresponds to the likelihood contributions from the binary, continuous and survival parts of the two-part joint model, respectively. With \mathbf{a}_i and \mathbf{b}_i the two vectors of random-effects following a multivariate normal distribution:

$$\begin{bmatrix} \mathbf{a}_i \\ \mathbf{b}_i \end{bmatrix} \sim MVN(\mathbf{0}, \mathbf{B}) \text{ with } \mathbf{B} = \begin{bmatrix} \Sigma_a^2 & \Sigma_{ab} \\ \Sigma_{ab} & \Sigma_b^2 \end{bmatrix}. \quad (1)$$

The set of parameters to estimate is $\Theta = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}, \lambda_0(t), \boldsymbol{\gamma}, \boldsymbol{\varphi})$.

Noting that

$$\text{Prob}(Y_{ij} > 0) = \frac{\exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)}{1 + \exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)}$$

We can deduce

$$\log(\text{Prob}(Y_{ij} > 0)) = \mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i - \log(1 + \exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)).$$

Finally,

$$\log(1 - \text{Prob}(Y_{ij} > 0)) = -\log(1 + \exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)).$$

We introduce $U_{ij} = I[Y_{ij} > 0]$, the likelihood contribution from the binary part can be expressed as

$$\begin{aligned} L_i^B(\cdot) &= \prod_{j=1}^{n_i} P(U_{ij} | \mathbf{a}_i) \\ &= \prod_{j=1}^{n_i} \text{Prob}(Y_{ij} > 0)^{U_{ij}} (1 - \text{Prob}(Y_{ij} > 0))^{(1-U_{ij})} \\ &= \prod_{j=1}^{n_i} \left(\frac{\text{Prob}(Y_{ij} > 0)}{1 - \text{Prob}(Y_{ij} > 0)} \right)^{U_{ij}} (1 - \text{Prob}(Y_{ij} > 0)) \\ &= \prod_{j=1}^{n_i} \exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)^{U_{ij}} \left(1 - \frac{\exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)}{1 + \exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)} \right) \end{aligned}$$

2

The continuous part contribution to the likelihood has a log-normal density

$$L_i^C(\cdot) = \prod_{j=1}^{n_i} \left\{ \frac{1}{Y_{ij} \sqrt{2\pi\sigma_\epsilon^2}} \exp\left(-\frac{(\log(Y_{ij}) - \mu_{ij})^2}{2\sigma_\epsilon^2}\right) \right\}^{U_{ij}}$$

With either the location parameter of the marginal TPJM

$$\begin{aligned} \mu_{ij} &= \mathbf{X}_{Cij}^\top \boldsymbol{\beta} + \mathbf{Z}_{Cij}^\top \mathbf{b}_i - \log(\text{Prob}(Y_{ij} > 0)) - \frac{\sigma_\epsilon^2}{2} \\ &= \mathbf{X}_{Cij}^\top \boldsymbol{\beta} + \mathbf{Z}_{Cij}^\top \mathbf{b}_i + \mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i - \log(1 + \exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)) - \sigma_\epsilon^2/2 \end{aligned}$$

or the location parameter of the conditional TPJM

$$\mu_{ij} = \mathbf{X}_{Cij}^\top \boldsymbol{\beta} + \mathbf{Z}_{Cij}^\top \mathbf{b}_i - \frac{\sigma_\epsilon^2}{2}.$$

The contribution to the likelihood from the survival part corresponds to a Cox proportional hazards model, with splines approximation of the baseline hazard

$$\begin{aligned} L_i^S(\cdot) &= \prod_{j=1}^{n_i} \lambda_i(T_i|a_i, b_i)^{\delta_i} S(T_i|a_i, b_i) \\ &= \prod_{j=1}^{n_i} \lambda_i(T_i|a_i, b_i)^{\delta_i} \exp\left(-\int_0^{T_i} \lambda_i(t|a_i, b_i) dt\right). \end{aligned}$$

Where $\lambda_i(t) = \lambda_0(t) \exp\{\mathbf{X}_{Si}(t)^\top \boldsymbol{\gamma} + h(\cdot)\varphi\}$.

The full likelihood of the M-TPJM is therefore given by

$$\begin{aligned} L_i(\cdot) &= \int_{a_i} \int_{b_i} \prod_{j=1}^{n_i} \left\{ \frac{\exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)}{(\sqrt{2\pi\sigma_\epsilon^2})} Y_{ij}^{-1} \exp\left(-\frac{(\log(Y_{ij}) - \mu_{ij})^2}{2\sigma_\epsilon^2}\right) \right\}^{U_{ij}} \\ &\quad \times \left(1 - \frac{\exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)}{1 + \exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)} \right) \\ &\quad \times \lambda_i(T_i|\Theta)^{\delta_i} \exp\left(-\int_0^{T_i} \lambda_i(t|\Theta) dt\right) p(\mathbf{a}_i, \mathbf{b}_i) db_i da_i \end{aligned}$$

and the log-likelihood

$$\begin{aligned}
\log(L_i(\Theta)) &= \int_{\mathbf{a}_i} \int_{\mathbf{b}_i} \sum_{j=1}^{n_i} \left\{ \mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i - \log(Y_{ij}) - \frac{\log(2\pi)}{2} - \log(\sigma) \right. \\
&\quad - \frac{1}{2\sigma_\epsilon^2} \left(\log(Y_{ij}) + \mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i - \log(1 + \exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)) \right. \\
&\quad \left. \left. + \frac{\sigma_\epsilon^2}{2} - \mathbf{X}_{Cij}^\top \boldsymbol{\beta} + \mathbf{Z}_{Cij}^\top \mathbf{b}_i \right)^2 \right\}^{U_{ij}} - \log(1 + \exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)) \\
&\quad + \delta_i \left(\log(\lambda_0(T_i | \Theta_{ij})) + \mathbf{X}_{Si}(T_i)^\top \boldsymbol{\gamma} + h(\cdot) \varphi \right) \\
&\quad - \left(\int_0^{T_i} \lambda_0(t | \Theta_{ij}) \exp(\mathbf{X}_{Si}(t)^\top \boldsymbol{\gamma} + h(\cdot) \varphi) dt \right) p(\mathbf{a}_i, \mathbf{b}_i) db_i da_i
\end{aligned}$$

Web Appendix B Interpretation of the treatment effect on the marginal mean value of the biomarker with the C-TPJM

The effect of treatment on the marginal mean of the biomarker involves parameters from both the binary $(\alpha_{trt_{int}}, \alpha_{trt_{slo}})$ and the continuous part $(\beta_{trt_{int}}, \beta_{trt_{slo}})$, assuming we include the effect of treatment at baseline and on the slope. Let $\mathbf{X}_{Bij(-trt)}$ and $\boldsymbol{\alpha}_{(-trt)}$ denote the set of covariates and the associated parameters other than treatment in the binary part. We have

$$E[Y_{ij}] = Prob(Y_{ij} > 0) E[Y_{ij} | Y_{ij} > 0] = \frac{\exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)}{1 + \exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)} \exp(\mathbf{X}_{Cij}^\top \boldsymbol{\beta} + \mathbf{Z}_{Cij}^\top \mathbf{b}_i)$$

The effect of a treatment (trt) on the unconditional mean is

$$\begin{aligned}
\frac{E[Y_{ij} | trt_i = 1]}{E[Y_{ij} | trt_i = 0]} &= \frac{Prob(Y_{ij} > 0 | trt_i = 1)}{Prob(Y_{ij} > 0 | trt_i = 0)} \times \frac{E[Y_{ij} | Y_{ij} > 0, trt_i = 1]}{E[Y_{ij} | Y_{ij} > 0, trt_i = 0]} \\
&= \exp \left(\log \left(\frac{Prob(Y_{ij} > 0 | trt_i = 1)}{Prob(Y_{ij} > 0 | trt_i = 0)} \right) + \log \left(\frac{E[Y_{ij} | Y_{ij} > 0, trt_i = 1]}{E[Y_{ij} | Y_{ij} > 0, trt_i = 0]} \right) \right)
\end{aligned}$$

4

where

$$\begin{aligned}
& \log \left(\frac{\text{Prob}(Y_{ij} > 0 | \text{trt}_i = 1)}{\text{Prob}(Y_{ij} > 0 | \text{trt}_i = 0)} \right) = \\
& \log \left(\frac{\frac{\exp\{\alpha_{\text{trt}_{int}} + \text{time}_j \alpha_{\text{trt}_{slo}} + \mathbf{X}_{Bij(-\text{trt})}^\top \boldsymbol{\alpha}_{(-\text{trt})} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i\}}{1 + \exp\{\alpha_{\text{trt}_{int}} + \text{time}_j \alpha_{\text{trt}_{slo}} + \mathbf{X}_{Bij(-\text{trt})}^\top \boldsymbol{\alpha}_{(-\text{trt})} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i\}}}{\frac{\exp\{\mathbf{X}_{Bij(-\text{trt})}^\top \boldsymbol{\alpha}_{(-\text{trt})} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i\}}{1 + \exp\{\mathbf{X}_{Bij(-\text{trt})}^\top \boldsymbol{\alpha}_{(-\text{trt})} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i\}}}} \right) \\
& = \log \left(\frac{\exp\{\alpha_{\text{trt}_{int}} + \text{time}_j \alpha_{\text{trt}_{slo}} + \mathbf{X}_{Bij(-\text{trt})}^\top \boldsymbol{\alpha}_{(-\text{trt})} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i\}}{1 + \exp\{\alpha_{\text{trt}_{int}} + \text{time}_j \alpha_{\text{trt}_{slo}} + \mathbf{X}_{Bij(-\text{trt})}^\top \boldsymbol{\alpha}_{(-\text{trt})} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i\}} \right) \\
& \quad - \log \left(\frac{\exp\{\mathbf{X}_{Bij(-\text{trt})}^\top \boldsymbol{\alpha}_{(-\text{trt})} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i\}}{1 + \exp\{\mathbf{X}_{Bij(-\text{trt})}^\top \boldsymbol{\alpha}_{(-\text{trt})} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i\}} \right) \\
& = \log \left(\exp\{\alpha_{\text{trt}_{int}} + \text{time}_j \alpha_{\text{trt}_{slo}} + \mathbf{X}_{Bij(-\text{trt})}^\top \boldsymbol{\alpha}_{(-\text{trt})} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i\} \right) \\
& \quad - \log \left(1 + \exp\{\alpha_{\text{trt}_{int}} + \text{time}_j \alpha_{\text{trt}_{slo}} + \mathbf{X}_{Bij(-\text{trt})}^\top \boldsymbol{\alpha}_{(-\text{trt})} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i\} \right) \\
& \quad - \left[\log \left(\exp\{\mathbf{X}_{Bij(-\text{trt})}^\top \boldsymbol{\alpha}_{(-\text{trt})} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i\} \right) - \log \left(1 + \exp\{\mathbf{X}_{Bij(-\text{trt})}^\top \boldsymbol{\alpha}_{(-\text{trt})} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i\} \right) \right] \\
& = \alpha_{\text{trt}_{int}} + \text{time}_j \alpha_{\text{trt}_{slo}} + \mathbf{X}_{Bij(-\text{trt})}^\top \boldsymbol{\alpha}_{(-\text{trt})} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i \\
& \quad - \log \left(1 + \exp\{\alpha_{\text{trt}_{int}} + \text{time}_j \alpha_{\text{trt}_{slo}} + \mathbf{X}_{Bij(-\text{trt})}^\top \boldsymbol{\alpha}_{(-\text{trt})} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i\} \right) \\
& \quad - \left(\mathbf{X}_{Bij(-\text{trt})}^\top \boldsymbol{\alpha}_{(-\text{trt})} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i \right) \\
& \quad + \log \left(1 + \exp\{\mathbf{X}_{Bij(-\text{trt})}^\top \boldsymbol{\alpha}_{(-\text{trt})} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i\} \right) \\
& = \alpha_{\text{trt}_{int}} + \text{time}_j \alpha_{\text{trt}_{slo}} + \log \left(\frac{1 + \exp\{\mathbf{X}_{Bij(-\text{trt})}^\top \boldsymbol{\alpha}_{(-\text{trt})} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i\}}{1 + \exp\{\alpha_{\text{trt}_{int}} + \text{time}_j \alpha_{\text{trt}_{slo}} + \mathbf{X}_{Bij(-\text{trt})}^\top \boldsymbol{\alpha}_{(-\text{trt})} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i\}} \right)
\end{aligned}$$

and

$$\log \left(\frac{E[Y_{ij} | Y_{ij} > 0, \text{trt}_i = 1]}{E[Y_{ij} | Y_{ij} > 0, \text{trt}_i = 0]} \right) = \beta_{\text{trt}_{int}} + \text{time}_j \beta_{\text{trt}_{slo}}$$

From this expression, we can deduct that the mean biomarker value is observed higher (lower) at baseline if both $\alpha_{\text{trt}_{int}}$ and $\beta_{\text{trt}_{int}}$ are positive (negative). The mean biomarker value increases (or decreases) over time if both $\alpha_{\text{trt}_{slo}}$ and $\beta_{\text{trt}_{slo}}$ are positive (negative). The interpretation of the marginal treatment effect on mean of the biomarker depends on specific values of the covariates other than treatment in the model ($\mathbf{X}_{Bij(-\text{trt})}$) and on the the random-effects. Further, to obtain a statistical test for this effect or confidence interval, the

Supporting Information for “A marginal two-part joint model for a longitudinal biomarker and a terminal event” 5

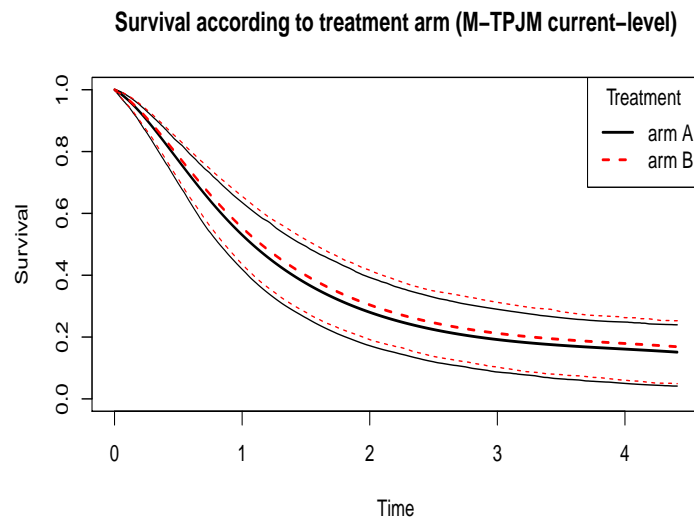
delta method or resampling techniques need to be used. With a marginal two-part model, the effect of treatment in the continuous part at baseline $\beta_{trt_{int}}$ and over time $\beta_{trt_{slo}}$ directly describe the marginal effect of the treatment on the mean biomarker baseline value and evolution over time, respectively. Besides, $\exp(\beta_{trt})$ can be interpreted as the multiplicative effect of the treatment on the unconditional marginal mean of the biomarker but it is not the case when considering the continuous part of the conditional two-part model. The standard errors and confidence intervals are therefore easily obtained as part of the standard model output (Smith et al. (2014)).

[Figure 1 about here.]

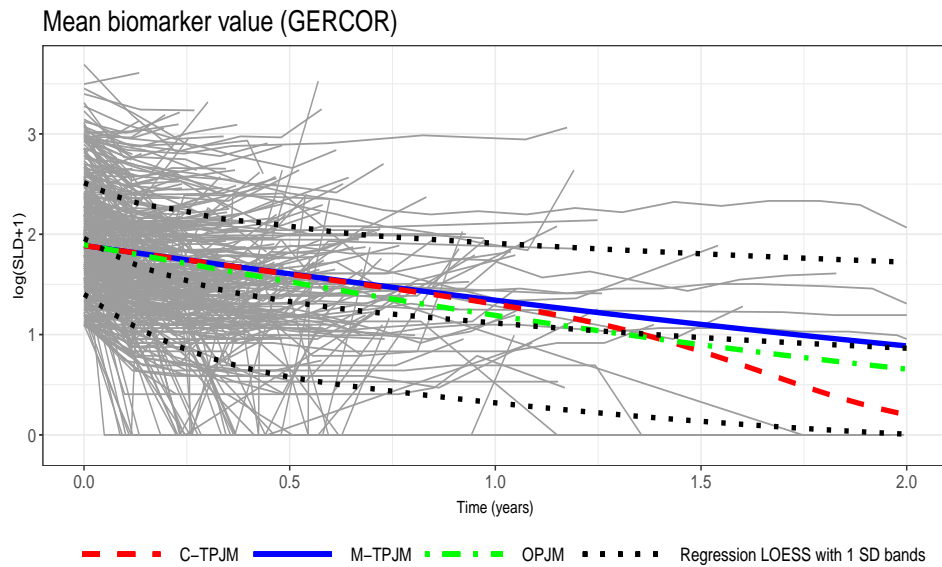
[Figure 2 about here.]

References

Smith, V. A., Preisser, J. S., Neelon, B., and Maciejewski, M. L. (2014). A marginalized two-part model for semicontinuous data. *Statistics in Medicine* **33**, 4891–4903.



Web Figure 1. Survival according to treatment arm from the real data application. The curves take into account the direct effect of treatment on the risk of death (hazard ratio in the survival part) and the time-dependent indirect effect captured in the marginal two-part model for the biomarker (i.e., treatment effect on the mean biomarker value) and shared through the CL association. The confidence intervals are obtained by resampling the parameters using the inverse Hessian matrix of the model, taking the 2.5% and 97.5% quantiles of 1000 simulated curves.



Web Figure 2. Individual biomarker trajectories from the GERCOR data with mean value estimated by the left-censoring OPJM (OPJM), the marginal TPJM (M-TPJM) and the conditional TPJM (C-TPJM). A local regression curve (locally estimated scatterplot smoothing, LOESS) represents the empirical mean biomarker value. Note that the LOESS curve does not take into account the correlation between the repeated measurements within an individual, informative drop-out and the semicontinuous distribution of the biomarker.

4.3 Additional remarks

4.3.1 On the software

The marginal two-part joint model was implemented as an extension of the *longiPenal* function in the R package **frailtypack**, along with the conditional two-part joint model. It is therefore easy to estimate both the conditional and marginal formulations of a model using a simple option in the function to switch from conditional to marginal formulation of the TPJM.

4.3.2 On the association structure

We developed the shared random effects and current value association structures for the M-TPJM, similarly as in Chapter 3 for the C-TPJM. However, the third association structure proposed in Chapter 3 involved a separate association for the current probability of positive value and for the expected value among positives with the risk of event. It was not proposed for the marginal formulation of the two-part joint model because it led to convergence issues. This is likely related to the fact that with the M-TPJM, the binary and continuous parts are not independent conditional on the random effects and this could induce some collinearity issue in the survival model.

Chapter 5

Bayesian Estimation of Two-Part Joint Models for a Longitudinal Semicontinuous Biomarker and a Terminal Event with R-INLA: Interests for Cancer Clinical Trial Evaluation

5.1 Introduction

The objective of this work is to extend the frequentist estimation of the conditional two-part joint model to a Bayesian inference approach. It is motivated by the limitations encountered in the frequentist framework. Indeed, the Levenberg-Marquardt algorithm we proposed for the estimation of the TPJM has strong convergence criteria (i.e., the difference between the log-likelihood, the estimated coefficients and the gradient of the log-likelihood of two consecutive iterations must be under 10^{-3}) and can fail to converge when maximizing the likelihood of complex models, especially for small sample size as the parameters only reflect the data distribution in the frequentist framework. When including covariates in each submodel of the TPJM, the sample must include a sufficient number of patients for each subcategory defined by the covariates in order to have a stable parameter estimation and reach convergence. Moreover, the numerical approximation of the integral over the random effects in the likelihood can represent a large computational burden, especially for high dimension (i.e., multiple correlated random effects), for the frequentist estimation. The computational burden is also an issue within the Bayesian framework with common methods such as MCMC. In this context, the integrated nested Laplace approximation (INLA) method implemented in the R package **R-INLA** is a promising alternative for approximate Bayesian inference, focusing on models that can be expressed as latent Gaussian

Markov random fields (GMRF). It gives fast and accurate estimates of posterior marginals and was recently introduced for joint models. We propose an estimation of the conditional TPJM with **R-INLA** and compare this estimation strategy with the initially proposed estimation with **frailtypack**. This work was motivated by two different phase III randomized clinical trials in colorectal metastatic cancer, the GERCOR study is used to contrast the Bayesian estimation of the conditional TPJM to the frequentist estimation while the PRIME study illustrates the robustness of the Bayesian estimation to convergence issues.

5.2 Article

Bayesian Estimation of Two-Part Joint Models for a Longitudinal Semicontinuous Biomarker and a Terminal Event with R-INLA: Interests for Cancer Clinical Trial Evaluation

Denis Rustand

Biostatistic Team, Bordeaux Population Health Center,
ISPED, Centre INSERM U1219, Bordeaux, France.

Janet van Niekerk

King Abdullah University of Science and Technology,
CEMSE Division, Saudi Arabia.

Håvard Rue

King Abdullah University of Science and Technology,
CEMSE Division, Saudi Arabia.

Christophe Tournigand

Hopital Henri Mondor, Creteil, France

Virginie Rondeau

Biostatistic Team, Bordeaux Population Health Center,
ISPED, Centre INSERM U1219, Bordeaux, France.

Laurent Briollais

Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital,
Dalla Lana School of Public Health (Biostatistics),
University of Toronto, 600 University Ave., Ontario M5G 1X5, Canada

Abstract

Two-part joint models for a longitudinal semicontinuous biomarker and a terminal event have been recently introduced based on frequentist estimation. The biomarker distribution is decomposed into a probability of positive value and the expected value among positive values. Shared random effects can represent the association structure between the biomarker and the terminal event. The computational burden increases compared to standard joint models with a single regression model for the biomarker. In this context, the frequentist estimation implemented in the R package **frailtypack** can be challenging for complex models (i.e., large number of parameters and dimension of the random effects). As an alternative, we propose a Bayesian estimation of two-part joint models based on the Integrated Nested Laplace Approximation (INLA) algorithm to alleviate the computational burden and fit more complex models. Our simulation studies show that **R-INLA** reduces the computation time substantially as well as the variability of the parameter estimates and improves the model convergence compared to **frailtypack**. We contrast the Bayesian and frequentist approaches in the analysis of two randomized cancer clinical trials (GERCOR and PRIME studies), where **R-INLA** suggests a stronger association between the biomarker and the risk of event. Moreover, the Bayesian approach was able to characterize subgroups of patients associated with different responses to treatment in the PRIME study while **frailtypack** had convergence issues. Our study suggests that the Bayesian approach using **R-INLA** algorithm enables broader applications of the two-part joint model to clinical applications.

1 Introduction

Estimation of joint models for longitudinal and time-to-event data were initially introduced using maximum likelihood estimation (Wulfsohn and Tsiatis (1997); Henderson et al. (2000); Song et al. (2002); Chi and Ibrahim (2006)). It was further developed within the Bayesian framework in situations where maximum likelihood estimation with asymptotic assumptions faces nonidentifiability issues. It allows flexible and more complex association structures and can handle multiple longitudinal outcomes (Andrinopoulou and Rizopoulos (2016)). Bayesian joint models can be fitted with the R package **JMbayes** (Rizopoulos et al. (2016)), which has been used in many biomedical researches (Lawrence Gould et al. (2015)), among other packages (e.g. **rstanarm**, Muth et al. (2018)). Bayesian estimation is usually based on MCMC techniques (Hanson et al. (2011); R. Brown and G. Ibrahim (2003); Xu and Zeger (2001); Rizopoulos and Ghosh (2011)), which can have slow convergence properties. The Integrated Nested Laplace Approximation (INLA) algorithm has been recently introduced as an alternative to MCMC techniques for latent Gaussian models (LGM) (Rue et al. (2009); Martins et al. (2013)). Many statistical models for spatial statistics, time series, etc., can be formulated as LGMs. A key feature of INLA is to provide approximations of the posterior marginals needed for Bayesian inference very efficiently and that still remain very accurate compared to MCMC methods (Rue et al. (2017)). By formulating complex joint models as LGMs, **R-INLA** can be used to fit these models as developed recently (Van Niekerk, Bakka, and Rue (2019); Van Niekerk, Bakka, Rue, and Schenk (2019)). For the two-part joint model, **R-INLA** is yet to be used for inference.

Two-part joint models (TPJMs) for a longitudinal semicontinuous biomarker and a terminal event have been recently introduced (Rustand, Briollais, Tournigand, and Rondeau (2020)) and applied to the joint analysis of survival times and repeated measurements of the Sum of the Longest Diameter of target lesions (SLD), which is a biomarker representative of tumor burden in cancer clinical trials. The TPJM uses a conditional two-part joint model that decomposes the biomarker distribution into a binary outcome (zero vs. positive value) fitted with a logistic mixed effects model and a continuous outcome (positive values only) fitted with either a linear mixed effect model on the log-transformed outcome or a lognormal mixed effects model (Rustand, Briollais, and Rondeau (2020)). The “conditional” form of the two-part model gives the effect of covariates on the mean biomarker value conditional on a positive value in the continuous part. An alternative marginal model has recently been proposed to get the effect of covariates on the (unconditional) mean of the biomarker. A drawback of the marginal two-part model is that it may lead to arbitrary heterogeneity and provide less interpretable estimates on the conditional mean of the biomarker among positive values (Smith et al. (2014)). In this article, we focus on the conditional two-part joint model, simply referred to as TPJM in what follows. The association with the survival model can be specified in terms of shared random effects, i.e., random effects that are shared between the different components of the models including the binary and continuous parts of the model and the survival component. An important limitation of such models is the estimation procedure that requires a numerical approximation of the random effects distribution, which can lead to long computation times and convergence issues with high-dimensional parameter settings and complex association structures between the different components of the TPJMs. In this article, we propose an efficient Bayesian estimation procedure for the TPJM which relies on INLA algorithm, as implemented in the R package **R-INLA**. The Bayesian inference is compared to the frequentist estimation of the TPJM available in the R package **frailtypack**. The remainder of the article is structured as follows: in Section 2, we describe the TPJM and introduce the frequentist and Bayesian estimations. In Section 3, we present a simulation study to assess the performance of these two estimation strategies in terms of bias, coverage probability, computation time and convergence rate. An application to two randomized clinical trials each comparing two treatment strategies in patients with metastatic colorectal cancer is proposed in Section 4 and we conclude with a discussion in Section 5.

2 Estimation of the conditional two-part joint model

2.1 Model specification

Let Y_{ij} denote the biomarker measurement for individual i ($i = 1, \dots, n$) at visit j ($j = 1, \dots, n_i$), T_i denotes the survival time and δ_i the censoring indicator for individual i . We use a logistic mixed effect model for

the probability of a positive value of the biomarker and a lognormal mixed effect model for the conditional expected biomarker value. A proportional hazards survival model specifies the effect of covariates on survival time, adjusted for the individual heterogeneity captured in the biomarker model. The complete model is defined as follows:

$$\begin{cases} \eta_{Bij} = \text{Logit}(\text{Prob}(Y_{ij} > 0)) = \mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i & \text{(Binary part),} \\ \eta_{Cij} = E[Y_{ij} | Y_{ij} > 0] = \exp(\mathbf{X}_{Cij}^\top \boldsymbol{\beta} + \mathbf{Z}_{Cij}^\top \mathbf{b}_i) & \text{(Continuous part),} \\ \lambda_i(t) = \lambda_0(t) \exp(\eta_{Si}) = \lambda_0(t) \exp(\mathbf{X}_i^\top \boldsymbol{\gamma} + \mathbf{a}_i^\top \boldsymbol{\varphi}_a + \mathbf{b}_i^\top \boldsymbol{\varphi}_b) & \text{(Survival part),} \end{cases}$$

where \mathbf{X}_{Bij} , \mathbf{X}_{Cij} and \mathbf{X}_i are vectors of covariates associated to the fixed effects $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively. Similarly, \mathbf{Z}_{Bij} and \mathbf{Z}_{Cij} are vectors of covariates associated to the random effects \mathbf{a}_i and \mathbf{b}_i in the binary and continuous parts. These two vectors of random effects follow a multivariate normal distribution. They are shared in the survival model, with association parameters $\boldsymbol{\varphi}_a$ and $\boldsymbol{\varphi}_b$, respectively. Therefore, the random effects account for both the association between the three components of the model and the correlation between the repeated measurements in the longitudinal process (observations are independent conditional on the random effects). The joint distribution assumes that the vectors of random effects underlies both the longitudinal and survival process, the joint distribution of the observed outcomes for individual i is defined by

$$\begin{aligned} p(T_i, \delta_i, \mathbf{Y}_i | \mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\Theta}) &= p(T_i, \delta_i | \mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\Theta}) \prod_{j=1}^{n_i} p(Y_{ij} | \mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\Theta}) \\ &= p(T_i, \delta_i | \mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\Theta}) \prod_{j=1}^{n_i} p(Y_{ij} | Y_{ij} > 0; \mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\Theta}) p(Y_{ij} > 0; \mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\Theta}) \end{aligned}$$

with $\boldsymbol{\Theta}$ the full parameter vector, including the parameters for the binary, continuous and survival outcomes, the baseline hazard function and the random effects covariance matrix, such that the full conditional distribution is given by

$$p(\mathbf{T}, \boldsymbol{\delta}, \mathbf{Y} | \mathbf{a}, \mathbf{b}; \boldsymbol{\Theta}) = \prod_{i=1}^n p(T_i, \delta_i, \mathbf{Y}_i | \mathbf{a}_i, \mathbf{b}_i; \boldsymbol{\Theta}).$$

The likelihood contribution for the i th subject can be formulated as follows

$$\begin{aligned} L_i(\boldsymbol{\Theta} | \mathbf{Y}_i, T_i, \delta_i) &= \int_{\mathbf{a}_i} \int_{\mathbf{b}_i} \prod_{j=1}^{n_i} \exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)^{U_{ij}} \left(1 - \frac{\exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)}{1 + \exp(\mathbf{X}_{Bij}^\top \boldsymbol{\alpha} + \mathbf{Z}_{Bij}^\top \mathbf{a}_i)} \right) \\ &\quad \times \left\{ \frac{1}{Y_{ij} \sqrt{2\pi\sigma_\epsilon^2}} \exp\left(-\frac{(\log(Y_{ij}) - \mu_{ij})^2}{2\sigma_\epsilon^2}\right) \right\}^{U_{ij}} \\ &\quad \times \lambda_i(T_i | \mathbf{a}_i, \mathbf{b}_i)^{\delta_i} \exp\left(-\int_0^{T_i} \lambda_i(t | \mathbf{a}_i, \mathbf{b}_i) dt\right) p(\mathbf{a}_i, \mathbf{b}_i) d\mathbf{b}_i d\mathbf{a}_i, \end{aligned}$$

where $U_{ij} = I[Y_{ij} > 0]$ and $\lambda_i(t) = \lambda_0(t) \exp\{\mathbf{X}_{Si}(t)^\top \boldsymbol{\gamma} + \mathbf{a}_i^\top \boldsymbol{\varphi}_a + \mathbf{b}_i^\top \boldsymbol{\varphi}_b\}$.

2.2 Bayesian estimation of the TPJM

Define $\mathbf{D} \equiv \{T_i, \delta_i, Y_{ij} : i = 1, \dots, n; j = 1, \dots, n_i\}$ the observation variables. The goal of the Bayesian inference is to estimate the posterior distribution $\pi(\boldsymbol{\Theta} | \mathbf{D})$. The joint posterior distribution $\pi(\boldsymbol{\Theta} | \mathbf{D})$ is given by Bayes theorem as

$$\pi(\boldsymbol{\Theta} | \mathbf{D}) = \frac{p(\mathbf{D} | \boldsymbol{\Theta}) \pi(\boldsymbol{\Theta})}{\pi(\mathbf{D})} \propto p(\mathbf{D} | \boldsymbol{\Theta}) \pi(\boldsymbol{\Theta}),$$

where $p(\mathbf{D} | \boldsymbol{\Theta})$ is the likelihood and $\pi(\boldsymbol{\Theta})$ is the joint prior. The marginal likelihood $\pi(\mathbf{D}) = \int_{\boldsymbol{\Theta}} p(\mathbf{D} | \boldsymbol{\Theta}) \pi(\boldsymbol{\Theta}) d\boldsymbol{\Theta}$ acts as a normalizing constant. The posterior marginal distribution of each parameter is then obtained by integrating out the other parameters of the model. In many cases, the posterior distribution is not analytically

tractable and sampling-based methods like MCMC can be used. Approximate methods like INLA, provide exact approximations to the posterior at lower cost than sampling-based methods. The INLA methodology is based on the assumption that the statistical model is a latent Gaussian model, which we show in the next section for the TPJM.

2.3 Formulation of the TPJM as a latent Gaussian model

Let $\mathbf{u} \equiv (\boldsymbol{\eta}_B, \boldsymbol{\eta}_C, \boldsymbol{\eta}_S, \mathbf{a}, \mathbf{b}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\varphi})$ be the set of latent Gaussian variables related to the TPJM, where $\boldsymbol{\lambda}$ is a vector of coefficients associated with a random walk order one or order two used to approximate the baseline hazard function $\lambda_0(t)$ of the survival model. Note that the first $\sum_{i=1}^n n_i + \sum_{i=1}^n n_i + n$ elements of \mathbf{u} are the linear predictors of the TPJM and the rest of the elements are the latent unobserved variables. For that reason, the random field \mathbf{u} is termed the latent field.

In particular, we assume $\mathbf{a}_i, \mathbf{b}_i | \mathbf{Q}_{ab} \sim \mathcal{N}(0, \mathbf{Q}_{ab}^{-1})$, $\boldsymbol{\alpha} \sim \mathcal{N}(0, \tau_\alpha \mathbf{I})$, $\boldsymbol{\beta} \sim \mathcal{N}(0, \tau_\beta \mathbf{I})$, $\boldsymbol{\gamma} \sim \mathcal{N}(0, \tau_\gamma \mathbf{I})$ and $\boldsymbol{\varphi} \sim \mathcal{N}(0, \tau_\varphi \mathbf{I})$. The coefficients of the baseline hazard $\boldsymbol{\lambda}$ are assumed to follow either a random walk one or random walk two model. These models are stochastic spline models with precision parameter τ_λ . Thus, the latent field \mathbf{u} is multivariate Gaussian with zero mean and precision matrix $\mathbf{Q}(\boldsymbol{\theta}_1)$, i.e.,

$$\mathbf{u} | \boldsymbol{\theta}_1 \sim \mathcal{N}(0, \mathbf{Q}^{-1}(\boldsymbol{\theta}_1)).$$

Note that $\mathbf{Q}(\boldsymbol{\theta}_1)$ is a sparse matrix indexed by a low dimension of parameters $\boldsymbol{\theta}_1$. This then implies that the latent field \mathbf{u} is a Gaussian Markov random field (GMRF).

The distribution of the observation variables \mathbf{D} is denoted by $\pi(\mathbf{D} | \mathbf{u}, \boldsymbol{\theta}_2)$ and depends on the set of hyperparameters $\boldsymbol{\theta}_2$ that influence the likelihood. They are assumed to be conditionally independent over the n individuals given the latent Gaussian random field \mathbf{u} and hyperparameters $\boldsymbol{\theta} \equiv (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$,

$$\mathbf{D} | \mathbf{u}, \boldsymbol{\theta} \sim \prod_{i=1}^n p(\mathbf{d}_i | u_i, \boldsymbol{\theta}).$$

Thus, assuming a prior $\pi(\boldsymbol{\theta})$ for the hyperparameters $\boldsymbol{\theta} \equiv (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, the posterior of $(\mathbf{u}, \boldsymbol{\theta})$ can be written as

$$\begin{aligned} \pi(\mathbf{u}, \boldsymbol{\theta} | \mathbf{D}) &\propto \pi(\boldsymbol{\theta}) \pi(\mathbf{u} | \boldsymbol{\theta}) \prod_{i=1}^n p(\mathbf{d}_i | u_i, \boldsymbol{\theta}), \\ &\propto \pi(\boldsymbol{\theta}) |\mathbf{Q}(\boldsymbol{\theta}_1)|^{n/2} \exp \left[\frac{1}{2} \mathbf{u}^T \mathbf{Q}(\boldsymbol{\theta}_1) \mathbf{u} + \sum_{i=1}^n \log \{ p(\mathbf{d}_i | u_i, \boldsymbol{\theta}) \} \right]. \end{aligned}$$

This construction then shows that the TPJM is in fact an LGM since the latent field is a Gaussian Markov field and each data contribution depends on only one element of the latent field.

The main aim of INLA is then to approximate the posterior marginals $\pi(u_i | \mathbf{D})$, $\pi(\boldsymbol{\theta} | \mathbf{D})$ and $p(\theta_j | \mathbf{D})$.

2.4 INLA

The INLA methodology introduced by Rue and Held (2005) is a major contribution to achieving efficient Bayesian inference, especially for complex or large models. INLA uses a unique combination of Laplace Approximations and conditional distributions to approximate the joint posterior density as well as the marginals of the latent field and hyperparameters. It is thus not a sampling based method like MCMC and such.

For the sake of brevity, the INLA methodology can be presented in the following three steps:

1. Approximate

$$\pi(\boldsymbol{\theta} | \mathbf{D}) = \frac{\pi(\mathbf{u}, \boldsymbol{\theta} | \mathbf{D})}{\pi(\mathbf{u} | \boldsymbol{\theta}, \mathbf{D})} \approx \frac{\pi(\boldsymbol{\theta}) \pi(\mathbf{u} | \boldsymbol{\theta}) \pi(\mathbf{D} | \mathbf{u}, \boldsymbol{\theta})}{\tilde{\pi}(\mathbf{u} | \boldsymbol{\theta}, \mathbf{D})} \Big|_{\mathbf{u}=\mathbf{u}^*(\boldsymbol{\theta})},$$

where the Gaussian or Laplace approximation is used to approximate the denominator at the mode $\mathbf{u}^*(\boldsymbol{\theta})$ of the latent field for a given configuration of $\boldsymbol{\theta}$.

2. Approximate

$$\pi(u_j|\boldsymbol{\theta}, \mathbf{D}) \propto \frac{\pi(\mathbf{u}, \boldsymbol{\theta}|\mathbf{D})}{\pi(\mathbf{u}_{-j}|u_j, \boldsymbol{\theta}, \mathbf{D})},$$

using a Gaussian approximation (option 1), or in a similar way as mentioned in step 1 (option 2) or by expanding the numerator and denominator up to a third order Taylor series expansion and then applying a Laplace approximation (option 3).

3. Use numerical integration to approximate

$$\pi(u_j|\mathbf{Y}) \approx \sum_{h=1}^H \tilde{\pi}(u_j|\boldsymbol{\theta}_h^*, \mathbf{Y}) \tilde{\pi}(\boldsymbol{\theta}_h^*|\mathbf{Y}) \Delta_h,$$

from steps 1 and 2. The integration points $\{\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_H^*\}$ are selected from a rotation using polar coordinates and based on the density at these points.

2.5 Priors for the hyperparameters

From the formulation of the TPJM as an LGM, the prior for the hyperparameters, $\pi(\boldsymbol{\theta})$, should be specified. This prior can assume any form while keeping the TPJM an LGM. Amidst the debate about priors, Simpson et al. (2017) proposed a framework to construct principled priors for hyperparameters, namely penalizing complexity (PC) priors. These priors are derived based on the distance from a complex model to a simpler (base) model, with a user-defined parameter that informs the strength of contraction towards the simpler model. This parameter defines whether the PC priors are vague, weakly informative, or strongly informative based on the departure from the base model measured by the Kullback-Leibler distance. It is based on the principle of parsimony, simplifying the interpretation of the results by ensuring that the priors do not overfit.

In our case we have various precision hyperparameters, $\{\mathbf{Q}_{ab}, \tau_\alpha, \tau_\beta, \tau_\gamma, \tau_\varphi, \tau_\lambda\}$. We assign weakly informative priors to the fixed effects such that $\tau_\alpha = \tau_\beta = \tau_\gamma = \tau_\varphi = 10^{-3}$. We thus need to formulate priors for the elements of \mathbf{Q}_{ab} and τ_λ . For all these hyperparameters (precision and correlation parameters), we assume the respective PC priors as given in Simpson et al. (2017).

As illustration we give the details for the precision of the first element of \mathbf{a} , τ_{a_0} . The PC prior is derived as

$$\pi(\tau_{a_0}) = \frac{\rho}{2} \tau_{a_0}^{-3/2} \exp(-\rho \tau_{a_0}^{-1/2}),$$

with the user-defined scaling parameter $\rho = -\frac{\ln(v)}{w}$. This parameter is chosen based on the desired tail behaviour (or strength of contraction towards the base model $\sigma_{a_0} = \tau_{a_0}^{-1/2} = 0$) in the sense that v and w are such that

$$P[\sigma_{a_0} > w] = v, \quad w > 0, 0 < v < 1.$$

Larger values of v and w results in higher prior density away from the base model, whereas smaller values of v places more density closer to the base model.

3 Simulation study

3.1 Settings

We designed simulation studies to compare the performances of **R-INLA** and **frailtypack** in terms of bias of the parameter estimates, coverage probabilities, computation time and convergence rates. The main factor driving the performance is the model complexity defined by the number of parameters. In particular, the number of correlated random effects defines the dimension of the integration that needs to be numerically approximated. We propose two simulation scenarios based on the results obtained from the real data analyses. The first scenario includes a random intercept in the binary and continuous parts of the TPJM that are correlated. The second simulation scenario includes an additional random-effect for the individual deviation from the mean slope in the continuous part, thus 3 correlated random effects. For each scenario, we generate

1000 datasets with 200 individuals each, corresponding to a small sample size commonly seen in randomized clinical trials. We first sample the positive longitudinal biomarker repeated measurements from a log-normal distribution and include the zero values sampled from a binomial distribution. The relationship between the probability of zero value and the positive values is given by the correlated random effects. Survival times for the terminal event are generated with an exponential baseline hazard function with a scale of 0.2, an administrative censoring is assumed to occur at the end of the follow-up (4 years). The rate of zeros is 8% (SD=1%), which is in between what we observed in our two real datasets (12% of zeros in application 1 and 4% in application 2). A zero value observation corresponds to a patient who experienced a complete disappearance of his/her target lesions and thus is extremely informative about treatment effect.

The baseline hazard function in the survival part of the model is approximated by a random walk model with **R-INLA** (Martino et al. (2011)) such that for m bins of the time axis,

$$\lambda_k - \lambda_{k-1} \sim N(0, \tau_\lambda),$$

where the PC prior (see Section 2.5) is used as the prior for τ_λ .

The random walk order one model is a stochastic smoothing spline that smooths based on first order differences. The number of bins are not influential (as opposed to knots of other splines) since an increase in bins only results in an estimate closer to the stochastic model. In the simulations and applications, we use the random walk order two model that provides a smoother spline since the smoothing is then done on the second order. See Van Niekerk et al. (2020) for more details on the use of these random walk models as Bayesian smoothing splines. This approximation is different with **frailtypack** that uses cubic M-splines with 5 knots. A penalization ensures that the baseline hazard is smooth (a smoothing parameter is chosen using an approximate cross-validation criterion from a separate Cox model). The Levenberg-Marquardt algorithm, a robust Newton-like algorithm maximizes the log-likelihood function with **frailtypack** (Marquardt (1963)). The convergence of the algorithm depends on three conditions: The difference between the log-likelihood, the estimated coefficients and the gradient of the log-likelihood of two consecutive iterations must be under 10^{-3} . We use a Monte-Carlo approximation for the approximation of the integrals over the random effects in the likelihood function, with 1000 integration points which is a reasonable tradeoff between the precision of the approximation and computation time. The simulation studies are performed with 80 CPUs, **frailtypack** uses Message Passing Interface (MPI) for parallel computation while the conjunction of **R-INLA** with the **PARDISO** library provides a high performance computing environment with parallel computing support (Schenk and Gärtner (2004)). In practice, the 80 CPUs are only useful to reduce the computation time with **frailtypack** because the computation time with INLA is very low regardless of the number of threads because of the small sample size and number of hyperparameters.

3.2 Results

In the results, we are comparing a Bayesian and frequentist method and for this we have to keep in mind that each has a different criteria for evaluation of the method. Frequentist bias is used to evaluate the results from **frailtypack** while the plausibility of the result based on 95% credible intervals are used to evaluate the results from **R-INLA** (Hespanhol et al. (2019)). However, we are interested in the Bayesian approximation of the MLE (i.e. non informative priors) and therefore provide an interpretation in this context.

3.2.1 Scenario 1: Two correlated random effects

The fixed effect parameters from the binary and continuous parts are properly estimated with both algorithms, with similar precision and coverage probabilities close to the expected 95% level. The parameter for the treatment effect in the survival part ($\gamma_1 = 0.2$) is associated to a larger variability with **frailtypack** ($\hat{\gamma}_1 = 0.22$, SD=0.35, CP=96%) compared to **R-INLA** ($\hat{\gamma}_1 = 0.19$, SD=0.27, CP=96%). The true value of the standard deviation of the random intercept in the binary part ($\sigma_a = 1$) is within the 95% credible interval with **R-INLA** ($\hat{\sigma}_a = 0.86$, SD=0.19), with a slightly lower posterior mean value compared to **frailtypack**'s estimate ($\hat{\sigma}_a = 0.97$, SD=0.22). The random intercept's standard deviation in the continuous part is found similar with both algorithms but the correlation between the random intercepts of the binary and continuous parts ($corr_{ab} = 0.5$) has a reduced variability estimate with **R-INLA** ($c\hat{orr}_{ab} = 0.48$, SD=0.10) compared to **frailtypack** ($c\hat{orr}_{ab} = 0.51$, SD=0.15). The main difference observed is the estimation of the parameters

for the association of the random effects with the risk of event, which links the biomarker to the terminal event. The association involving the random intercept from the binary part ($\varphi_a = 1$) has much lower variability with **R-INLA** ($\hat{\varphi}_a = 1.00$, SD=0.11, CP=99%) and is unbiased with good coverage with **frailtypack** ($\hat{\varphi}_a = 1.08$, SD=0.82, CP=93%). The association involving the random intercept from the continuous ($\varphi_b = 1$) part is biased upwards with **frailtypack** with large variability ($\hat{\varphi}_b = 1.33$, SD=1.13, CP=92%), while **R-INLA**'s posterior estimate recovers the true value ($\hat{\varphi}_b = 1.05$, SD=0.15, CP=99%). This could be due to the small sample size problems that cause more convergence issues under the frequentist framework. Although **R-INLA** yields accurate posterior estimates with small variability for these parameters, the coverage probabilities are higher than the expected 95%. The computation times are much lower with **R-INLA** (14 seconds per model, SD=1) compared to **frailtypack** (66 seconds per model, SD=26). Finally, all models converged with **R-INLA** while 11% of the 1000 models did not reach convergence with **frailtypack**.

Table 1: Simulations with two correlated random effects

Package	R-INLA		frailtypack	
	Est.*	(SD [†]) [CP [‡]]	Est. (SD)	[CP]
Binary part (SLD>0 versus SLD=0)				
intercept	$\alpha_0 = 4$	3.96 (0.35) [94%]	4.02 (0.38)	[95%]
time (year)	$\alpha_1 = -0.5$	-0.51 (0.11) [95%]	-0.51 (0.12)	[95%]
treatment (B/A)	$\alpha_2 = -0.5$	-0.49 (0.45) [96%]	-0.50 (0.47)	[95%]
time:treatment (B/A)	$\alpha_3 = 0.5$	0.50 (0.18) [94%]	0.51 (0.18)	[95%]
Continuous part ($E[Y_{ij} Y_{ij} > 0]$)				
intercept	$\beta_0 = 2$	2.00 (0.05) [95%]	2.00 (0.06)	[92%]
time (years)	$\beta_1 = -0.3$	-0.30 (0.01) [95%]	-0.30 (0.01)	[95%]
treatment (B/A)	$\beta_2 = -0.3$	-0.30 (0.08) [94%]	-0.30 (0.09)	[91%]
time:treatment (B/A)	$\beta_3 = 0.3$	0.30 (0.02) [94%]	0.30 (0.02)	[95%]
residual S.E.	$\sigma_\varepsilon = 0.3$	0.30 (0.01) [89%]	0.30 (0.01)	[94%]
Death risk				
treatment (B/A)	$\gamma_1 = 0.2$	0.19 (0.27) [96%]	0.22 (0.35)	[96%]
Association				
Intercept (binary part)	$\varphi_a = 1$	1.00 (0.11) [99%]	1.08 (0.82)	[93%]
Intercept (continuous part)	$\varphi_b = 1$	1.05 (0.15) [99%]	1.33 (1.13)	[92%]
Random effects's standard deviation				
intercept (binary part)	$\sigma_a = 1$	0.86 (0.19)	0.97 (0.22)	
intercept (continuous part)	$\sigma_b = 0.5$	0.50 (0.03)	0.50 (0.03)	
	$corr_{ab} = 0.5$	0.48 (0.10)	0.51 (0.15)	
Computation time				
80 CPUs - Intel Xeon E5-4627 v4 2.60 GHz		14 sec. (1)	66 sec. (26)	
Convergence rate		100%	89%	

* Posterior mean, [†] Standard deviation of the posterior mean, [‡] Coverage probability

3.2.2 Scenario 2: Three correlated random effects

With an additional random effect parameter compared to scenario 1, the fixed effects parameters are still properly estimated with **R-INLA**. The coverage probabilities are low with **frailtypack** for the slope and treatment by slope parameters in the continuous part ($\beta_1 = -0.3$ and $\beta_3 = 0.3$), while the parameter estimates remain unbiased. The variability for these two parameters is lower with **R-INLA** ($\hat{\beta}_1 = -0.30$, SD=0.06, CP=94% and $\hat{\beta}_3 = 0.30$, SD=0.08, CP=95%) compared to **frailtypack** ($\hat{\beta}_1 = -0.25$, SD=0.11, CP=46% and $\hat{\beta}_3 = 0.29$, SD=0.14, CP=44%). As observed in the first scenario, the treatment effect's posterior estimate in the survival model has lower variability with **R-INLA** ($\hat{\gamma}_1 = 0.20$, SD=0.30, CP=95%), moreover the coverage probability for this parameter is lower than observed with **frailtypack** ($\hat{\gamma}_1 = 0.24$, SD=0.49, CP=84%). For the random effects covariance structure estimation, the posterior mean from **R-INLA** is slightly lower than the true value of the random intercept's standard deviation in the binary part ($\hat{\sigma}_a = 0.86$, SD=0.15) with lower variability for the standard deviation and correlation terms overall. The association parameters ($\varphi_a = 1$, $\varphi_{b_0} = 1$, $\varphi_{b_1} = 1$) are recovered well and have much lower variability with **R-INLA** ($\hat{\varphi}_a = 1.03$, SD=0.13, CP=98%, $\hat{\varphi}_{b_0} = 1.07$, SD=0.14, CP=98%, $\hat{\varphi}_{b_1} = 1.07$, SD=0.14, CP=98%) compared to **frailtypack** ($\hat{\varphi}_a = 0.87$, SD=1.86, CP=91%, $\hat{\varphi}_{b_0} = 1.03$, SD=1.82, CP=89%, $\hat{\varphi}_{b_1} = 1.44$, SD=1.70, CP=91%), but still with conservative coverage probabilities. Computation times remain much lower with **R-INLA** (19 seconds per model, SD=2) compared to **frailtypack** (159 seconds per model, SD=52) for which the time increased substantially when adding the third random-effect. Moreover, the convergence rate of the model is reduced with **frailtypack** for this scenario (82%), because the model complexity increased while it remains 100% with **R-INLA**.

Table 2: Simulations with three correlated random effects

Package	R-INLA		frailtypack	
	Est. *	(SD) [†] [CP] [‡]	Est. (SD)	[CP]
Binary part (SLD>0 versus SLD=0)				
intercept	$\alpha_0 = 4$	3.95 (0.35) [94%]	4.03 (0.39)	[94%]
time (year)	$\alpha_1 = -0.5$	-0.52 (0.12) [94%]	-0.51 (0.12)	[95%]
treatment (B/A)	$\alpha_2 = -0.5$	-0.51 (0.47) [95%]	-0.50 (0.50)	[94%]
time:treatment (B/A)	$\alpha_3 = 0.5$	0.51 (0.18) [95%]	0.50 (0.18)	[96%]
Continuous part ($E[Y_{ij} Y_{ij} > 0]$)				
intercept	$\beta_0 = 2$	2.00 (0.05) [96%]	1.99 (0.06)	[88%]
time (years)	$\beta_1 = -0.3$	-0.30 (0.06) [94%]	-0.25 (0.11)	[46%]
treatment (B/A)	$\beta_2 = -0.3$	-0.30 (0.08) [96%]	-0.29 (0.09)	[87%]
time:treatment (B/A)	$\beta_3 = 0.3$	0.30 (0.08) [95%]	0.29 (0.14)	[44%]
residual S.E.	$\sigma_\varepsilon = 0.3$	0.29 (0.01) [88%]	0.30 (0.01)	[96%]
Death risk				
treatment (B/A)	$\gamma_1 = 0.2$	0.20 (0.30) [95%]	0.24 (0.49)	[84%]
Association				
Intercept (binary part)	$\varphi_a = 1$	1.03 (0.13) [98%]	0.87 (1.86)	[91%]
Intercept (continuous part)	$\varphi_{b_0} = 1$	1.07 (0.14) [98%]	1.03 (1.82)	[89%]
Slope (continuous part)	$\varphi_{b_1} = 1$	1.07 (0.14) [98%]	1.44 (1.70)	[91%]
Random effects's standard deviation				
intercept (binary part)	$\sigma_a = 1$	0.86 (0.15)	1.07 (0.24)	
intercept (continuous part)	$\sigma_{b_0} = 0.5$	0.50 (0.03)	0.50 (0.04)	
slope (continuous part)	$\sigma_{b_1} = 0.5$	0.49 (0.03)	0.58 (0.10)	
	$corr_{ab_0} = 0.5$	0.47 (0.10)	0.48 (0.16)	
	$corr_{ab_1} = 0.5$	0.46 (0.12)	0.58 (0.16)	
	$corr_{b_0b_1} = -0.2$	-0.19 (0.09)	-0.14 (0.19)	
Computation time				
80 CPUs - Intel Xeon E5-4627 v4 2.60 GHz)		19 sec. (2)	159 sec. (52)	
Convergence rate				
		100%	82%	

* Posterior mean, [†] Standard deviation of the posterior mean, [‡] Coverage probability

3.3 Conclusions

Our method comparison suggests that the frequentist approach, implemented in **frailtypack**, reaches some limitations when fitting the more complex TPJMs, compared to the Bayesian approach implemented in **R-INLA**. Convergence rates are lower and estimation of the association parameters is highly variable with **frailtypack**. However, a representation of the baseline survival curves estimated under both scenarios is displayed in Figure 1. The median of the estimated survival curves is slightly lower than the true survival with **R-INLA** although the credible interval contains the true curve, while the point estimate from **frailtypack** is closer to the true curve but yields much larger confidence intervals.

4 Application

We applied the Bayesian TPJM to two cancer clinical trials, the GERCOR and the PRIME studies. A comparison with **frailtypack** is provided only for the GERCOR data since this approach did not converge on the PRIME study. We used the same parameterizations for **R-INLA** and **frailtypack**, as detailed in the simulation studies. In the context of a Bayesian approximation of the MLE, we provide indications of the p-value for both **frailtypack** and **INLA** to ease the interpretation and the comparison of the results.

Table 3: Description of the GERCOR and PRIME study datasets

Study	GERCOR		PRIME	
	arm A	arm B	arm A	arm B
Treatment	FOLFIRI/FOLFOX6	FOLFOX6/FOLFIRI	FOLFOX4	Panitumumab/FOLFOX4
Number of patients enrolled	109	111	593	590
Number of patients for the analysis	101	104	223	219
number of repeated measurements of the SLD	748	727	1192	1081
Number of zero values (%)	118 (16.2%)	56 (7.5%)	47 (3.8%)	52 (4.6%)
Number of death (%)	83 (82.2%)	82 (78.8%)	164 (73.5%)	164 (74.9%)
Median OS (years)	1.8 (1.4-2.3)	1.8 (1.5-2.2)	1.7 (1.5-1.9)	1.4 (1.3-1.7)
KRAS exon 2 at codons 12 and 13				
Nonmutated			132 (59.2%)	128 (58.4%)
Mutated			91 (40.8%)	91 (41.6%)
Not available	101 (100%)	104 (100%)		

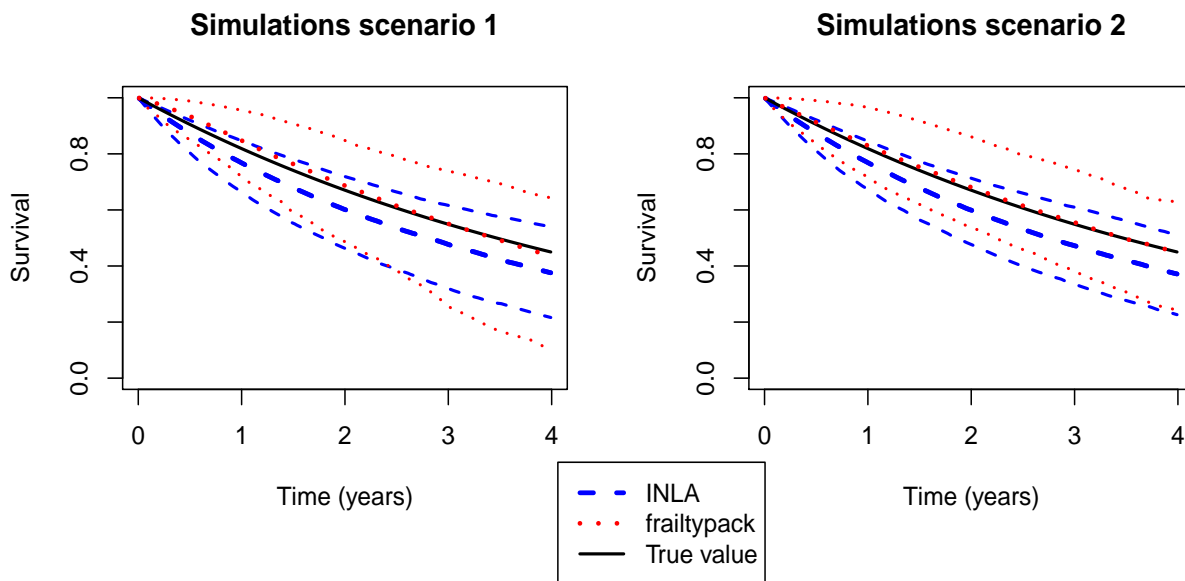


Figure 1: Quantiles at 2.5%, 50% and 97.5% of the baseline survival curves estimated with **frailtypack** and **R-INLA** in the simulations (Section 3).

4.1 GERCOR study

4.1.1 Description

It is a randomized clinical trial investigating two treatment strategies that included a total of 220 patients with metastatic colorectal cancer. The reference strategy (arm A) corresponds to FOLFIRI (irinotecan) followed by FOLFOX6 (oxaliplatin) while arm B involves the reverse sequence. Patients were randomly assigned from December 1997 to September 1999 and the date chosen to assess overall survival was August 30, 2002. Complete data are available on 205 individuals for data analysis. Among them, 165 (80%) died during the follow-up. There are 1475 repeated measurements for the biomarker, 174 of which are zero values (12%). A summary of the dataset structure is given in Table 3. Our model uses death as the terminal event and the repeated SLD measurements (in centimeters) as the semicontinuous biomarker. Additional baseline covariates collected at the start of the study are also included, including performance status (0/1/2), lung metastatic site (Y/N), previous adjuvant radiotherapy (Y/N), previous surgery (no surgery/curative/palliative) and metastases (metachronous/synchronous). The first analysis of this dataset (Tournigand et al. (2004)) did not find any significant difference between the two treatment strategies using classic survival analysis methods (i.e. log-rank tests). A trivariate joint model has been applied to this study for the simultaneous analysis of the longitudinal SLD, recurrent events (progression of lesions not included in the SLD or new lesions) and the terminal event (Król et al. (2018)). A flexible mechanistic model using ordinary differential equation was proposed to fit the biomarker dynamics. The results shows a greater decline of the SLD for treatment arm A compared to treatment arm B. Moreover, the model finds a strong association between the biomarker model and the risk of terminal event. However the interpretation of this treatment effect is difficult due to the non-linear transformation applied to the outcome (Box-Cox) and the use of a non-linear mechanistic model. Finally, a conditional two-part joint model was recently proposed (Rustand et al. (2020)), which showed a significant treatment effect on the positive values of the biomarker (and no treatment effect on the probability of zero value). The model was able to show that when taking into account this treatment effect on the biomarker, the risk of terminal event is not significantly different between the two treatment arms. However, the interpretation of the treatment effect on the biomarker value could have been impacted by a logarithm transformation used on the outcome. Instead, we use a GLM with log link function here for the continuous part of the biomarker.

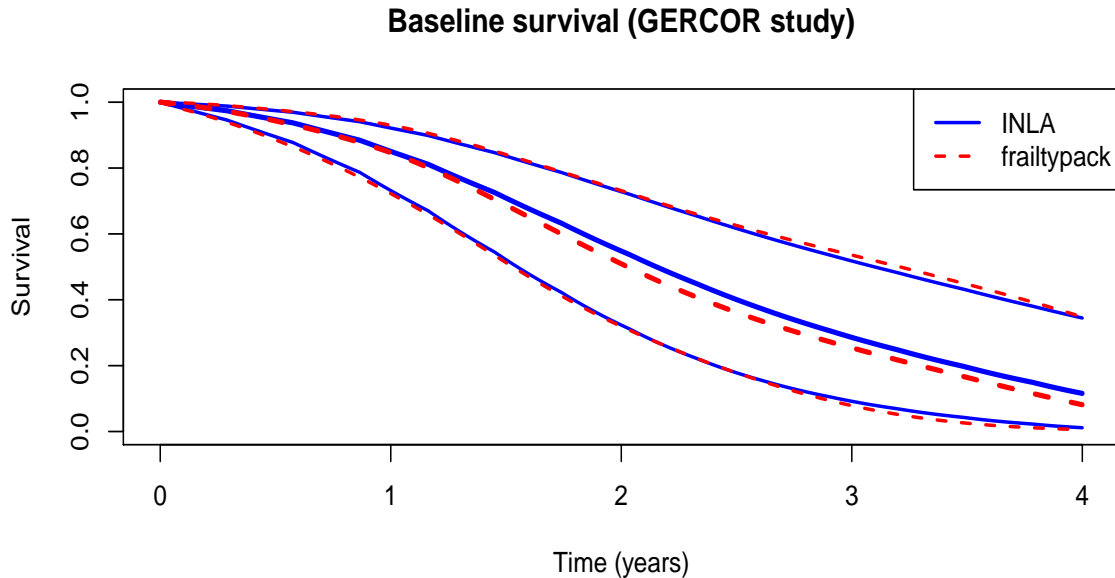


Figure 2: Baseline survival curves and their 95% confidence and credible intervals obtained from the application of the TPJM to the GERCOR study with **frailtypack** and **R-INLA**, respectively.

4.1.2 Results

As presented in Table 4, the fixed effect parameter estimates in the binary, continuous and survival parts are quite similar between the frequentist and Bayesian approaches, with a slightly lower variability for the parameters estimated with **R-INLA**. Treatment arm A (reference treatment) is associated with a significant reduction in the SLD conditional on a positive value with **R-INLA** ($\hat{\beta}_1 = -0.33$, $SD=0.07$) and **frailtypack** ($\hat{\beta}_1 = -0.35$, $SD=0.08$), compared to treatment arm B. This is because the treatment by time interaction balances out the negative slope obtained, i.e., with **R-INLA** ($\hat{\beta}_9 = 0.30$, $SD=0.10$) and with **frailtypack** ($\hat{\beta}_9 = 0.33$, $SD=0.10$). Therefore, conditional on a positive value of the SLD, treatment arm A is associated with a reduction of $\sim 26\%$ of the SLD per year ($1 - \exp(-0.30)$) while treatment arm B is associated with no change over time. The hazard ratio of treatment arm B versus treatment arm A that evaluates the change in the risk of death was similar between **R-INLA** ($HR=1.30$, $CI\ 0.84 - 1.92$) and **frailtypack** ($HR=1.26$, $CI\ 0.85 - 1.79$). The main difference between **R-INLA** and **frailtypack** is in the estimation of the parameters for the association between the two-part model for the biomarker and the survival model. There is a positive and significant association between the random intercept ($\hat{\varphi}_a = 0.11$, $SD=0.03$) from the binary part, the random intercept ($\hat{\varphi}_{b_0} = 0.66$, $SD=0.07$) and the random slope ($\hat{\varphi}_{b_1} = 0.83$, $SD=0.26$) from the continuous part and the risk of event with **R-INLA**. This association has a slightly lower effect size and much larger variability with **frailtypack** ($\hat{\varphi}_a = 0.13$, $SD=0.13$, $\hat{\varphi}_{b_0} = 0.46$, $SD=0.37$ and $\hat{\varphi}_{b_1} = 0.55$, $SD=0.60$), so that the effects are not significant. This is in line with our simulation results (scenario 1) where the association structure was estimated with better precision with **R-INLA**. The computation time is much longer with **frailtypack** compared to **R-INLA**, the latter fits the data in less than a minute. The computation time for **frailtypack** increases quickly with the sample size and the model complexity (number of parameters and dimension of the random effects). The model was estimated in 60 minutes with **frailtypack** with 8 CPUs and this reduces to 10 minutes with 80 CPUs. The differences found in the association structure estimates is important when assessing the relationship between the biomarker dynamics and the risk of event. For instance, let's assume a clinician is interested in the top 15% patients who had the largest SLD increase during follow-up compared to the average patient. Their random effect b_{1i} should be higher than 1 standard deviation, that is from Table 4, $b_{1i} > 0.51$ with **R-INLA** (respectively $b_{1i} > 0.52$ with **frailtypack**). Conditional on $b_{1i} > 0.51$

(respectively $b_{1i} > 0.52$), the mean values of the random effects can be derived by sampling from a conditional multivariate normal distribution with correlation matrix given in Table 4. These conditional means are 2.35, -0.14 and 0.77 for a , b_0 and b_1 , respectively (2.25, -0.20 and 0.80 with **frailtypack**). Therefore, these top 15% individuals increase their chance to have the terminal event (i.e., to die) measured by an hazard ratio of $HR = \exp(0.11 * 2.35 + 0.66 * (-0.14) + 0.83 * 0.77) = \exp(0.81) = 2.24$, $CI=1.46 - 3.51$, compared to a patient who has an average longitudinal SLD profile. **Frailtypack** underestimates this risk as we obtain $HR = \exp(0.13 * 2.25 + 0.46 * (-0.20) + 0.55 * 0.80) = \exp(0.64) = 1.90$, $CI=1.21 - 3.05$. The confidence intervals were obtained by sampling parameters from the Hessian matrix with **frailtypack** and the credible intervals from the posterior distribution of the parameters with **R-INLA**. Figure 2 shows a similar estimation of the baseline survival curves obtained with **frailtypack** and **R-INLA**.

Table 4: Application of the Bayesian and frequentist two-part joint models with shared random effects to the GERCOR study with the R packages **R-INLA** and **frailtypack**

Package		R-INLA Est. † (SD ‡)	frailtypack Est. (SD)
Binary part (SLD>0 versus SLD=0)			
intercept	α_0	5.05*** (0.68)	5.10*** (0.69)
time (year)	α_1	-2.11*** (0.39)	-2.14*** (0.41)
treatment (B/A)	α_2	-1.17 (0.69)	-1.24 (0.69)
PS (1 vs. 0)	α_3	1.83** (0.57)	1.97*** (0.58)
PS (2 vs. 0)	α_4	1.72 (1.08)	1.72 (1.18)
Previous_radio (Y/N)	α_5	0.82 (0.70)	0.85 (0.72)
Lung (Y/N)	α_6	1.84** (0.67)	2.14** (0.75)
time:treatment (B/A)	α_7	0.30 (0.46)	0.31 (0.47)
Continuous part ($E[Y_{ij} Y_{ij} > 0]$)			
intercept	β_0	1.99*** (0.16)	2.04*** (0.23)
time (years)	β_1	-0.33*** (0.07)	-0.35*** (0.08)
treatment (B/A)	β_2	-0.28** (0.10)	-0.35*** (0.10)
PS (1 vs. 0)	β_3	0.42*** (0.11)	0.42*** (0.11)
PS (2 vs. 0)	β_4	0.53** (0.17)	0.55** (0.17)
Previous_surgery (curative)	β_5	-0.53** (0.20)	-0.61** (0.23)
Previous_surgery (palliative)	β_6	0.00 (0.15)	-0.02 (0.19)
Previous_radio (Y/N)	β_7	-0.25* (0.12)	-0.22 (0.13)
Metastases (metachronous vs. synchronous)	β_8	0.43** (0.17)	0.46** (0.16)
time:treatment (B/A)	β_9	0.30** (0.10)	0.33** (0.10)
residual S.E.	σ_ε	0.39*** (0.00)	0.42*** (0.01)
Death risk			
treatment (B/A)	γ_1	0.24 (0.21)	0.21 (0.19)
PS (1 vs. 0)	γ_2	0.81*** (0.22)	0.78*** (0.21)
PS (2 vs. 0)	γ_3	1.59*** (0.34)	1.58*** (0.33)
Previous_surgery (curative)	γ_4	-0.93* (0.42)	-0.97* (0.42)
Previous_surgery (palliative)	γ_5	-0.51 (0.30)	-0.53 (0.30)
Metastases (metachronous vs. synchronous)	γ_6	0.95** (0.35)	0.99** (0.34)
Association			
Intercept (binary part)	φ_a	0.11*** (0.03)	0.13 (0.13)
Intercept (continuous part)	φ_{b_0}	0.66*** (0.07)	0.46 (0.37)
Slope (continuous part)	φ_{b_1}	0.83** (0.26)	0.55 (0.60)
Random effects's standard deviation			
intercept (binary part)	σ_a	2.67	2.81
intercept (continuous part)	σ_{b_0}	0.67	0.71
slope (continuous part)	σ_{b_1}	0.51	0.52
	$corr_{ab_0}$	0.49	0.55
	$corr_{ab_1}$	0.57	0.53
	$corr_{b_0b_1}$	-0.14	-0.18
Computation time (Intel Xeon E5-4627 v4 2.60 GHz)			
8 CPUs		< 1 minute	60 minutes
80 CPUs		< 1 minute	10 minutes

† Posterior mean, ‡ Posterior standard deviation, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

4.2 PRIME study

4.2.1 Description

The Panitumumab Randomized Trial in Combination with Chemotherapy for Metastatic Colorectal Cancer to Determine Efficacy (PRIME) study is a more challenging application for fitting the TPJM because it includes information about the KRAS mutation status (exon 2 codons 12/13), which has been shown to impact the clinical response to treatment in metastatic colorectal cancer patients (Van Cutsem et al. (2008); Normanno et al. (2009); Bokemeyer et al. (2008)). It is therefore an important risk modifier and clinicians

are interested to assess treatment by mutation interaction in order to tailor treatment to patients' genetic risk (Marabelle et al. (2020)). This dataset is freely available on ProjectDataSphere.org.

The PRIME study is a randomized clinical trial that compares the efficacy and safety of panitumumab (anti-EGFR) in combination with FOLFOX4 (chemotherapy) with those of FOLFOX4 alone in the first-line treatment of patients, according to KRAS exon 2 status (Wild type or Mutant type). Between August 2006 and February 2008, 1183 patients were randomly assigned to receive treatment arm A (FOLFOX4 alone) or treatment arm B (panitumumab + FOLFOX4). The data for analysis includes a subset of 442 patients (i.e., 741 excluded from the publicly available dataset). There are 2372 repeated measurements of the SLD, 99 of which are zero values (4%). The small rate of zero measurements in the SLD distribution leads to a large variability in the binary part, however zeros corresponds to patients with a complete shrinkage of their target lesions, which is a very relevant information for clinicians about treatment effect. The number of individual repeated measurements for this biomarker varies between 1 and 24 with a median of 5. The death rate is 74%, corresponding to 328 deaths. Summary statistics of the dataset are given in Table 3. Additional baseline covariates collected at the start of the study are also included, including metastases to liver at study entry (Y/N), the number of baseline metastases sites (1/2/3/4+), age (<60/60-69/>=70) and baseline ECOG performance status (0/1/2). We used a global backward selection procedure for each component of the model to select the covariates to include in the final joint model. The conclusions of the study are presented in Douillard et al. (2013) and show the importance of taking into account the mutation status when assessing treatment effect. Among patients without mutated KRAS, treatment arm B was associated with a slightly significant reduced risk of death compared to treatment arm A. For patients with mutated KRAS, treatment arm B was associated with a non-significant increase in the risk of death compared to treatment arm A. In the results, the mean parameters and their standard deviation are obtained by taking the ML estimates and the inverse Hessian matrix with **frailtypack** while the posterior mean and standard deviation of the posterior distribution were used with **R-INLA**.

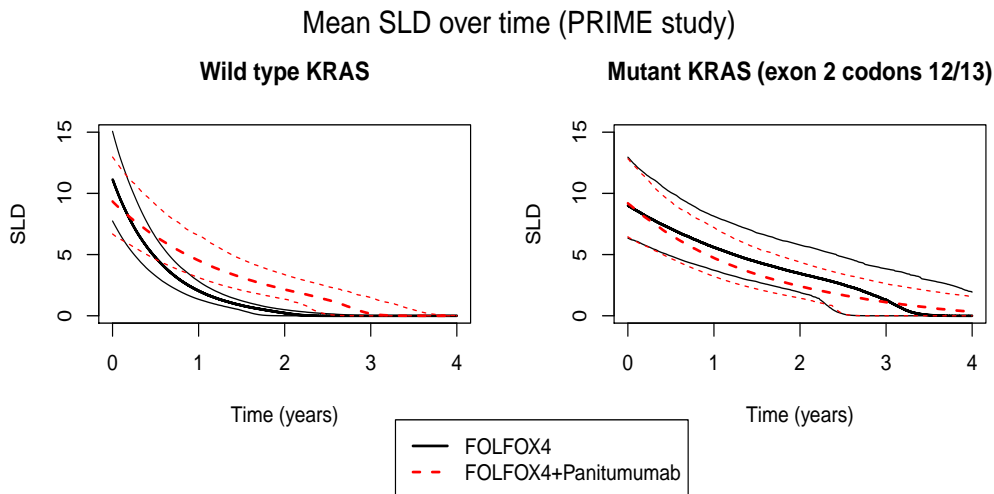


Figure 3: Mean biomarker value according to treatment received for patients with wild type KRAS status (left) and mutant KRAS status (right). The 95% credible intervals are obtained by resampling from the posterior parameter distributions.

4.2.2 Results

As presented in Table 5, in the binary part of the TPJM, the intercept is very large ($\hat{\alpha}_0 = 21.32$, $SD=3.99$), corresponding to a high probability of positive value at baseline. This probability is increased for patients with mutated KRAS ($\hat{\alpha}_3 = 5.63$, $SD=5.73$) and patients receiving treatment arm B ($\hat{\alpha}_2 = 2.41$, $SD=4.70$) but with large standard deviations so that these effects are not significant. The slope parameter with time is negative and significant ($\hat{\alpha}_1 = -10.43$, $SD=1.95$), meaning that patients without mutated KRAS and

receiving treatment A have a higher odds of zero SLD value over time, i.e., complete response to treatment. This odds decreases, but not significantly, among patients with either mutated KRAS ($\hat{\alpha}_6 = 1.68$, $SD=2.90$) or receiving treatment B ($\hat{\alpha}_5 = 1.99$, $SD=2.24$) and is slightly attenuated in patients with both mutated KRAS and receiving treatment arm B ($\hat{\alpha}_7 = -1.38$, $SD=5.32$).

In the continuous part of the TPJM, patients with the wild type KRAS status and in treatment arm A are associated with a decrease in the SLD value over time conditional on a positive SLD value ($\hat{\beta}_1 = -1.71$, $SD=0.09$). This reduction of SLD over time is attenuated in patients with mutated KRAS ($\hat{\beta}_{12} = 1.22$, $SD=0.14$) or receiving treatment B ($\hat{\beta}_{11} = 0.98$, $SD=0.13$). Patients with the KRAS mutation and who received treatment B have a similar SLD trend over time as patients with the KRAS mutation who received treatment A or patients who received treatment B but with the wild type KRAS status because of the negative interaction term between time, treatment and KRAS status ($\hat{\beta}_{13} = -1.16$, $SD=0.20$).

In the survival part, the model shows no significant difference between treatment arms for the risk of death ($\hat{\gamma}_1 = 0.10$, $SD=0.16$). Besides, patients with mutated KRAS have similar risk of death compared to patients with the wild type ($\hat{\gamma}_2 = 0.21$, $SD=0.17$), so do patients with mutated KRAS receiving treatment B ($\hat{\gamma}_3 = 0.04$, $SD=0.23$). The random effect from the binary part and the random slope from the continuous part are not associated to the risk of death ($\hat{\varphi}_a = 0.00$, $SD=0.01$ and $\hat{\varphi}_{b1} = 0.13$, $SD=0.15$) but the random intercept from the continuous part ($\hat{\varphi}_{b0} = 0.36$, $SD=0.09$) have a positive and highly significant association with the risk of event. This means that conditional on a positive value, the individual deviation from the mean baseline value of the SLD is predictive of the risk of event. Similarly to the GERCOR study, we can compare the top 15% patients with the smallest SLD at baseline to the average patient, their risk of death is reduced by 32% ($HR=0.68$, $CI=0.61 - 0.78$).

In conclusion, we did not find a direct effect of treatment B vs. A on the risk of death while the initial study (Douillard et al. (2013)) finds a slightly significant improvement in overall survival for patients with wild type KRAS status ($HR=0.78$, $CI=0.62 - 0.99$), likely because of the reduced sample size available for our analysis (publicly available dataset only includes 37% of the original set of patients). Interestingly, the analysis of the continuous part of the TPJM suggests that the reduction of the SLD over time conditional on a positive value is attenuated with treatment B compared to treatment A for patients with wild type KRAS status. A graphical representation of the mean biomarker evolution over time according to KRAS mutation status and treatment received is depicted in Figure 3. It confirms the suggested significant difference between treatment arms for patients with wild type KRAS status and shows no treatment effect for patients with mutant KRAS.

5 Discussion

In this article, we developed a Bayesian estimation approach based on the INLA algorithm for two-part joint models for a longitudinal semicontinuous biomarker and a terminal event. We also provided a comparison with a frequentist alternative approach previously implemented into the **frailtypack** package, using small sample sizes as seen in cancer clinical trial evaluation. The frequentist estimation raises several limitations both in terms of model complexity and computation time. The Bayesian estimation proposed in the R package **R-INLA** has been recently introduced to fit complex joint models (Van Niekerk, Bakka, Rue, and Schenk (2019)) but to our knowledge, has never been proposed for TPJMs. Accounting for the semicontinuous nature of the biomarker, i.e. the SLD, and being able to fit joint models with more complex association structures between the biomarker and the terminal event, can be quite relevant in clinical applications by providing critical insights into the direct and indirect effect of a treatment on the event of interest. This was illustrated in our simulations and applications to two randomized cancer clinical trials.

In our simulation studies, the estimation with **R-INLA** was found superior to **frailtypack** in terms of computation time and precision of the fixed effects estimation. The point estimates from **frailtypack** yielded closer results to the true values of the random effects' standard deviations, the residual error term and the baseline hazard function than the posterior mean from **R-INLA**, even though **R-INLA** recovered all parameters well based on the estimated credible intervals.

Our first application to the GERCOR randomized clinical trial investigating two treatment lines to treat metastatic colorectal cancer shows some differences between the two estimation approaches. In line with our simulations, the variability of the parameter estimates is reduced with **R-INLA**, in particular for the

Table 5: Application of the Bayesian two-part joint model with shared random effects to the PRIME study with the R package **R-INLA**

Package		R-INLA Est. [†] (SD [‡])
Binary part (SLD>0 versus SLD=0)		
intercept	α_0	21.32*** (3.99)
time (year)	α_1	-10.43*** (1.95)
treatment (B/A)	α_2	2.41 (4.70)
kras (MT/WT)	α_3	5.63 (5.73)
treatment (B/A):kras (MT/WT)	α_4	3.66 (8.43)
time:treatment (B/A)	α_5	1.99 (2.24)
time:kras (MT/WT)	α_6	1.68 (2.90)
time:treatment (B/A):kras (MT/WT)	α_7	-1.38 (5.32)
Continuous part $E[Y_{ij} Y_{ij} > 0]$		
intercept	β_0	2.41*** (0.17)
time (years)	β_1	-1.71*** (0.09)
treatment (B/A)	β_2	-0.17 (0.10)
kras (MT/WT)	β_3	-0.20 (0.11)
liver metastases (Y/N)	β_4	0.63*** (0.14)
ECOG (symptoms but ambulatory vs. fully active)	β_5	0.19* (0.08)
ECOG (in bed less than 50% of the time vs. fully active)	β_6	0.52** (0.18)
baseline metastases sites (2 vs. 1)	β_7	0.08 (0.12)
baseline metastases sites (3 vs. 1)	β_8	0.26* (0.12)
baseline metastases sites (4+ vs. 1)	β_9	0.19 (0.12)
treatment (B/A):kras (MT/WT)	β_{10}	0.19 (0.15)
time:treatment (B/A)	β_{11}	0.98*** (0.13)
time:kras (MT/WT)	β_{12}	1.22*** (0.14)
time:treatment (B/A):kras (MT/WT)	β_{13}	-1.16*** (0.20)
residual S.E.	σ_ϵ	0.31 (0.01)
Death risk		
treatment (B/A)	γ_1	0.10 (0.16)
kras (MT/WT)	γ_2	0.21 (0.17)
treatment (B/A):kras (MT/WT)	γ_3	0.04 (0.23)
age (60-69 vs. <60)	γ_4	0.08 (0.13)
age (70+ vs. <60)	γ_5	0.22 (0.14)
liver metastases (Y/N)	γ_6	0.03 (0.23)
ECOG (symptoms but ambulatory vs. fully active)	γ_7	0.30** (0.12)
ECOG (in bed less than 50% of the time vs. fully active)	γ_8	0.81** (0.26)
baseline metastases sites (2 vs. 1)	γ_9	0.12 (0.20)
baseline metastases sites (3 vs. 1)	γ_{10}	0.32 (0.20)
baseline metastases sites (4+ vs. 1)	γ_{11}	0.43* (0.21)
Association		
Intercept (binary part)	φ_a	0.00 (0.01)
Intercept (continuous part)	φ_{b_0}	0.36*** (0.09)
Slope (continuous part)	φ_{b_1}	0.13 (0.15)
Random effects's standard deviation		
intercept (binary part)	σ_a	11.17
intercept (continuous part)	σ_{b_0}	0.74
slope (continuous part)	σ_{b_1}	0.74
	$corr_{ab_0}$	0.03
	$corr_{ab_1}$	0.84
	$corr_{b_0b_1}$	-0.13

[†] Posterior mean, [‡] Posterior standard deviation, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

association parameters between the biomarker and the survival outcome where **R-INLA** concluded to a strong association unlike **frailtypack**, which found a non-significant association with an attenuated effect size. The second application to the PRIME study reflects upon the concept that treatment response might depend on genetic alterations or tumor biomarker status (DNA/RNA/protein features). There is now a great interest in identifying subgroups of patients with specific patterns of responses however most methods provide an average effect of covariates. Instead, our model can distinguish complete responders (i.e. SLD=0) from partial responders (i.e. SLD >0). This leads also to an increase in model complexity as additional covariates and random effects are included in each submodel of the TPJM. The frequentist approach proposed in **frailtypack** can have convergence issues or sometimes cannot be fitted at all as this was the case for the PRIME study. Only the Bayesian approach could be used in that situation. Interestingly, the analysis of the continuous part of the TPJM suggested that the subgroup of patients with the KRAS mutation receiving treatment B had a similar decrease of the SLD over time compared to the KRAS mutation group receiving treatment A or patients who received treatment B with wild type KRAS status. Therefore, the lack of response to the addition of anti-EGFR to FOLFOX4 chemotherapy was not fully explained by the KRAS mutation status. This could motivate further investigations of the interaction between KRAS mutation and

anti-EGFR therapies to treat advanced colorectal cancer patients, in particular by including information on other somatic tumour mutations (e.g., BRAF or NRAS mutations).

Our work has several limitations. Our applications focused on clinical trials of very advanced cancers, which often have high death rates and small proportions of complete responses (i.e. SLD=0). In situations where we have a higher proportion of complete responders, the relative performances of **R-INLA** vs. **frailtypack** could be different. The conclusions might be different for different settings (i.e. with higher zero rate and reduced censoring). For instance a meta-analysis evaluating the responses among non-Hodgkin's Lymphoma patients estimated complete response rates (i.e. SLD= 0) ranging from 1.2% to 84% in the different pooled clinical trials (Mangal et al. (2018)). We also notice that the two models estimated with **R-INLA** and **frailtypack** are not completely comparable because of the difference in the approximation of the baseline hazard function. Besides the shared random effects, other association structures have also been proposed such as the current value association, i.e., it uses the current level of the biomarker, and is available in **frailtypack**. For the TPJMs, the current value of the biomarker is defined as $E[Y_{ij}] = Prob(Y_{ij} > 0)E[Y_{ij}|Y_{ij} > 0]$, which is non linear. It cannot be directly defined as part of the latent gaussian model and more work is warranted to be included in **R-INLA**. It would be also interesting to consider a Bayesian development for the marginal TPJM we recently proposed (Rustand, Briollais, and Rondeau (2020)). Finally, the definition of the hyperparameter prior distributions are an important aspect of Bayesian estimation. In this work, the PC priors provided a general setting for the priors since they provide a natural avenue to incorporate knowledge from the practitioner about the expected size of the parameter and they are constructed to be proper and avoid overfitting. Alternative prior choices for the hyperparameters can be used in **R-INLA** if the practitioner possesses motivation for it from expert or prior knowledge.

The reduction in the computation times with **R-INLA** was beyond our expectations. It improves drastically the applicability of the Bayesian estimation for complex models such as the TPJMs and other families of joint models, such as a bivariate joint model for recurrent events and a terminal event or a trivariate joint model for a longitudinal biomarker, recurrent events and a terminal event, which are currently available in **frailtypack**. Finally, **R-INLA** can accommodate multiple longitudinal outcomes while **frailtypack** is currently limited to a single longitudinal outcome.

Acknowledgements

This publication is based on research using information obtained from www.projectdatasphere.org, which is maintained by Project Data Sphere, LLC. Neither Project Data Sphere, LLC nor the owner(s) of any information from the web site have contributed to, approved or are in any way responsible for the contents of this publication.

References

- Andrinopoulou, E.-R. and D. Rizopoulos (2016). Bayesian shrinkage approach for a joint model of longitudinal and survival outcomes assuming different association structures. *Statistics in medicine* 35(26), 4813–4823.
- Bokemeyer, C., I. Bondarenko, J. Hartmann, F. De Braud, C. Volovat, J. Nippgen, C. Stroh, I. Celik, and P. Koralewski (2008). Kras status and efficacy of first-line treatment of patients with metastatic colorectal cancer (mrcr) with folfox with or without cetuximab: The opus experience. *Journal of Clinical Oncology* 26(15_suppl), 4000–4000.
- Chi, Y.-Y. and J. G. Ibrahim (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics* 62(2), 432–445.
- Douillard, J.-Y., K. S. Oliner, S. Siena, J. Tabernero, R. Burkes, M. Barugel, Y. Humblet, G. Bodoky, D. Cunningham, J. Jassem, et al. (2013). Panitumumab–folfox4 treatment and ras mutations in colorectal cancer. *New England Journal of Medicine* 369(11), 1023–1034.
- Hanson, T. E., A. J. Branscum, and W. O. Johnson (2011). Predictive comparison of joint longitudinal-survival modeling: a case study illustrating competing approaches. *Lifetime data analysis* 17(1), 3–28.

- Henderson, R., P. Diggle, and A. Dobson (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* 1(4), 465–480.
- Hespanhol, L., C. S. Vallio, L. M. Costa, and B. T. Saragiotto (2019). Understanding and interpreting confidence and credible intervals around effect estimates. *Brazilian journal of physical therapy* 23(4), 290–301.
- Król, A., C. Tournigand, S. Michiels, and V. Rondeau (2018). Multivariate joint frailty model for the analysis of nonlinear tumor kinetics and dynamic predictions of death. *Statistics in Medicine* 37(13), 2148–2161.
- Lawrence Gould, A., M. E. Boye, M. J. Crowther, J. G. Ibrahim, G. Quartey, S. Micallef, and F. Y. Bois (2015). 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.
- Mangal, N., A. H. Salem, M. Li, R. Menon, and K. J. Freise (2018). Relationship between response rates and median progression-free survival in non-hodgkin’s lymphoma: A meta-analysis of published clinical trials. *Hematological Oncology* 36(1), 37–43.
- Marabelle, A., M. Fakih, J. Lopez, M. Shah, R. Shapira-Frommer, K. Nakagawa, H. C. Chung, H. L. Kindler, J. A. Lopez-Martin, W. H. Miller Jr, et al. (2020). Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 keynote-158 study. *The Lancet Oncology* 21(10), 1353–1365.
- Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics* 11(2), 431–441.
- Martino, S., R. Akerkar, and H. Rue (2011). Approximate bayesian inference for survival models. *Scandinavian Journal of Statistics* 38(3), 514–528.
- Martins, T. G., D. Simpson, F. Lindgren, and H. Rue (2013). Bayesian computing with inla: new features. *Computational Statistics & Data Analysis* 67, 68–83.
- Muth, C., Z. Oravecz, and J. Gabry (2018). User-friendly bayesian regression modeling: A tutorial with rstanarm and shinystan. *Quantitative Methods for Psychology* 14(2), 99–119.
- Normanno, N., S. Tejpar, F. Morgillo, A. De Luca, E. Van Cutsem, and F. Ciardiello (2009). Implications for kras status and egfr-targeted therapies in metastatic crc. *Nature reviews Clinical oncology* 6(9), 519.
- R. Brown, E. and J. G. Ibrahim (2003). A bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* 59(2), 221–228.
- Rizopoulos, D. et al. (2016). The r package jmbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *Journal of Statistical Software* 72(i07).
- Rizopoulos, D. and P. Ghosh (2011). A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in medicine* 30(12), 1366–1380.
- Rue, H. and L. Held (2005). *Gaussian Markov random fields: theory and applications*. CRC press.
- Rue, H., S. Martino, and N. Chopin (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71(2), 319–392.
- Rue, H., A. Riebler, S. H. Sørbye, J. B. Illian, D. P. Simpson, and F. K. Lindgren (2017). Bayesian computing with inla: A review. *Annual Review of Statistics and Its Application* 4(1), 395–421.
- Rustand, D., L. Briollais, and V. Rondeau (2020). A marginal two-part joint model for a longitudinal biomarker and a terminal event with application to advanced head and neck cancers. (Under submission).

- Rustand, D., L. Briollais, C. Tournigand, and V. Rondeau (2020). Two-part joint model for a longitudinal semicontinuous marker and a terminal event with application to metastatic colorectal cancer data. *Biostatistics*. kxaa012.
- Schenk, O. and K. Gärtner (2004). Solving unsymmetric sparse systems of linear equations with pardiso. *Future Generation Computer Systems* 20(3), 475–487.
- Simpson, D., H. Rue, A. Riebler, T. G. Martins, S. H. Sørbye, et al. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science* 32(1), 1–28.
- Smith, V. A., J. S. Preisser, B. Neelon, and M. L. Maciejewski (2014). A marginalized two-part model for semicontinuous data. *Statistics in Medicine* 33(28), 4891–4903.
- Song, X., M. Davidian, and A. A. Tsiatis (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics* 58(4), 742–753.
- Tournigand, C., T. André, E. Achille, G. Lledo, M. Flesh, D. Mery-Mignard, E. Quinaux, C. Couteau, M. Buyse, G. Ganem, B. Landi, P. Colin, C. Louvet, and A. de Gramont (2004). Folfiri followed by folfox6 or the reverse sequence in advanced colorectal cancer: A randomized gercor study. *Journal of Clinical Oncology* 22(2), 229–237.
- Van Cutsem, E., I. Lang, G. D’haens, V. Moiseyenko, J. Zaluski, G. Folprecht, S. Tejpar, O. Kisker, C. Stroh, and P. Rougier (2008). Kras status and efficacy in the first-line treatment of patients with metastatic colorectal cancer (mrcr) treated with folfiri with or without cetuximab: The crystal experience. *Journal of Clinical Oncology* 26(15-suppl), 2–2.
- Van Niekerk, J., H. Bakka, and H. Rue (2019). Joint models as latent gaussian models-not reinventing the wheel. *arXiv:1901.09365*.
- Van Niekerk, J., H. Bakka, and H. Rue (2020). Stable non-linear generalized bayesian joint models for survival-longitudinal data. *Sankhya A (Accepted)*.
- Van Niekerk, J., H. Bakka, H. Rue, and L. Schenk (2019). New frontiers in bayesian modeling using the inla package in r. *arXiv:1907.10426*.
- Wulfsohn, M. S. and A. A. Tsiatis (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 330–339.
- Xu, J. and S. L. Zeger (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 50(3), 375–387.

5.3 Additional remarks

5.3.1 On the association structure

In this Chapter, we only focused on the shared random effects association structure presented in Chapter 2 and proposed in Chapter 3 and 4. We also investigated the possibility to use a current level association, which is available in **frailtypack** as described in Chapter 3. The current level association is also available with **R-INLA** in the context of standard joint models. However, the “current value” of a two-part model is non linear as illustrated in Chapter 4 and therefore the two-part joint model with a current value association cannot be directly formulated as a latent GMRF model. Further developments are required to use a current value association for the two-part joint model with **R-INLA**.

5.3.2 On the marginal two-part joint model

Similarly, the marginal two-part model uses a modified lognormal distribution to account for the probability of positive value and therefore requires further developments to be estimated with **R-INLA**. We were therefore only interested in a conditional TPJM in this thesis when focusing on **R-INLA**.

5.3.3 On the prior distributions

When the data does not provide sufficient information to distinguish each subgroup of the population defined by each category of the covariates included in the model, the Bayesian estimation of the parameters relies mostly on the non informative prior (with minimal influence on the inference) and concludes to no significant covariate effect while the frequentist estimation faces unstability (because the parameter estimates only depend on the observed data distribution), as illustrated in the application to the PRIME study.

Chapter 6

General discussion

6.1 Conclusion on the thesis work

The objective of this thesis was to develop statistical methods for the analysis of cancer clinical trials with a focus on the repeated tumor measurements along with survival times. We propose a joint modeling approach for the simultaneous study of the repeated measurements of the tumor size of target lesions (i.e., SLD) and the risk of death. The particular distribution of the biomarker (i.e., semicontinuous) requires an appropriate methodology.

Firstly, we proposed a method for the application of a conditional two-part joint model to analyze the repeated measurements of the SLD and the survival times in a cancer clinical trial. The biomarker distribution is decomposed in order to evaluate both the effect of covariates on the probability of a positive value and the conditional mean of the positive values. We provided an efficient estimation method of the parameters using a maximum penalized likelihood approach. We proposed three different forms for the association structure that allow to answer different questions of interest. With the shared random effects association, the proportional hazards submodel that evaluates the effect of covariates on the risk of death is adjusted for individual heterogeneity of the population, captured by the random effects of the two-part model for the biomarker. Moreover, we can evaluate the effect on the risk of death of an individual's deviation from the population mean distribution of the biomarker. The second proposed association structure evaluates the separate effect of the probability of positive values and the expected value conditional on a positive value on the risk of death. It is helpful to clinicians interested to assess the effect of a complete response of target lesions (i.e., $SLD=0$) versus a partial response (i.e., $SLD>0$) on the risk of terminal event. Moreover, it could help understanding the complex pattern of responses of the tumors to treatment, such as the long term effect of a complete removal of target lesions on the risk of death, as discussed in Section (1.2.2). Finally, the current value association measures the association of the expected value of the SLD (on the log scale) on the risk of terminal event. We compared the two-part joint model with a standard joint model (i.e., a single linear regression model for the biomarker) and a left-censoring joint model (i.e.,

single regression model and zeros are assumed to be censored). The best model choice depends on the research question of interest, the standard and left-censoring joint models give the effect of covariates on the marginal mean of the biomarker while the two-part joint model gives the effect of covariates on both the probability of positive values and the conditional mean of positive values. In a simulation study, we show how a negative treatment effect in the binary part of the two-part joint model can bias negatively the overall treatment effect captured by a standard or a left-censoring joint model and can lead to spurious results. We show in the application how the TPJM is relevant for cancer clinical trials because the subset of complete responders to treatment for whom the disease disappeared, resulting in a zero biomarker value during follow-up, is of particular interest in addition to the distribution of the positive values of the SLD and the survival times.

In the second part, we developed a marginal two-part joint model for a longitudinal semi-continuous biomarker and a terminal event that gives the population average effect of covariates on the biomarker value. It is an alternative formulation of the conditional two-part joint model which gives the effect of covariates on the marginal mean value of the biomarker instead of the conditional mean of positive values. The interpretation of the continuous part of the marginal TPJM is therefore similar to the left-censoring OPJM. We compared the marginal TPJM with the conditional TPJM and the left-censoring OPJM in terms of bias and coverage probabilities in simulation studies. The marginal TPJM is robust to misspecification and provides an accurate inference about the biomarker and covariate effects on the biomarker. However, when the distribution of the biomarker over time is not linear on the log scale, the marginal TPJM can lack flexibility and can provide biased parameter estimates. When there is only interest in the effect of covariates on the marginal mean of the biomarker and the biomarker's distribution is assumed to be censored, resulting in the zero excess, the left-censoring OPJM is appropriate. When true zeros are observed and there is an interest in the marginal mean of the biomarker or the probability of positive versus zero value, the marginal TPJM should be preferred. If there is an interest in the positive values of the biomarker (i.e., zeros excluded) and the probability of positive values, the conditional TPJM must be used. An application of the marginal TPJM to a randomized clinical trial of advanced head and neck cancer illustrates these differences. Moreover, an appropriate model choice criterion (i.e., LCV) shows that the marginal TPJM fitted the data better than the conditional TPJM.

In the final part, we addressed the limitations of the frequentist framework for the estimation of a conditional two-part joint model. This model had convergence issues in the second part of this thesis when misspecified (scenario 3 of the simulation studies) because of unstable parameter estimations due to the model complexity (number of parameters, dimension of the random effects). We proposed a Bayesian estimation of the conditional two-part joint model, using the recently introduced Integrated Nested Laplace Approximation (INLA) method. This is an efficient alternative to Markov Chain Monte Carlo methods that provide accurate approximations of the posterior marginals needed for Bayesian inference. We compared the INLA method implemented in the **R-INLA** package with a frequentist alternative approach previously implemented into the **frailtypack** package through simulation studies. The comparison of the frequentist and Bayesian frameworks can be tricky as we use maximum likelihood estimation in the frequentist

framework while the Bayesian framework involves prior knowledge in addition to the likelihood of the observed data. However, we assume non informative prior distributions for the parameters and therefore evaluate the Bayesian estimation as an approximation of the MLE. The estimation with **R-INLA** was shown to reduce significantly the computation time and improve the precision of the fixed effects estimation, compared to **frailtypack**. The parameters are all recovered with **R-INLA** based on the estimated credible intervals but **frailtypack** yielded closer results to the true values of the random effects' standard deviations, the residual error term and the baseline hazard function. Our application to the GERCOR clinical trial exhibits similar differences between the two estimation methods as the simulation studies and **R-INLA** concluded to a strong association between the semicontinuous biomarker and the survival times unlike **frailtypack**, which found a non-significant association with an attenuated effect size. An additional application shows how the Bayesian estimation with **INLA** avoids convergence issues. It illustrates the ability of the conditional TPJM to describe subgroups of patients associated with specific patterns of response to treatment and subsequently provides insights into the direct and indirect effect of a treatment on the event of interest, which is relevant in clinical applications. The Bayesian approach proposed in **R-INLA** overcomes the limits of the frequentist framework and allows complex models with high dimensional random effects to be fitted. Moreover, **R-INLA** can accommodate multiple longitudinal outcomes while **frailtypack** is currently limited to a single longitudinal outcome because of the computational burden added when including additional longitudinal outcomes.

6.2 Critical insights and perspectives

The new approach using a two-part model for jointly modeling tumor response and survival proposed in this thesis contributes to the scientific discussion on elucidating objective response criteria for cancer clinical trials. We illustrated this new approach on the basis of the GERCOR study and the PRIME study, both randomized clinical trials in colorectal metastatic cancer as well as the SPECTRUM study, a randomized clinical trial in recurrent and/or metastatic head and neck cancer. These trials only involve advanced (i.e., metastatic) cancer for which the occurrence of zero values for the biomarker is unlikely, therefore resulting in small zero rates. Another relevant application could be to study early-onset cancers as the zero rate is much higher but on the side note, the death rate might be reduced and the survival model would have less statistical power for hypothesis testing. As discussed in this thesis, the use of joint modeling for the tumor response evaluation can be of particular interest for immunotherapies where the traditional response criteria are not well adapted. In perspective, we might apply the proposed two-part joint model to data from an immunotherapy clinical trial in order to analyze the effect of tumor size changes on survival while accounting for the subset of individuals with a complete response of their target lesions.

In our analyses, we used the sum of the longest diameters as the measure of tumor size. It is subject to limitations, in particular due to the unidimensionality of the measure and because it relies on anatomical aspects only. The individual deviations in size change of each lesion instead of the sum of the target lesions might be of interest, using a multivariate biomarker with separate longitudinal outcomes for each target lesion. The Bayesian estimation would be suitable because

the computational burden would increase importantly due to larger number of parameters and higher dimensions of integrals approximated numerically. Moreover, the evolution of the different target lesions is likely to be highly correlated and might induce colinearity in the survival model if each lesion is associated to the risk of death separately. For the tumor size, two or three dimensional measures or volumetric measures could be considered to increase the precision of the response but in currently available data, there is usually no information other than the unidimensional measurement of target lesions. Beyond target lesions, additional non-target lesions can provide information on the disease evolution and could be of interest, they are however not measured according to RECIST criteria. Similarly, new lesions appearing during follow-up are not considered as target lesions but could capture important information on the patient's response to treatment. New lesions and non-target lesions could be taken into account jointly as recurrent events as proposed in Król et al. (2016). However, the progression of non-target lesions and the appearance of new lesions only show a modest improvement in the context of OS prediction and are highly correlated with target lesions response (Litière et al. (2014)).

6.3 General conclusion

In this thesis, we have extended the statistical methods available for the joint analysis of a longitudinal semicontinuous marker and the time to an event with the development of the two-part joint model. Our developments were motivated by the evaluation of anti-tumor therapies in cancer clinical trials for which both the survival and the tumor burden are outcomes of primary interest. The relationship between these outcomes can be assessed using joint modeling in order to take advantage of these two sources of information about treatment effect, simultaneously. This methodological approach allows a better understanding of the relationship between the tumor response to treatment and the risk of death. Therefore, it contributes to clinical research by providing an innovative method for treatment evaluation in cancer clinical trials that overcomes the limitations of standard response criteria (i.e., RECIST). The new statistical model developed is a general method for the joint modeling of a longitudinal semicontinuous biomarker and a terminal event. It is applicable beyond cancer clinical trials since semicontinuous distributions of longitudinal markers are common to many scientific areas (e.g., daily quantity of rainfall or snowfall, goods, food or drug consumption or expenditure, gene expression data, microbiome compositional data).

Bibliography

- Al-Tassan, N. A., N. Whiffin, F. J. Hosking, C. Palles, S. M. Farrington, S. E. Dobbins, R. Harris, M. Gorman, A. Tenesa, B. F. Meyer, et al. (2015). A new gwas and meta-analysis with 1000genomes imputation identifies novel risk variants for colorectal cancer. *Scientific reports* 5, 10442.
- Bokemeyer, C., I. Bondarenko, J. Hartmann, F. De Braud, C. Volovat, J. Nippgen, C. Stroh, I. Celik, and P. Koralewski (2008). Kras status and efficacy of first-line treatment of patients with metastatic colorectal cancer (merc) with folfox with or without cetuximab: The opus experience. *Journal of Clinical Oncology* 26(15_suppl), 4000–4000.
- Borcoman, E., Y. Kanjanapan, S. Champiat, S. Kato, V. Servois, R. Kurzrock, S. Goel, P. Be-dard, and C. Le Tourneau (2019). Novel patterns of response under immunotherapy. *Annals of Oncology* 30(3), 385–396.
- Branchoux, S., C. Bellera, A. Italiano, D. Rustand, A.-F. Gaudin, and V. Rondeau (2019). Immune-checkpoint inhibitors and candidate surrogate endpoints for overall survival across tumour types: A systematic literature review. *Critical reviews in oncology/hematology* 137, 35–42.
- Bray, F., J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* 68(6), 394–424.
- Breslow, N. E. (1972). Discussion of professor cox’s paper. *J Royal Stat Soc B* 34, 216–217.
- Brown, E. R., J. G. Ibrahim, and V. DeGruttola (2005). A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics* 61(1), 64–73.
- Brown, K. F., H. Rungay, C. Dunlop, M. Ryan, F. Quartly, A. Cox, A. Deas, L. Elliss-Brookes, A. Gavin, L. Hounsome, et al. (2018). The fraction of cancer attributable to modifiable risk factors in england, wales, scotland, northern ireland, and the united kingdom in 2015. *British journal of cancer* 118(8), 1130–1141.

- Chai, H., H. Jiang, L. Lin, and L. Liu (2018). A marginalized two-part beta regression model for microbiome compositional data. *PLoS computational biology* 14(7), e1006329.
- Champiat, S., L. Derle, S. Ammari, C. Massard, A. Hollebecque, S. Postel-Vinay, N. Chanut, A. Eggermont, A. Marabelle, J.-C. Soria, et al. (2017). Hyperprogressive disease is a new pattern of progression in cancer patients treated by anti-pd-1/pd-l1. *Clinical Cancer Research* 23(8), 1920–1928.
- Chan, T. A., M. Yarchoan, E. Jaffee, C. Swanton, S. A. Quezada, A. Stenzinger, and S. Peters (2019). Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Annals of Oncology* 30(1), 44–56.
- Chen, E. Z. and H. Li (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics* 32(17), 2611–2617.
- Choi, J.-H., M.-J. Ahn, H.-C. Rhim, J.-W. Kim, G.-H. Lee, Y.-Y. Lee, and I.-S. Kim (2005). Comparison of who and recist criteria for response in metastatic colorectal carcinoma. *Cancer Research and Treatment: Official Journal of Korean Cancer Association* 37(5), 290.
- Chopra, S. S. (2003). Industry funding of clinical trials: benefit or bias? *Jama* 290(1), 113–114.
- Commenges, D. and H. Jacqmin-Gadda (2015). *Dynamical biostatistical models*, Volume 86. CRC Press.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 34(2), 187–202.
- Crowther, M. J. (2013). Stjm: Stata module to fit shared parameter joint models of longitudinal and survival data.
- Crowther, M. J. (2018). merlin-a unified modelling framework for data analysis and methods development in stata. *arXiv preprint arXiv:1806.01615*.
- Dancey, J. E., P. L. Bedard, N. Onetto, and T. J. Hudson (2012). The genetic basis for cancer treatment decisions. *Cell* 148(3), 409–420.
- De Gruttola, V. G., P. Clax, D. L. DeMets, G. J. Downing, S. S. Ellenberg, L. Friedman, M. H. Gail, R. Prentice, J. Wittes, and S. L. Zeger (2001). Considerations in the evaluation of surrogate endpoints in clinical trials: summary of a national institutes of health workshop. *Controlled clinical trials* 22(5), 485–502.
- Duan, N., W. G. Manning, C. N. Morris, and J. P. Newhouse (1983). A comparison of alternative models for the demand for medical care. *Journal of business & economic statistics* 1(2), 115–126.
- Eisenhauer, E. A., P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, et al. (2009). New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer* 45(2), 228–247.

- Faucett, C. L. and D. C. Thomas (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A gibbs sampling approach. *Statistics in Medicine* 15(15), 1663–1685.
- Finak, G., A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, et al. (2015). Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology* 16(1), 1–13.
- Fisher, L. D. and D. Y. Lin (1999). Time-dependent covariates in the cox proportional-hazards regression model. *Annual Review of Public Health* 20(1), 145–157. PMID: 10352854.
- Fisher, R., L. Pusztai, and C. Swanton (2013). Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer* 108(3), 479–485.
- Fiteni, F., V. Westeel, X. Pivot, C. Borg, D. Vernerey, and F. Bonnetain (2014). Endpoints in cancer clinical trials. *Journal of visceral surgery* 151(1), 17–22.
- Fitzmaurice, G. M., G. Molenberghs, and S. R. Lipsitz (1995). Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society: Series B (Methodological)* 57(4), 691–704.
- Franko, J., Q. Shi, J. P. Meyers, T. S. Maughan, R. A. Adams, M. T. Seymour, L. Saltz, C. J. Punt, M. Koopman, C. Tournigand, et al. (2016). Prognosis of patients with peritoneal metastatic colorectal cancer given systemic therapy: an analysis of individual patient data from prospective randomised trials from the analysis and research in cancers of the digestive system (arcad) database. *The Lancet Oncology* 17(12), 1709–1719.
- Fuentes-Antrás, J., M. Provencio, and E. Díaz-Rubio (2018). Hyperprogression as a distinct outcome after immunotherapy. *Cancer Treatment Reviews* 70, 16–21.
- Galluzzi, L., T. A. Chan, G. Kroemer, J. D. Wolchok, and A. López-Soto (2018). The hallmarks of successful anticancer immunotherapy. *Science translational medicine* 10(459).
- García-Hernandez, A., T. Pérez, M. d. C. Pardo, and D. Rizopoulos (2020). Mmrm vs joint modeling of longitudinal responses and time to study drug discontinuation in clinical trials using a “de jure” estimand. *Pharmaceutical Statistics*.
- Gibbs, R. A., J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch’ang, W. Huang, B. Liu, Y. Shen, et al. (2003). The international hapmap project.
- Guedj, J., R. Thiébaud, and D. Commenges (2011). Joint modeling of the clinical progression and of the biomarkers’ dynamics using a mechanistic model. *Biometrics* 67(1), 59–66.
- Han, D., L. Liu, X. Su, B. Johnson, and L. Sun (2019). Variable selection for random effects two-part models. *Statistical methods in medical research* 28(9), 2697–2709.
- Hanahan, D. and R. Weinberg (2011). Hallmarks of cancer: The next generation. *Cell* 144(5), 646 – 674.

- Hansen, E., R. J. Woods, and A. F. Read (2017, 02). How to use a chemotherapeutic agent when resistance to it threatens the patient. *PLOS Biology* 15(2), 1–21.
- Hargadon, K. M., C. E. Johnson, and C. J. Williams (2018). Immune checkpoint blockade therapy for cancer: an overview of fda-approved immune checkpoint inhibitors. *International immunopharmacology* 62, 29–39.
- Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* 72(358), 320–338.
- Hatfield, L. A., M. E. Boye, and B. P. Carlin (2011). Joint modeling of multiple longitudinal patient-reported outcomes and survival. *Journal of biopharmaceutical statistics* 21(5), 971–991.
- Hu, F., J. Goldberg, D. Hedeker, B. Flay, and M. Pentz (1998, January). Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American journal of epidemiology*.
- Ibrahim, J. G., H. Chu, and L. M. Chen (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 28(16), 2796–2801.
- Islami, F., A. Goding Sauer, K. D. Miller, R. L. Siegel, S. A. Fedewa, E. J. Jacobs, M. L. McCullough, A. V. Patel, J. Ma, I. Soerjomataram, et al. (2018). Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the united states. *CA: a cancer journal for clinicians* 68(1), 31–54.
- Jacqmin-Gadda, H., R. Thiébaud, G. Chêne, and D. Commenges (2000, 12). Analysis of left-censored longitudinal data with application to viral load in HIV infection . *Biostatistics* 1(4), 355–368.
- Kamada, T., Y. Togashi, C. Tay, D. Ha, A. Sasaki, Y. Nakamura, E. Sato, S. Fukuoka, Y. Tada, A. Tanaka, et al. (2019). Pd-1+ regulatory t cells amplified by pd-1 blockade promote hyper-progression of cancer. *Proceedings of the National Academy of Sciences* 116(20), 9999–10008.
- Kareva, I. (2018). Chapter 2 - tumor dormancy. In I. Kareva (Ed.), *Understanding Cancer from a Systems Biology Point of View*, pp. 7 – 25. Academic Press.
- Kemp, R. and V. Prasad (2017). Surrogate endpoints in oncology: when are they acceptable for regulatory and clinical decisions, and are they currently overused? *BMC medicine* 15(1), 134.
- Keroui, M., F. Mercier, J. Bertrand, C. Tardivon, R. Bruno, J. Guedj, and S. Desmée (2020). Bayesian inference using hamiltonian monte-carlo algorithm for nonlinear joint modeling in the context of cancer immunotherapy. *Statistics in Medicine*. sim.8756.
- Kim, C. and V. Prasad (2015, 12). Cancer Drugs Approved on the Basis of a Surrogate End Point and Subsequent Overall Survival: An Analysis of 5 Years of US Food and Drug Administration Approvals. *JAMA Internal Medicine* 175(12), 1992–1994.

- Koch, A. L. (1966). The logarithm in biology 1. mechanisms generating the log-normal distribution exactly. *Journal of theoretical biology* 12(2), 276–290.
- Król, A., L. Ferrer, J.-P. Pignon, C. Proust-Lima, M. Ducreux, O. Bouché, S. Michiels, and V. Rondeau (2016). 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.
- Król, A., C. Tournigand, S. Michiels, and V. Rondeau (2018). Multivariate joint frailty model for the analysis of nonlinear tumor kinetics and dynamic predictions of death. *Statistics in Medicine* 37, 2148–2161.
- Kurland, B. F. and P. J. Heagerty (2005). Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. *Biostatistics* 6(2), 241–258.
- Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics* 38(4), 963–974.
- Lawrence Gould, A., M. E. Boye, M. J. Crowther, J. G. Ibrahim, G. Quartey, S. Micallef, and F. Y. Bois (2015). 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.
- Lesaffre, E. and B. Spiessens (2001). 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.
- Lichtenstein, P., N. V. Holm, P. K. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, E. Pukkala, A. Skytthe, and K. Hemminki (2000). Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from sweden, denmark, and finland. *New England journal of medicine* 343(2), 78–85.
- Lievre, A., J.-B. Bachet, D. Le Corre, V. Boige, B. Landi, J.-F. Emile, J.-F. Côté, G. Tomasic, C. Penna, M. Ducreux, et al. (2006). Kras mutation status is predictive of response to cetuximab therapy in colorectal cancer. *Cancer research* 66(8), 3992–3995.
- Lin, H., C. E. McCulloch, B. W. Turnbull, E. H. Slate, and L. C. Clark (2000). A latent class mixed model for analysing biomarker trajectories with irregularly scheduled observations. *Statistics in Medicine* 19(10), 1303–1318.
- Lin, W. M., A. C. Baker, R. Beroukhim, W. Winckler, W. Feng, J. M. Marmion, E. Laine, H. Greulich, H. Tseng, C. Gates, et al. (2008). Modeling genomic diversity and tumor dependency in malignant melanoma. *Cancer research* 68(3), 664–673.
- Litière, S., S. Collette, E. G. de Vries, L. Seymour, and J. Bogaerts (2017). Recist—learning from the past to build the future. *Nature Reviews Clinical Oncology* 14(3), 187–192.

- Litière, S., E. G. De Vries, L. Seymour, D. Sargent, L. Shankar, J. Bogaerts, R. Committee, et al. (2014). The components of progression as explanatory variables for overall survival in the response evaluation criteria in solid tumours 1.1 database. *European Journal of Cancer* 50(10), 1847–1853.
- Liu, L. (2009). Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Statistics in Medicine* 28(6), 972–986.
- Liu, L., M. R. Conaway, W. A. Knaus, and J. D. Bergin (2008). A random effects four-part model, with application to correlated medical costs. *Computational Statistics & Data Analysis* 52(9), 4458–4473.
- Liu, L., J. Z. Ma, and B. A. Johnson (2008). A multi-level two-part random effects model, with application to an alcohol-dependence study. *Statistics in Medicine* 27(18), 3528–3539.
- Liu, L., R. L. Strawderman, M. E. Cowen, and Y.-C. T. Shih (2010). A flexible two-part random effects model for correlated medical costs. *Journal of health economics* 29(1), 110–123.
- Liu, L., R. L. Strawderman, B. A. Johnson, and J. M. O’Quigley (2016). Analyzing repeated measures semi-continuous data, with application to an alcohol dependence study. *Statistical Methods in Medical Research* 25(1), 133–152.
- Liu, L., R. A. Wolfe, and X. Huang (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics* 60(3), 747–756.
- Mangal, N., A. H. Salem, M. Li, R. Menon, and K. J. Freise (2018). Relationship between response rates and median progression-free survival in non-hodgkin’s lymphoma: A meta-analysis of published clinical trials. *Hematological Oncology* 36(1), 37–43.
- Manning, W. G., C. N. Morris, J. P. Newhouse, L. L. Orr, N. Duan, E. B. Keeler, A. Leibowitz, K. H. Marquis, M. S. Marquis, and C. E. Phelps (1981). A two-part model of the demand for medical care: preliminary results from the health insurance study. *Health, economics, and health economics*, 103–123.
- McDavid, A., G. Finak, P. K. Chattopadhyay, M. Dominguez, L. Lamoreaux, S. S. Ma, M. Roederer, and R. Gottardo (2013). Data exploration, quality control and testing in single-cell qpcr-based gene expression experiments. *Bioinformatics* 29(4), 461–467.
- Miller, A., B. Hoogstraten, M. Staquet, and A. Winkler (1981). Reporting results of cancer treatment. *cancer* 47(1), 207–214.
- Molenberghs, G. and G. Verbeke (2001). A review on linear mixed models for longitudinal data, possibly subject to dropout. *Statistical Modelling* 1(4), 235–269.
- Muth, C., Z. Oravecz, and J. Gabry (2018). User-friendly bayesian regression modeling: A tutorial with rstanarm and shinystan. *Quantitative Methods for Psychology* 14(2), 99–119.

- Nielsen, M., J. L. Thomsen, S. Primdahl, U. Dyreborg, and J. A. Andersen (1987, December). Breast cancer and atypia among young and middle-aged women: a study of 110 medicolegal autopsies. *British Journal of Cancer* 56(6), 814–819.
- Normanno, N., S. Tejpar, F. Morgillo, A. De Luca, E. Van Cutsem, and F. Ciardiello (2009). Implications for kras status and egfr-targeted therapies in metastatic crc. *Nature reviews Clinical oncology* 6(9), 519.
- Olsen, M. K. and J. L. Schafer (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association* 96(454), 730–745.
- Pardoll, D. M. (2012). The blockade of immune checkpoints in cancer immunotherapy. *Nature Reviews Cancer* 12(4), 252–264.
- Piedbois, P. and J. M. Croswell (2008). Surrogate endpoints for overall survival in advanced colorectal cancer: a clinician’s perspective. *Statistical Methods in Medical Research* 17(5), 519–527. PMID: 18285441.
- Pomerantz, M. M. and M. L. Freedman (2011). The genetics of cancer risk. *Cancer Journal (Sudbury, Mass.)* 17(6), 416.
- Prasad, V., C. Kim, M. Burotto, and A. Vandross (2015, 08). The Strength of Association Between Surrogate End Points and Survival in Oncology: A Systematic Review of Trial-Level Meta-analyses. *JAMA Internal Medicine* 175(8), 1389–1398.
- Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine* 8(4), 431–440.
- Proust-Lima, C., M. Séne, J. M. Taylor, and H. Jacqmin-Gadda (2014). Joint latent class models for longitudinal and time-to-event data: A review. *Statistical methods in medical research* 23(1), 74–90.
- Rizopoulos, D. (2010). Jm: An r package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software (Online)* 35(9), 1–33.
- Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC press.
- Rizopoulos, D. et al. (2016). The r package jmbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *Journal of Statistical Software* 72(i07).
- Rizopoulos, D. and P. Ghosh (2011). A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in medicine* 30(12), 1366–1380.
- Rondeau, V., J. R. Gonzalez, A. Yassin Mazroui, A. D. Mauguen, A. Laurent, M. Lopez, A. Krol, C. L. Sofeu, J. Dumerc, D. Rustand, et al. (2020). Package ‘frailtypack’.
- Rondeau, V., S. Mathoulin-Pelissier, H. Jacqmin-Gadda, V. Brouste, and P. Soubeyran (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics* 8(4), 708–721.

- Rouanet, A., C. Helmer, J.-F. Dartigues, and H. Jacqmin-Gadda (2019). Interpretation of mixed models and marginal models with cohort attrition due to death and drop-out. *Statistical methods in medical research* 28(2), 343–356.
- Royston, P. and M. K. Parmar (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in medicine* 21(15), 2175–2197.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63(3), 581–592.
- Rue, H., A. Riebler, S. H. Sørbye, J. B. Illian, D. P. Simpson, and F. K. Lindgren (2017). Bayesian computing with inla: A review. *Annual Review of Statistics and Its Application* 4(1), 395–421.
- Rustand, D., L. Briollais, C. Tournigand, and V. Rondeau (2020, 04). Two-part joint model for a longitudinal semicontinuous marker and a terminal event with application to metastatic colorectal cancer data. *Biostatistics*. kxaa012.
- Sakr, W., G. Haas, B. Cassin, J. Pontes, and J. Crissman (1993). The frequency of carcinoma and intraepithelial neoplasia of the prostate in young male patients. *Journal of Urology* 150(2 Part 1), 379–385.
- Seymour, L., J. Bogaerts, A. Perrone, R. Ford, L. H. Schwartz, S. Mandrekar, N. U. Lin, S. Litière, J. Dancey, A. Chen, et al. (2017). irect: guidelines for response criteria for use in trials testing immunotherapeutics. *The Lancet Oncology* 18(3), e143–e152.
- Smith, V. A., M. L. Maciejewski, and M. K. Olsen (2018). Modeling semicontinuous longitudinal expenditures: a practical guide. *Health services research* 53, 3125–3147.
- Smith, V. A., J. S. Preisser, B. Neelon, and M. L. Maciejewski (2014). A marginalized two-part model for semicontinuous data. *Statistics in Medicine* 33(28), 4891–4903.
- Sofeu, C. L., T. Emura, and V. Rondeau (2019). One-step validation method for surrogate endpoints using data from multiple randomized cancer clinical trials with failure-time endpoints. *Statistics in medicine* 38(16), 2928–2942.
- Sofeu, C. L., T. Emura, and V. Rondeau (2020). A joint frailty-copula model for meta-analytic validation of failure time surrogate endpoints in clinical trials. *Biometrical Journal*.
- Stearns, S. C. and R. Medzhitov (2015, August). *Evolutionary Medicine* (First Edition ed.). Oxford, New York: Oxford University Press.
- Strickland, K. C., B. E. Howitt, S. A. Shukla, S. Rodig, L. L. Ritterhouse, J. F. Liu, J. E. Garber, D. Chowdhury, C. J. Wu, A. D. D’Andrea, et al. (2016). Association and prognostic significance of brca1/2-mutation status with neoantigen load, number of tumor-infiltrating lymphocytes and expression of pd-1/pd-l1 in high grade serous ovarian cancer. *Oncotarget* 7(12), 13587.
- Tazdait, M., L. Mezquita, J. Lahmar, R. Ferrara, F. Bidault, S. Ammari, C. Balleyguier, D. Planchard, A. Gazzah, J. Soria, et al. (2018). Patterns of responses in metastatic nscl during pd-1 or pdl-1 inhibitor therapy: comparison of recist 1.1, irrecist and irect criteria. *European Journal of Cancer* 88, 38–47.

- Thiesse, P., L. Ollivier, D. Di Stefano-Louineau, S. Négrier, J. Savary, K. Pignard, C. Lasset, and B. Escudier (1997). Response rate accuracy in oncology trials: reasons for interobserver variability. groupe français d'immunothérapie of the fédération nationale des centres de lutte contre le cancer. *Journal of Clinical Oncology* 15(12), 3507–3514. PMID: 9396404.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, 24–36.
- Tooze, J. A., G. K. Grunwald, and R. H. Jones (2002). Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research* 11(4), 341–355.
- Tsiatis, A. A., V. Degruittola, and M. S. Wulfsohn (1995). 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.
- Tucker, K., J. Branson, M. Dilleen, S. Hollis, P. Loughlin, M. J. Nixon, and Z. Williams (2016). Protecting patient privacy when sharing patient-level data from clinical trials. *BMC medical research methodology* 16(1), 77.
- Van Cutsem, E., I. Lang, G. D'haens, V. Moiseyenko, J. Zaluski, G. Folprecht, S. Tejpar, O. Kisker, C. Stroh, and P. Rougier (2008). Kras status and efficacy in the first-line treatment of patients with metastatic colorectal cancer (mcr) treated with folfiri with or without cetuximab: The crystal experience. *Journal of Clinical Oncology* 26(15_suppl), 2–2.
- Van Niekerk, J., H. Bakka, and H. Rue (2019). Joint models as latent gaussian models-not reinventing the wheel. *arXiv:1901.09365*.
- Van Niekerk, J., H. Bakka, H. Rue, and L. Schenk (2019). New frontiers in bayesian modeling using the inla package in r. *arXiv:1907.10426*.
- Verbeke, G. (1997). *Linear Mixed Models for Longitudinal Data*, pp. 63–153. New York, NY: Springer New York.
- Verbeke, G. and G. Molenberghs (2000). *How Ignorable Is Missing At Random?*, pp. 375–386. New York, NY: Springer New York.
- Vermorken, J. B., J. Stöhlmacher-Williams, I. Davidenko, L. Licitra, E. Winqvist, C. Villanueva, P. Foa, S. Rottey, K. Skladowski, M. Tahara, et al. (2013). Cisplatin and fluorouracil with or without panitumumab in patients with recurrent or metastatic squamous-cell carcinoma of the head and neck (spectrum): an open-label phase 3 randomised trial. *The lancet oncology* 14(8), 697–710.
- Wang, Q., J. Gao, and X. Wu (2018). Pseudoprogression and hyperprogression after checkpoint blockade. *International immunopharmacology* 58, 125–135.
- Wang, Y. and J. M. G. Taylor (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association* 96(455), 895–905.

- Whiteman, D. C., P. M. Webb, A. C. Green, R. E. Neale, L. Fritschi, C. J. Bain, D. M. Parkin, L. F. Wilson, C. M. Olsen, C. M. Nagle, et al. (2015). Cancers in australia in 2010 attributable to modifiable factors: summary and conclusions. *Australian and New Zealand journal of public health* 39(5), 477–484.
- Wulfsohn, M. S. and A. A. Tsiatis (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* 53(1), 330–339.
- Xing, J., R. E. Myers, X. He, F. Qu, F. Zhou, X. Ma, T. Hyslop, G. Bao, S. Wan, H. Yang, et al. (2011). Gwas-identified colorectal cancer susceptibility locus associates with disease prognosis. *European Journal of Cancer* 47(11), 1699–1707.
- Xu, C., P. Z. Hadjipantelis, and J.-L. Wang (2020). Semi-parametric joint modeling of survival and longitudinal data: The r package jsm. *Journal of Statistical Software* 93(2).
- Zhang, D., M.-H. Chen, J. G. Ibrahim, M. E. Boye, and W. Shen (2016). Jmfit: a sas macro for joint models of longitudinal and survival data. *Journal of statistical software* 71(3).

Appendices

R code for the estimation of the conditional TPJM with frailtypack

The code "TPJM_sim.R" is available at github.com/DenisRustand/TPJM_sim.

```
1 # The following code is decomposed into 3 parts:
2 # 1. Simulation of a dataset assuming a conditional two-part joint model
3 #   (with current-level association structure)
4 # 2. Estimation of the true model as long as a joint naive and left-censored model
5 # 3. Plot of the conditional survival taking into account the effect of treatment
6 #   on the risk of event through the biomarker in addition to the direct treatment
7 #   effect on the risk of event.
8
9 suppressMessages(library(frailtypack))
10 suppressMessages(library(PermAlgo)) # for generation of death times and censoring times
11 suppressMessages(library(mvtnorm)) # multivariate normal generation
12
13 # Method for the numerical approximation of the integral over random-effects
14 methodInt="Monte-carlo" # ("Standard" for standard gauss-hermite quadrature)
15 # association structure for estimation ("Current-level" / "Two-part" / "Random-effects")
16 assoc="Current-level"
17 MAXITER=100 # max iterations for Marquardt algorithm
18 estim_TPJM <- T # Two-part joint model estimation
19 estim_JMn <- F # Naive joint model estimation
20 estim_JMlc <- F # Left-censoring joint model estimation
21 plotsRes=F # plot results (survival conditional on treatment)
22 set.seed(1) # seed for data generation
23 seed_MC=1 # seed for Monte-carlo integration method
24
25 #####
26 # 1 # data simulation
27 #####
28 # Need to sync random number generator because of some changes in R 3.6
29 if("Rounding"%in%RNGkind() | "Rejection"%in%RNGkind()){
30   suppressWarnings(RNGkind(sample.kind = "Rounding"))
31 }
32
33 nsujet=150 # number of individuals
34 numInt=500 # number of integration points
35
36 # binary part fixed effects
37 alpha_0=6 # intercept
38 alpha_1=-4 # slope
39 alpha_2=-1 # baseline treatment
40 alpha_3=1 # treatment X time
41
42 # continuous part fixed effects
43 beta_0=4 # intercept
```

```

44 beta_1=-0.5 # slope
45 beta_2=1 # baseline treatment
46 beta_3=1 # treatment X time
47
48 # survival part
49 gamma_1=0.3 # fixed effect of treatment on the risk of event
50
51 sigma_e=0.5 # error term (gaussian)
52
53 gapLongi=0.07 # gap between longi measurements
54 gap=0.001 # used to generate a lot of time points because the permutation
55 # algorithm choses among those time points to define survival times
56
57 assocCL=0.30 # current-level association between two-part model and survival model
58 kno=5 # knots for splines / baseline hazard
59 followup=4 # follow-up time
60
61 # random effects variance and covariance
62 sigma_b=sqrt(0.5625) # continuous intercept
63 sigma_bt=sqrt(0.5625) # continuous slope
64 sigma_a=sqrt(4) # binary intercept
65 cor_bbt=-0.2 # correlation continuous intercept X slope
66 cor_ba=0.2 # correlation continuous intercept X binary intercept
67 cor_bta=0.7 # correlation continuous slope X binary intercept
68 cov_bbt <- sigma_b*sigma_bt*cor_bbt
69 cov_ba <- sigma_b*sigma_a*cor_ba
70 cov_bta <- sigma_bt*sigma_a*cor_bta
71
72 Sigma=matrix(c(sigma_b^2,cov_bbt,cov_ba,
73               cov_bbt,sigma_bt^2,cov_bta,
74               cov_ba,cov_bta,sigma_a^2),ncol=3,nrow=3)
75
76 fsurv <- Surv(deathTimes, d)~trt # survival model formula
77 flon <- Y~timej*trtY # continuous model formula
78 fbin <- Y~timej*trtY # binary model formula
79
80 mestime=seq(0, followup, gap) # measurement times
81 timej=rep(mestime, nsujet) # time column
82 nmesindiv=followup/gap+1 # number of individual measurements
83
84 nmesy= nmesindiv*nsujet # number of longi measurements
85 idY<-as.factor(rep(1:nsujet, each=nmesindiv)) # id
86
87 ### begin data generation
88 # random effects generation
89 MVnorm <- mvtnorm::rmvnorm(nsujet, rep(0, 3), Sigma)
90
91 a_i = MVnorm[,3] # binary intercept
92 a_iY <- rep(a_i, each=nmesindiv)
93
94 b_i = MVnorm[,1] # continuous intercept
95 b_iY <- rep(b_i, each=nmesindiv)
96 bt_i = MVnorm[,2] # continuous slope
97 bt_iY <- rep(bt_i, each=nmesindiv)
98
99 e_ij = rnorm(nmesy, mean=0, sigma_e) # error
100
101 trt=rbinom(nsujet, 1, 0.5) # treatment covariate
102 trtY=rep(trt, each=nmesindiv)
103
104 ## binary part generation
105 # linear predictor (binary)
106 linPredBin <- alpha_0+a_iY+alpha_1*timej+alpha_2*trtY+alpha_3*timej*trtY
107 probaBin <- exp(linPredBin)/(1+exp(linPredBin)) # proba of zero
108 B <- rbinom(nmesy, 1, probaBin) # zero values (binomial)
109
110 ## generation of longitudinal measurements of outcome
111 # linear predictor (continuous)
112 linPredCont <- beta_0+b_iY+(beta_1+bt_iY)*timej+beta_2*trtY+beta_3*timej*trtY+e_ij
113 # linear predictor (free from error term, for the association)
114 linPredContTrue <- beta_0+b_iY+(beta_1+bt_iY)*timej+beta_2*trtY+beta_3*timej*trtY
115 # in case of negative generated continuous measurements (rarely happening)

```

```

116 linPredContP <- ifelse(linPredCont<(0), 0, linPredCont)
117 linPredContTrueP <- ifelse(linPredCont<(0), 0, linPredContTrue)
118
119 # include zeros in the biomarker distribution
120 Yobs = (ifelse(B==1, linPredContP, 0))
121 Ytrue = (ifelse(B==1, linPredContTrueP, 0))
122
123 #longitudinal dataset
124 id <- as.integer(idY)
125 longDat <- data.frame(id, timej, trtY, Yobs, B, Ytrue)
126
127 # longi measurements to generate survival times with permutation algorithm
128 matX=matrix(ncol=3, nrow=nsujet*nmesindiv)
129 # treatment covariate (to evaluate treatment effect on the risk of event)
130 matX[,1] <- longDat[, "trtY"]
131 # true value of the biomarker (to evaluate effect of the biomarker on the risk of event)
132 matX[,2] <- longDat[, "Ytrue"]
133 # observed value of the biomarker
134 matX[,3] <- longDat[, "Yobs"]
135 eventRandom <- round(rexp(nsujet, 0.0012)+1,0) # ~80% death
136 censorRandom=runif(nsujet,1,nmesindiv) # uniform censoring
137 Ttemp <- permlgorithm(nsujet, nmesindiv, Xmat=matX, eventRandom = eventRandom,
138                       censorRandom=censorRandom, XmatNames=c("trtY", "Ytrue", "Yobs"),
139                       betas=c(gamma_1, assocCL, 0) )
140
141 # extract last line of each individual (= death/censoring time)
142 ligne=NULL
143 for(i in 1:(dim(Ttemp)[1]-1)){
144   if(Ttemp[i, "Id"]!=Ttemp[i+1, "Id"]) ligne <- c(ligne, i)
145 }
146 ligne <-c(ligne, dim(Ttemp)[1])
147
148 Ttemp2=Ttemp[ligne, c("Id", "Event", "Stop", "trtY")] # one line per individual
149 Ttemp2$deathTimes <- mestime[Ttemp2$Stop+1] # deathtimes
150 survDat <- Ttemp2[, c("Id", "deathTimes", "Event", "trtY")] # survival dataset
151 names(survDat) <- c("id", "deathTimes", "d", "trt")
152
153 longDat2 <- Ttemp[,c("Id", "Start", "trtY", "Yobs")]
154 longDat2$timej <- mestime[longDat2$Start+1] # measurements times of the biomarker
155 longDat3 <- longDat2[, c("Id", "timej", "trtY", "Yobs")]
156 names(longDat3) <- c("id", "timej", "trtY", "Y")
157 timesLongi=mestime[which(mestime-round(mestime/gapLongi,0)*gapLongi==0)] # visit times
158 longDat <- longDat3[longDat3$timej%in%timesLongi,]
159
160 survDat$id <- as.integer(survDat$id)
161 longDat$id <- as.integer(longDat$id)
162
163 # Datasets generated are also stored in the frailtypack
164 #load(survDat)
165 #load(longDat)
166
167 ### end data generation
168
169 print(head(longDat, 20))
170 print(head(survDat, 20))
171 print(str(survDat))
172 print(str(longDat))
173 print(summary(survDat))
174 print(summary(longDat))
175
176 #####
177 # 2 # Model estimation
178 #####
179
180 # kappa value (smoothing) chosen by cross-validation with an univariate Cox model
181 tte <- frailtyPenal(fsurv, n.knots=kno, kappa=0, data=survDat, cross.validation = T)
182 kap <- round(tte$kappa, 2); kap # smoothing parameter
183 if(estim_TPJM){ # computation takes ~12min with an Intel i7-4790 (8 cores, 3.60 GHz)
184   TPJM <- longiPenal(fsurv, flon, data=survDat, data.Longi = longDat,
185                     random = c("1", "timej"), formula.Binary=fbin,
186                     random.Binary=c("1"), timevar="timej", id = "id",
187                     link = assoc, n.knots = kno, kappa = kap,

```

```

188         hazard="Splines-per", method.GH=methodInt,
189         n.nodes=numInt, seed.MC=seed_MC);TPJM
190 }
191
192 ## joint naive model
193 if(estim_JMn){
194     JMn <- longiPenal(fsurv, flon, data=survDat, data.Longi = longDat,
195                     random = c("1","timej"), timevar="timej",id = "id",
196                     link = assoc, n.knots = kno,
197                     kappa = kap,hazard="Splines-per",
198                     method.GH=methodInt, n.nodes=numInt, seed.MC=seed_MC);JMn
199 }
200
201 ## joint left-censored model
202 if(estim_JMlc){
203     # censoring threshold (just below smallest observed positive value)
204     TRE <- min(longDat[longDat$Y!=min(longDat$Y),"Y"] -
205              (min(longDat[longDat$Y!=min(longDat$Y),"Y"])/1000)
206     JMlc <- longiPenal(fsurv, flon, data=survDat, data.Longi = longDat,
207                      random = c("1","timej"), timevar="timej", id = "id",
208                      link = assoc, left.censoring = TRE, n.knots = kno,
209                      kappa = kap,hazard="Splines-per",
210                      method.GH=methodInt, n.nodes=numInt, seed.MC=seed_MC);JMlc
211 }
212
213 #####
214 # 3 # plots results (only Two-part model / conditional on treatment arm)
215 #####
216
217 if(plotsRes){
218     # Plot conditional survival from a model estimated with frailtypack
219     # M-splines for the baseline hazard risk and
220     # I-splines for the baseline survival (Ispline=integral(Msplines))
221     # We estimate baseline survival with a numerical approximation
222
223     # load models as R objects
224     #load("~/TPJM.RData")
225     TP=TPJM
226     #-----mspline-----#
227     #' this function generates M_i
228     #' @param x time x for estimation
229     #' @param tp timepoint of length n+k
230     #' @param n.knot total number of knots
231     #' @param k order of the spline function
232     mspline = function(x,tp,n.knot,k=4){
233         if(k==1){
234             region = cbind(tp[1:(length(tp)-1)],tp[2:length(tp)])
235             bool = as.integer(x>region[,1] & x<region[,2])
236             return((1/diff(tp))[as.logical(bool)]*bool)
237         }
238         else{
239             n=length(tp)-k
240             region = cbind(tp[1:n],tp[(k+1):(k+n)])
241             bool = I(x>region[,1] & x<region[,2])
242             M=k*(x-tp[1:n])*(mspline(x,tp,n.knot,k-1)[1:n])+
243                 (tp[(k+1):(k+n)]-x)*(mspline(x,tp,n.knot,k-1)[2:(n+1)])/
244                 ((k-1)*(tp[(k+1):(k+n)]-tp[1:n]))
245             M_final = rep(0,n)
246             M_final[bool]=M[bool]
247             return(M_final)
248         }
249     }
250
251     tpoints=seq(0,max(TP$xD),len=1000) # time points for splines estimation and biomarker values
252     BH=TP$b[1:7]^2 # parameters associated to splines (n.knots+2)
253     M_i=apply(as.matrix(tpoints), 1,mspline,tp=TP$zi,n.knot = 5,k=4) # M-splines
254     hazardEst=t(M_i)%*%as.matrix(BH) # baseline hazard risk
255
256     weights=rep(tpoints[2]-tpoints[1],len=length(tpoints)) # for the integral approximation
257     hCUM=cumsum(hazardEst)*weights # baseline cumulative risk
258     survEst <- exp(-hCUM) # baseline survival
259

```

```

260 # biomarker value
261 coefTP <- TP$coef # model parameters
262 res <- NULL
263 TwoPart <- function(t,trt){
264   BinLinPred <- coefTP[6]+coefTP[7]*t+coefTP[8]*trt+coefTP[9]*t*trt
265   ConLinPred <- coefTP[2]+coefTP[3]*t+coefTP[4]*trt+coefTP[5]*t*trt
266   Prob <- exp(BinLinPred)/(1+exp(BinLinPred))
267   res <- Prob*ConLinPred
268   return(res)
269 }
270
271 survx <- tpoints # abscissas
272 survy <- survEst~exp(TwoPart(survx,0)*TP$eta) # survival (arm A)
273 survytrt <- survEst~exp(coefTP[1]+TwoPart(survx,1)*TP$eta) # (arm B)
274
275 # confidence intervals (Monte-carlo method)
276 nloop=1000 # number of Monte-Carlo curves
277 Hess <- TP$varHIHtotal # Hessian matrix (splines for the baseline hazard included)
278 # generation of the random points
279 isCoef <- mvtnorm::rmvnorm(nloop, TP$b, Hess)
280 mc_BH=isCoef[,1:7]^2 # parameters for the splines (survival)
281 M_i=apply(as.matrix(tpoints), 1,mspline,tp=TP$zi,n.knot = 5,k=4) # M-splines
282 mc_hazardEst=apply(mc_BH,1,function(x) t(M_i)%*%as.matrix(x))# baseline hazard
283 mc_hCUM=apply(mc_hazardEst,2,cumsum)
284
285 mc_hCUMfinal = apply(mc_hCUM,2, function(x) x*weights)
286 mc_survEst <- exp(-mc_hCUMfinal) # baseline survival for all the Monte-carlo curves
287
288 # biomarker
289 survMC=NULL
290 survMctrtr=NULL
291 for(i in 1:nloop){
292   curve_i <- mc_survEst[,i]^exp(TwoPart(survx,0)*isCoef[i,8])
293   curve_itrtr <- mc_survEst[,i]^exp(isCoef[i,16]+TwoPart(survx,1)*isCoef[i,8])
294   survMC <- cbind(survMC, curve_i) # survival (arm A)
295   survMctrtr <- cbind(survMctrtr, curve_itrtr) # survival (arm B)
296 }
297
298 # quantiles
299 QL <- function(x) quantile(x,prob=0.025)
300 QU <- function(x) quantile(x,prob=0.975)
301 SCL <- apply(survMC,1,QL) # ref lower
302 SCU <- apply(survMC,1,QU) # ref upper
303 SCLtrtr <- apply(survMctrtr,1,QL) # trt lower
304 SCUtrtr <- apply(survMctrtr,1,QU) # trt upper
305
306 # plot
307 par(mfrow=c(1,1))
308 plot(survx, survy, lwd=2, xlab="time", ylab="survival", ylim=c(0,1), type='l',
309      main="Survival conditional on treatment arm (TPJM current-level)")
310 lines(survx, survytrt, col='red', lwd=2, lty=2)
311 lines(survx, SCL)
312 lines(survx, SCU)
313 lines(survx, SCLtrtr, col='red', lty=2)
314 lines(survx, SCUtrtr, col='red', lty=2)
315 legend("topright", title = "Treatment", c("arm A", "arm B"), lty=c(1,1),
316       lwd=c(2,2), col=c("black", "red"))
317 }

```

R code for the estimation of the marginal TPJM with **frailtypack**

The code "MTPJM_sim.R" is available at github.com/DenisRustand/TPJM_sim.

```

1 # 1- This code shows how to simulate a dataset assuming a marginal two-part joint model
2 # 2- The estimation of the marginal two-part joint model is then done with frailtypack
3
4 library(frailtypack)
5 library(PermAlgo)
6 library(mvtnorm)

```

```

7
8 #####
9 ### 1 ### Simulation of a dataset
10 #####
11 nsujet=150 # number of individuals
12 # fixed effects of the model
13 ## Binary part
14 alpha_0=6 # intercept
15 alpha_1=-3 # slope
16 alpha_2=1 # baseline treatment
17 alpha_3=-2 # treatment x slope
18 ## Continuous part
19 beta_0=1.5 # intercept
20 beta_1=-0.5 # slope
21 beta_2=0.3 # baseline treatment
22 beta_3=0.3 # treatment x slope
23 sigma_e=0.3 # error term
24 ## Survival part
25 gamma_1=-0.2 # treatment
26
27 gapLongi=0.25 # gap between repeated measurements of the biomarker
28 gap=0.001 # used to generates a lot of biomarker measurements because
29 # the permutation algorithm choses among those measuremnts to define survival time
30
31 assocCL=0.08 # current-level association between two-part model and survival model
32
33 followup=4 # duration of the study
34
35 # random effects variance and covariance
36 sigma_b=sqrt(0.5625) # continuous intercept
37 sigma_bt=sqrt(0.5625) # continuous slope
38 sigma_a=sqrt(4) # binary intercept
39 cor_bbt=-0.2 # correlation continuous intercept X slope
40 cor_ba=0.2 # correlation continuous intercept X binary intercept
41 cor_bta=0.7 # correlation continuous slope X binary intercept
42 cov_bbt <- sigma_b*sigma_bt*cor_bbt
43 cov_ba <- sigma_b*sigma_a*cor_ba
44 cov_bta <- sigma_bt*sigma_a*cor_bta
45
46 Sigma=matrix(c(sigma_b^2,cov_bbt,cov_ba,
47               cov_bbt,sigma_bt^2,cov_bta,
48               cov_ba,cov_bta,sigma_a^2),ncol=3,nrow=3)
49
50 mestime=seq(0,followup,gap) # measurement times
51 timej=rep(mestime, nsujet) # time column
52 nmesindiv=followup/gap+1 # number of individual measurements
53
54 nmesy= nmesindiv*nsujet# number of longi measurements
55 idY<-as.factor(rep(1:nsujet, each=nmesindiv)) # id
56
57 # random effects generation
58 MVnorm <- rmvnorm(nsujet, rep(0, 3), Sigma)
59 a_i = MVnorm[,3] # binary intercept
60 a_iY <- rep(a_i, each=nmesindiv)
61 b_i = MVnorm[,1] # continuous intercept
62 b_iY <- rep(b_i, each=nmesindiv)
63 bt_i = MVnorm[,2] # continuous slope
64 bt_iY <- rep(bt_i, each=nmesindiv)
65
66 e_ij = rnorm(nmesy,mean=0, sigma_e) # error term (continuous part)
67
68 trt=rbinom(nsujet,1, 0.5) # treatment covariate
69 trtY=rep(trt, each=nmesindiv)
70
71 ## binary part generation
72 # linear predictor (binary)
73 linPredBin <- alpha_0+a_iY+alpha_1*timej+alpha_2*trtY+alpha_3*timej*trtY
74 probaBin <- exp(linPredBin)/(1+exp(linPredBin)) # proba of zero
75 B <- rbinom(nmesy,1, probaBin) # zero values (binomial)
76
77 ## generation of longitudinal measurements of outcome
78 # linear predictor (continuous)

```

```

79 linPredCont <- beta_0+b_iY+(beta_1+bt_iY)*timej+beta_2*trtY+beta_3*timej*trtY+e_ij
80 # linear predictor (free from error term, for the association with survival)
81 linPredContTrue <- beta_0+b_iY+(beta_1+bt_iY)*timej+beta_2*trtY+beta_3*timej*trtY
82 # location parameter of the lognormal distribution of positive values
83 mu=linPredCont-log(probaBin)-sigma_e^2/2
84 muTrue=linPredCont-log(probaBin)
85 Y <- rlnorm(length(mu), meanlog = mu, sdlog = sigma_e) # observed positive value
86 Yt <- rlnorm(length(mu), meanlog = muTrue, sdlog = 0) # error-free positive value
87 # include zeros in the biomarker distribution
88 Yobs = (ifelse(B==1, Y, 0))
89 Ytrue = (ifelse(B==1, Yt, 0))
90
91 #longitudinal dataset
92 id <- as.integer(idY)
93 longDat <- data.frame(id, timej, trtY, Yobs, B, Ytrue)
94
95 # longi measurements to generate survival times with permutation algorithm
96 matX=matrix(ncol=3, nrow=nsujet*nmesindiv)
97 # treatment covariate (to evaluate treatment effect on the risk of event)
98 matX[,1] <- longDat[, "trtY"]
99 # true value of the biomarker (to evaluate effect of the biomarker on the risk of event)
100 matX[,2] <- longDat[, "Ytrue"]
101 # observed value of the biomarker
102 matX[,3] <- longDat[, "Yobs"]
103 eventRandom <- round(rexp(nsujet, 0.0012)+1,0) # ~80% death
104 censorRandom=runif(nsujet,1,nmesindiv) # uniform random censoring
105 Ttemp <- permalgorithm(nsujet,nmesindiv,Xmat=matX,eventRandom = eventRandom,
106                       censorRandom=censorRandom,XmatNames=c("trtY", "Ytrue", "Yobs"),
107                       betas=c(gamma_1,assocCL, 0) )
108
109 # extract last line of each individual (= death/censoring time)
110 ligne=NULL
111 for(i in 1:(dim(Ttemp)[1]-1)){
112   if(Ttemp[i,"Id"]!=Ttemp[i+1,"Id"]) ligne <- c(ligne, i)
113 }
114 ligne <-c(ligne, dim(Ttemp)[1])
115
116 Ttemp2=Ttemp[ligne, c("Id","Event","Stop", "trtY")] # one line per individual
117 Ttemp2$deathTimes <- mestime[Ttemp2$Stop+1] # deathtimes
118 survDat <- Ttemp2[, c("Id", "deathTimes", "Event", "trtY")] # survival dataset
119 names(survDat) <- c("id", "deathTimes", "d", "trt")
120
121 longDat2 <- Ttemp[,c("Id", "Start", "trtY", "Yobs")]
122 longDat2$timej <- mestime[longDat2$Start+1] # measurements times of the biomarker
123 longDat3 <- longDat2[, c("Id", "timej", "trtY", "Yobs")]
124 names(longDat3) <- c("id", "timej", "trtY", "Y")
125 timesLongi=mestime[which(round(mestime,3) %in% round(c(seq(0, followup, by=gapLongi),3)))] # visit times
126 longDat <- longDat3[longDat3$timej%in%timesLongi,]
127 survDat$id <- as.integer(survDat$id)
128 longDat$id <- as.integer(longDat$id)
129
130 print(head(longDat, 20))
131 print(head(survDat, 20))
132 print(str(survDat))
133 print(str(longDat))
134 print(summary(survDat))
135 print(summary(longDat))
136
137 #####
138 ## 2 ## Estimation of the marginal two-part joint model
139 #####
140 numInt=500 # number of integration points
141 fsurv <- Surv(deathTimes, d)~trt # survival model formula
142 flon <- Y~timej*trtY # continuous model formula
143 fbin <- Y~timej*trtY # binary model formula
144 # kappa value (smoothing) chosen by cross-validation
145 tte <- frailtyPenal(fsurv,
146                   n.knots=5,kappa=0, data=survDat,cross.validation = T)
147 kap <- round(tte$kappa,2)
148
149 MTPJM <- longiPenal(fsurv, flon, data=survDat,data.Longi = longDat, random = c("1", "timej"),
150                   formula.Binary=fbin, random.Binary=c("1"),

```



```

151     GLMlog=T, # logarithm link for the distribution of positive values
152     MTP=T, # Trigger for marginal two-part model (set to FALSE for a conditional two-part model)
153     timevar="timej",id = "id", link = "Current-level", left.censoring = F,seed.MC=1,
154     n.knots = 5, kappa = kap,hazard="Splines-per",maxit=200,
155     method.GH="Monte-carlo", n.nodes=numInt    )
156 print(MTPJM)

```

R code for the estimation of the conditional TPJM with R-INLA

The code "TPJM_INLA.R" is available at github.com/DenisRustand/TPJM_sim.

```

1
2 # 1- This code shows how to simulate a dataset assuming a conditional two-part joint model
3 # 2- The estimation of the conditional two-part joint model is then done with INLA
4
5 set.seed(1)
6 library(INLA)
7 inla.setOption(mkl=TRUE)
8
9 #####
10 ### 1 ### Simulation of a dataset
11 #####
12
13 library(mvtnorm) # for multivariate normal generation (random-effects)
14 nsujet=200 # number of individuals
15 #binary part
16 alpha_0=4 # Intercept
17 alpha_1=-0.5 # slope
18 alpha_2=-0.5 # treatment
19 alpha_3=0.5 # treatment x time
20 #continuous part
21 beta_0=2 # Intercept
22 beta_1=-0.3 # slope
23 beta_2=-0.3 # treatment
24 beta_3=0.3 # treatment x time
25 sigma_e=0.3 # error term (standard error)
26 gamma_1=0.2 # treatmentt effect on survival
27 # Shared random effects association between the two-part model for the biomarker and survival
28 phi_a=1 # random intercept (binary)
29 phi_b=1 # random intercept (continuous)
30 phi_bt=1 # random slope (continuous)
31 # baseline hazard scale (to generate exponential death times)
32 baseScale=0.2
33 gap=0.4 # gap between longitudinal repeated measurements
34 followup=4 # study duration
35 # correlated random-effects
36 sigma_a=1 # random intercept (binary)
37 sigma_b=0.5 # random intercept (continuous)
38 sigma_bt=0.5 # random slope (continuous)
39 cor_ba=0.5 # correlation intercept (binary)/intercept (continuous)
40 cor_bta=0.5 # correlation intercept (binary)/slope (continuous)
41 cor_bbt=-0.2 # correlation continuous intercept/slope
42 cov_ba <- sigma_b*sigma_a*cor_ba # covariance
43 cov_bta <- sigma_bt*sigma_a*cor_bta
44 cov_bbt <- sigma_b*sigma_bt*cor_bbt
45 Sigma=matrix(c(sigma_a^2,cov_ba,cov_bta, # variance-covariance matrix
46               cov_ba,sigma_b^2,cov_bbt,
47               cov_bta,cov_bbt,sigma_bt^2),ncol=3,nrow=3)
48 mestime=seq(0,followup,gap) # measurement times
49 timej=rep(mestime, nsujet) # time column
50 nmesindiv=followup/gap+1 # number of individual measurements
51 nmesy= nmesindiv*nsujet # number of longi measurements
52 id<-as.factor(rep(1:nsujet, each=nmesindiv)) # patient id
53 # random effects generation
54 MVnorm <- mvtnorm::rmvnorm(nsujet, rep(0, 3), Sigma)
55 a_i = MVnorm[,1] # binary intercept
56 a_iY <- rep(a_i, each=nmesindiv) # binary intercept (repeated for longi dataset)
57 b_i = MVnorm[,2] # continuous intercept
58 b_iY <- rep(b_i, each=nmesindiv)

```

```

59 bt_i = MVnorm[,3] # continuous slope
60 bt_iY <- rep(bt_i, each=nmesindiv)
61
62 treated <- sample(1:nsujet, nsujet/2, replace=F)
63 treatedFull <- NULL
64 for(i in 1:nsujet){
65   treatedFull <- c(treatedFull, ifelse(i%in%treated, 1, 0))
66 }
67 trt= treatedFull# treatment covariate
68 trtY=rep(trt, each=nmesindiv)
69
70 ## linear predictor (binary part)
71 linPredBin <- alpha_0+a_iY+alpha_1*timej+alpha_2*trtY+alpha_3*timej*trtY
72 probaBin <- exp(linPredBin)/(1+exp(linPredBin)) # proba of positive value
73 B <- rbinom(nmesy,1, probaBin) # observed zero values
74
75 ## linear predictor (continuous part)
76 linPredCont <- beta_0+b_iY+(beta_1+bt_iY)*timej+beta_2*trtY+beta_3*timej*trtY
77 mu=linPredCont-sigma_e^2/2 # lognormal mean
78 Ypos <- rlnorm(length(mu), meanlog = mu, sdlog = sigma_e) # observed biomarker values
79 Y = (ifelse(B==1, Ypos, 0)) # include zeros in the biomarker distribution
80
81 ## longitudinal biomarker dataset
82 longDat <- data.frame(id, timej, trtY, Y)
83
84 ## generation of exponential death times
85 u <- runif(nsujet) # uniform distribution for survival times generation
86 deathTimes <- -(log(u) / (baseScale * exp(trt * gamma_1 + a_i*phi_a + b_i*phi_b + bt_i*phi_bt)))
87 d <- as.numeric(deathTimes<followup) # deathtimes indicator
88 ## censoring individuals at end of follow-up (not at random)
89 deathTimes[deathTimes>=followup]=followup
90 ids <- as.factor(1:nsujet)
91 survDat <- data.frame(id=ids,deathTimes, d, trt) # survival times dataset
92
93 ## removing longi measurements after death
94 ind <- rep(NA, nsujet*length(mestime))
95 for (i in 1:nsujet){
96   for(j in 1:length(mestime)){
97     if(longDat[(i-1)*length(mestime)+j, "timej"]<=survDat[i,"deathTimes"]) ind[(i-1)*length(mestime)+j
98     ]=1
99   }
100 }
101 longDat <- longDat[!is.na(ind),]
102 survDat$trt <- as.factor(survDat$trt)
103 longDat$trtY <- as.factor(longDat$trtY)
104 longDat$id <- as.integer(longDat$id)
105 survDat$id <- as.integer(survDat$id)
106 ## Summary of the longitudinal and survival datasets
107 print(summary(survDat))
108 print(summary(longDat))
109
110 #####
111 ### 2 ### Estimation of a conditional two-part joint model with R-INLA
112 #####
113 # create dataset with positive values only for the continuous part
114 longDatlog <- longDat[longDat$Y>0,]
115 nB <- length(longDat$Y) # length of binary part
116 nC <- length(longDatlog$Y) # length of continuous part
117 ns=dim(survDat)[1] # number of individuals
118
119 longDat$B <- ifelse(longDat$Y==0,0,1) # zero value indicator (binary part outcome)
120 longDatlog$sld <- longDatlog$Y # positive values only (continuous part outcome)
121 yy <- matrix(NA, ncol = 2, nrow = nB+nC)
122 yy[1:nB,1] <- longDat$B # binary outcome
123 yy[nB+(1:nC),2] <- longDatlog$Y # continuous outcome
124 yB = yy[,1]
125 yC = yy[,2]
126
127 #####Add all survival covariates
128 # set up unique identifiers for the random-effects
129 longDatlog$idl <- ns+as.integer(longDatlog$id)

```

```

130 longDatlog$idl2 <- ns+ns+as.integer(longDatlog$id)
131 survDat$id1 <- ns+as.integer(survDat$id)
132 survDat$idl2 <- ns+ns+as.integer(survDat$id)
133
134 cox_TRTs = as.factor(c(ifelse(survDat$trt=="0", "ref", "trt")))
135 cox_IntS = rep(1, length(cox_TRTs))
136 surv_inla_obj = inla.surv(time=survDat$deathTimes,event = survDat$d)
137 cox_ext = inla.coxph(surv_inla_obj ~ 1+TRTs,
138                   control.hazard=list(model="rw2",
139                                       scale.model=TRUE,
140                                       diagonal=1e-4,
141                                       constr=F,
142                                       hyper=list(prec=list(prior="pc.prec",
143                                                         param=c(1,0.01))),
144                   data = c(list(surv_inla_obj = surv_inla_obj,
145                                 TRTs = cox_TRTs,
146                                 IDs = survDat$id1,
147                                 IDsb = as.integer(survDat$id),
148                                 IDs2 = survDat$idl2), as.list(survDat)))
149 ns_cox = dim(cox_ext$data)[1] # for extended dataframe for poisson regression
150
151 ###For other parts without survival part
152 # fixed effects
153 linear.covariate <- data.frame(
154   InteB = c(rep(1,nB), rep(0,nC)), # intercept (binary part)
155   InteC = c(rep(0,nB), rep(1,nC)), # intercept (continuous part)
156   TIME = c(rep(0,nB),longDatlog$timej), # time (continuous part)
157   TIMEb = c(longDat$timej,rep(0,nC)), # time (binary part)
158   TRTc = c(rep(0,nB),as.numeric(longDatlog$strty)-1), # treatment (continuous)
159   TRTb = c(as.numeric(longDat$strty)-1,rep(0,nC)) # treatment (binary)
160 # random-effects
161 random.covariate<-list(ID1=c(rep(NA,nB),longDatlog$id1), # random intercept (continuous)
162                       IDb=c(as.integer(longDat$id),rep(NA,nC)), # random intercept (binary)
163                       ID12=c(rep(NA,nB),as.integer(longDatlog$idl2)), # random slope (continuous)
164                       slopeCont=c(rep(NA,nB),longDatlog$timej)) # weight for random slope (continuous
165 )
166 jointdf = data.frame(linear.covariate, random.covariate, yB, yC)
167 joint.data_cox <- c(as.list(inla.rbind.data.frames(jointdf, cox_ext$data)),
168                   cox_ext$data.list)
169 Yjoint = cbind(joint.data_cox$yB, joint.data_cox$yC, joint.data_cox$y..coxph) # outcomes
170 joint.data_cox$Y <- Yjoint
171
172 # conditional two-part joint model formula - update from the cox expansion
173 formulaJ= update(cox_ext$formula, Yjoint ~ . + InteB+InteC + TIME*TRTc+TIMEb*TRTb+
174                 f(IDb, model="iid3d", n=3*ns,constr=F)+
175                 f(ID1, copy="IDb")+
176                 f(ID12, slopeCont,copy="IDb")+
177                 f(IDsb, copy="IDb",fixed=F)+
178                 f(IDs, copy="ID1",fixed=F)+
179                 f(IDs2, copy="ID12",fixed=F))
180 #Fit model with INLA ()
181 TPinla <- inla(formulaJ,family = c("binomial", "gamma", cox_ext$family),
182              data=joint.data_cox,
183              Ntrials=c(rep(1,length(longDat$Y)),rep(NA,nC),rep(NA,ns_cox)),
184              control.predictor=list(compute=TRUE,link=1),#error gaussian
185              E = joint.data_cox$E..coxph,
186              control.family=list(list(control.link = list(model = "logit")),
187                                list(link="log",hyper = list(prec = list(initial = 2, fixed=FALSE))
188                                ),
189                                list()),#variant = 1
190              control.inla = list(strategy="adaptive"),
191              control.fixed=list(remove.names="(Intercept)"),
192              verbose=F)
193 print(summary(TPinla))

```

Abstract

Assessing the effectiveness of cancer treatments in clinical trials raises multiple methodological problems that need to be properly addressed in order to produce a reliable estimate of treatment effects. The purpose of this research project is to propose a new modeling strategy within the joint modeling framework to study simultaneously the evolution of tumor size (biomarker) and the risk of death (terminal event). An excess of zero values characterize the distribution of the tumor size measurements, corresponding to patients responding well to a treatment that observe a complete shrinkage of their tumors. The two-part model has been proposed with the idea to decompose the distribution of the biomarker into a binary outcome (zero values vs. positive values) and a continuous outcome, both outcomes usually being modeled with mixed effects regression models. We developed a two-part joint model for which the binary part captures the effect of covariates on the probability of zero value of the biomarker while the continuous part gives the effect of covariates either on the expected value of the biomarker among positives (conditional form) or the marginal expected value of the biomarker (marginal form), both answering different clinical questions of interest. We established it provides unbiased parameter estimations by simulations and compared this new model with alternative approaches such as ignoring the zero excess by not decomposing the biomarker's distribution or considering zeros as censored values (i.e., too small to be measured). We show how the two-part approach is more appropriate in presence of true zeros (i.e., not censored). This new model allows to use both the tumor size repeated measurements and the survival times to compare several treatment lines, which could impact the final clinical decisions. We illustrated these developments on the basis of real data from randomized cancer clinical trials. Finally, we extended the frequentist estimation that we implemented into the R package **frailtypack** to a Bayesian framework within the R package **INLA** in order to reduce the computation time and solve convergence issues when dealing with more complex correlation structures. The software and code for both the frequentist and Bayesian estimations of this new model are freely available to ensure that these tools are easily disseminated to epidemiologists, statisticians or biomedical researchers. Semicontinuous distributions are common in biomedical research, e.g., when quantifying exposure or measuring symptoms of a disease, in genomics (microbiome, epigenetics), so that the proposed work could lead to a wide spectrum of applications beyond cancer research.

Key words: cancer; clinical trial; joint model; longitudinal analysis; semicontinuous distribution; survival analysis; tumor response; two-part model.

Résumé

Évaluer l'efficacité des traitements dans les essais cliniques en oncologie soulève de multiples problèmes méthodologiques qui doivent être correctement traités afin de produire une estimation fiable des effets du traitement. Le but de ce projet de recherche est de proposer une nouvelle stratégie de modélisation dans le cadre de la modélisation conjointe pour étudier simultanément l'évolution de la taille tumorale (biomarqueur) et le risque de décès (événement terminal). Un excès de zéros caractérise la distribution des mesures de taille tumorale, correspondant à des patients ayant une réponse au traitement qui se traduit par la disparition des tumeurs. Le modèle two-part a été proposé avec l'idée de décomposer la distribution du biomarqueur en une partie binaire (zéros vs. valeurs positives) et une partie continue, les deux étant généralement modélisés avec des modèles de régression à effets mixtes. Nous avons développé un modèle conjoint two-part pour lequel la partie binaire donne l'effet de covariables sur la probabilité de valeur nulle du biomarqueur tandis que la partie continue donne l'effet de covariables soit sur la valeur du biomarqueur parmi les positifs (forme conditionnelle) ou la valeur marginale du biomarqueur (forme marginale), tous deux répondant à différentes questions cliniques d'intérêt. Nous avons établi à l'aide de simulations que ce modèle fournit des estimations de paramètres non biaisées et nous l'avons comparé avec des approches alternatives telles qu'ignorer l'excès de zéro en ne décomposant pas la distribution du biomarqueur ou considérer les zéros comme des valeurs censurées (i.e., trop petites pour être mesurées). Nous montrons comment l'approche two-part est plus appropriée en présence de vrais zéros (i.e., non censurés). Ce nouveau modèle permet d'utiliser à la fois les mesures répétées de taille tumorale et les temps de survie pour comparer plusieurs lignes de traitement, ce qui pourrait impacter les décisions cliniques finales. Nous avons illustré ces développements sur la base de données réelles issues d'essais cliniques randomisés en cancérologie. Enfin, nous avons étendu l'estimation fréquentiste que nous avons implémentée dans le package R **frailtypack** à un cadre Bayésien avec le package R **INLA** afin de réduire le temps de calcul et résoudre les problèmes de convergence observés pour des structures de corrélation plus complexes. Les logiciels et codes pour l'estimation fréquentiste et Bayésienne de ce nouveau modèle sont publiquement disponibles pour s'assurer que ces outils sont facilement diffusés aux épidémiologistes, aux statisticiens ou chercheurs en sciences biomédicales. Les distributions semi-continues sont courantes dans la recherche biomédicale, par exemple lorsque l'on quantifie une exposition ou mesure les symptômes d'une maladie, notamment en génomique (microbiome, épigénétique), de sorte que le modèle proposé pourrait ouvrir un large spectre d'applications au-delà de la recherche relative au cancer.

Mots-clés : analyse de survie; analyse longitudinale; cancer; distribution semi-continue; essai clinique; modèle conjoint; modèle two-part; réponse tumorale.