Développement de méthodes pour la validation de critères de substitution en survie : méta-analyses de cancer

Casimir Sofeu

université de BORDEAUX

THÈSE PRÉSENTÉE
POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

.....

École doctorale Sociétés, Politique, Santé Publique
Spécialité Santé Publique, Option Biostatistique
Thèse co-financée par l'INSERM et la Région d'Aquitaine

Par Casimir Ledoux SOFEU

Développement de méthodes pour la validation des critères de substitution en survie : méta-analyses de cancer.

.....

Sous la direction de Virginie Rondeau

Soutenue le 12 décembre 2019

Membres du jury :

Mme	JACQMIN-GADDA Hélène	Dr, INSERM U1219, Bordeaux	Présidente
Mme	LEGRAND Catherine	Pr, UCL, Louvain-la-Neuve	Rapporteure
M.	PAOLETTI Xavier	Pr, Institut Curie, Paris	Rapporteur
Mme	BELLERA Carine	Dr, INSERM U1219, Bordeaux	Examinatrice
M.	GIORGI Roch	Pr, SESSTIM, Marseille	Examinateur
Mme	RONDEAU Virginie	Dr, INSERM U1219, Bordeaux	Directrice de thèse

Titre : Développement de méthodes pour la validation de critères de substitution en survie : méta-analyses de cancer.

Résumé : Les critères de substitution peuvent être utilisés à la place du critère de jugement le plus pertinent pour évaluer l'efficacité d'un nouveau traitement. Dans un contexte de méta-analyse, l'approche classique pour la validation d'un critère de substitution est basée sur une stratégie d'analyse en deux étapes. Pour des critères de jugement à temps d'évènements, cette approche est souvent sujette à des problèmes d'estimations. Nous proposons une approche de validation en une étape s'appuyant sur des modèles conjoints à fragilités et à copules. Ces modèles incluent à la fois des effets aléatoires au niveau essai et au niveau individuel ou des fonctions de copule. Nous considérons des fonctions de risque de base non paramétriques à l'aide des splines. Les paramètres des modèles et les fonctions de risque de base ont été estimés par une méthode semi-paramétrique, par maximisation de la vraisemblance marginale pénalisée, considérant différentes méthodes d'intégration numérique. La validation des critères de substitution à la fois au niveau individuel et au niveau essai a été faite à partir du tau de Kendall et du coefficient de détermination. Les études de simulation ont été faites pour évaluer les performances de nos modèles. Les modèles ont été appliqués aux données individuelles issues des méta-analyses sur le cancer afin de rechercher de potentiels critères de substitution à la survie globale. Les modèles étaient assez robustes avec réduction des problèmes de convergence et d'estimation rencontrés dans l'approche en deux étapes. Nous avons développé un package `R` convivial implémentant les nouveaux modèles.

Mots clés : cancer, copules, critères de substitution, intégration numérique, méta-analyse d'essais cliniques, modèle conjoint à fragilités

---

Title : Development of methods for the validation of time-to-event surrogate endpoints: meta-analysis of cancer

Abstract : Surrogate endpoint can be used instead of the most relevant clinical endpoint to assess the efficiency of a new treatment. In a meta-analysis framework, the classical approach for the validation of surrogate endpoint is based on a two-step analysis. For failure time endpoints, this approach often raises estimation issues. We propose a one-step validation approach based on a joint frailty and a joint frailty-copula model. The models include both trial-level and individual-level random effects or copula functions. We chose a non-parametric form of the baseline hazard functions using splines. We estimated parameters and hazard functions using a semi-parametric penalized marginal likelihood method, considering various numerical integration methods. Both individual level and trial level surrogacy were evaluated using Kendall's τ and coefficient of determination. The performance of the estimators was evaluated using simulation studies. The models were applied to individual patient data meta-analyses in cancer clinical trials for assesing potentiel surrogate endpoint to overall survival. The models were quite robust with a reduction of convergence and model estimation issues encountered in the two-step approach. We developed a user friendly `R` package implementing the models.

Keywords : cancer, copulas, joint frailty model, meta-analysis of clinical trials, numerical integration, surrogate endpoint

---

INSERM U1219 (Biostatistique), ISPED, Université de Bordeaux, 146 rue Léo Saignat, 33076 Bordeaux, France

Remerciements

Je remercie ma Directrice de Thèse, Dr Virginie Rondeau pour sa confiance et sa disponibilité durant ces trois années de thèse. Tu as été une Directrice de thèse sans pareil, ce qui m'a permis d'apprendre énormément de choses sur les plans scientifique, professionnel et même personnel. Ton côté affectueux m'a réellement marqué car à chaque fois que j'ai eu un problème personnel ou familial, tu as toujours été là pour me rassurer que tout était rentré dans l'ordre. Ce qui m'a permis effectivement de gagner en confiance dans cet environnement de travail qui attribue une place importante à l'être humain. Merci pour nos réunions hebdomadaires, qui sans être très longues étaient toujours efficaces et me donnaient chaque fois l'envie d'aller plus loin dans la recherche. Tellement j'ai appris de toi que j'espère que nous continuerons cette collaboration.

Je remercie l'INSERM et la région Nouvelle Aquitaine, l'ARC et l'INCA pour le financement de mes travaux de thèses.

Je remercie le Dr Takeshi Emura pour la collaboration que nous avons développée durant cette thèse. J'ai beaucoup appris de ton expérience sur les modèles de copules, surtout durant le séjour de recherche que j'ai passé avec toi à Taiwan. Cela a été une expérience formidable.

Je remercie mes rapporteurs de thèse, le Prof Catherine Legrand et Prof Xavier Paoletti pour avoir accepté d'évaluer ce travail. Votre rigueur et la pertinence de vos observations m'ont permis d'améliorer la qualité de ce travail.

Au Dr Carine Bellera, merci d'avoir accepté d'examiner ce travail.

Au Dr Hélène Jacqmin-Gadda, je te remercie pour ces petits moments d'échanges et pour avoir accepté d'examiner ce travail. Merci pour ce que tu nous apporte en tant que chef de l'équipe Biostatistique. Même si je n'ai pas souvent eu l'occasion de travailler avec toi, j'ai beaucoup appris de tes interventions lors des séminaires de l'équipe. J'espère pouvoir collaborer avec toi à l'avenir.

Au professeur Roch Giorgi, je ne saurais comment te remercier pour la confiance que tu as mise en moi depuis nos premiers échanges lorsqu'il fallait que je rédige mon projet de stage de Master 2 MQERS dont tu as la responsabilité. Grâce à toi j'ai pu réaliser mon rêve de poursuivre mes études en France et me voici aujourd'hui Docteur. Merci pour tout ce que tu m'as apporté notamment la qualité de tes enseignements, ta rigueur et ton sens du travail bien fait, l'encadrement durant le stage de Master et tes encouragements pour mon projet de thèse. Merci d'avoir accepté d'examiner ce travail. J'espère que nous aurons l'opportunité de collaborer davantage à l'avenir.

Au Dr Mathurin Cyrille Tejiokem je te remercie pour les 5 années que nous avons passées ensemble au sein du service d'épidémiologie et de santé publique du Centre Pasteur du Cameroun. Tu as été un chef formidable, bien que ta rigueur dans le travail été parfois incomprise

par tes collaborateurs dont je faisais partie. Tu m'as donné l'envie de faire de la Biostatistique et tu m'as beaucoup encouragé dans ce projet qui qui trouve son aboutissement aujourd'hui. J'espère pouvoir collaborer davantage avec toi sur les aspects plus techniques de la cohorte ANRS PEDIACAM dont tu es le coordinateur.

A toute la famille BIOSTAT-SISTM de l'ISPED, merci pour ces multiples séminaires et réunions durant lesquels j'ai beaucoup appris. Merci à Rodolphe et à Pierre pour m'avoir accepté dans vos cours de M2 Biostatistique. A Cécile, merci pour les échanges lors des séminaires (j'espère que j'aurai l'opportunité de mieux profiter de ton expérience à l'avenir sur des questions méthodologiques). Alioum, avec Pierre vous m'avez permis de faire cette expérience d'enseignement durant ces deux dernières années avec l'ISPED, merci pour cela. Daniel, je n'ai réellement pas eu l'occasion d'échanger avec toi pendant ma thèse, toutefois, je retiens que tu es la personne à qui revient l'honneur de la fondation de l'équipe Biostatistique, et donc de façon indirecte tu as également été là pour ma thèse et je t'en remercie.

A tous mes compagnons du RU le midi, merci pour ces moments de partage même si je n'ai pas toujours été avec vous cette dernière année: Corentin (oui CS11, ces trois années passées avec vous resteront à jamais graver dans ma mémoire), Camille (CS9 tout comme avec CS11, nous formions un bon trio avec des échanges toujours plaisants), Maude, Jocelyn le tout nouveau post doc très sympa (tu a assuré avec enthousiasme mes déplacements avec ma fille lors de petites sorties chez les copains), Anaïs, Sophie, Perrine, Virginie, Aline, Laura.

Les collègues de bureau: Viviane (Tu étais toujours prête à échanger avec moi sur mes questions de code, même lorsqu'à priori tu n'avais aucune solution. Merci pour ta sympathie, sans oublier le vélo du samedi; ces moments ont été formidables), Loic (dans mes pires moments où je me demandais comment faire du bootstrap avec mes modèles gourmands en temps de calcul pour l'estimation des intervalles de confiance du τ de Kendall dans les simulations, tu m'as permis de me rendre compte que je pouvais faire du bootstrap paramétrique :) I was very happy for that), Myriam (Merci pour l'initiation à frailtypack), Emilie pour tes encouragements et tes précieux conseils qui à chaque fois me permettaient de voir la vie du bon côté (tu es très galvanisante) merci aussi pour la relecture de ma thèse, Anthony, Tiphaine, Kateline notre jeune stagiaire très pointue dans son travail. Agnieszka pour m'avoir aidé dans le déchiffrage de l'existant dans frailtypack. Bachiro pour toutes ces longs moments d'échanges que ce soit sur le plan professionnel, académique ou familial; tu es une personne aimable aux cotés de qui j'ai beaucoup profité, surtout dans les pires moments où j'avais du mal à faire converger mes modèles.

Autres collègues sympa de l'ISPED à qui j'adresse mes remerciements: Soufiane et sa bonne humeur (je te dirais quand le papier de Plos One sera publié ou accepté, tu mérites ce retour de par ta curiosité galvanisante, tu procures de la bonne humeur et du courage), Aaron, Adrien

(tu as pu me faire partir en montagne) Thierry (pour nos longs échanges sur les détails administratifs et le partage d'expérience), Robin pour tes orientations dans le développement du package et l'utilisation de Git Hub (tu étais toujours prêt à me recevoir dans ton bureau à chaque fois que j'avais mes soucis de codes, tu es vraiment génial), Antoine (rencontré à peine arrivé à l'ISPED, tu as été là quand j'avais le plus besoin de ton aide, tu m'as permis de comprendre l'approche Bayesienne et avec toi j'ai pu implémenter l'approche de Renfro et al. 2012, afin de répondre à une question des reviewers de mon article dans Stat Med, je t'ai d'ailleurs remercié à ce sujet dans mon second papier soumis dans Plos One. Ta pédagogie et ton intégration n'ont été qu'à mon avantage, Boris pour m'avoir accepté dans ton cours de M2 sur le Bayésien et aussi pour ces quelques moments d'échanges que nous avons eu dans mes premières années à l'ISPED. Bruno et ses cheveux (oui nous avons eu à prendre quelques verres ensemble), Marta pour ces deux années que nous avons passé ensemble à construire le cours sur R commander (Merci pour ta confiance), Fleu, Valerie, Frantz et Gayo pour l'accompagnement dans les cours d'informatique que j'ai dispensé en L2 Médecine, l'équipe du CREDIM et Curta pour la logistique sur le plan informatique, et toutes les personnes que j'ai pu croiser dans les couloirs de l'ISPED durant ces trois belles années. Sébastien et Denis merci pour ces moments d'échanges.

J'exprime ma gratitude à l'endroit de Sébastien et Louise de l'ex-Capionis, vous m'avez permis de faire mes premier pas dans le privé en France, avec cette mission d'expertise durant ma première année de thèse. Merci pour votre confiance. J'espère avoir l'occasion de vous côtoyer de nouveau, cette fois sur un plan purement professionnel.

Côté logistique le bureau d'accueil des chercheurs internationaux, particulièrement Nicolas et Frédérique avec qui j'ai constamment été en contact depuis le début de ma thèse. Oui, merci Sandrine et Ludivine pour tout ce que vous faites pour les nouveaux arrivants et le suivi que vous nous apportez tout au long de la thèse. Merci enfin au personnel du service des ressources humaines de l'INSERM Nathalie et Malika pour avoir toujours été là quand j'avais besoin de vous.

Toutes les personnes que j'ai eu la chance de côtoyer au sein du SESSTIM à Marseille (Célia, Juste et Andréa, Camelia, Toussaint, Guindo, Sokhna, Yasouko, Mme Le Noah, Farida) ainsi que l'équipe enseignante du Master de santé publique spécialité MQERS de l'Université d'Aix-Marseille.

Tout le personnel du service d'épidémiologie du CPC ainsi que du projet ANRS-PEDIACAM : Pascal, Gaëtan, Albert, Josiane, Ida, Suzy, Ange, Tatiana, Félicité, Pascaline, Etienne, Angèle, William, Paul Alain, Robert, Francis Youya, Francis Ateba, Audrey, Verlaine, Francine, Derboise, Kelly)

Mes connaissances en France (Christine, Fabrice, Esper, Sam), j'ai passé de bon moments

à chaque vois que j'ai été en votre compagnie. Un merci spécial à Christine pour son accueil à Marseille et tous ces beaux moments que nous avons passé ensemble, tu as été plus qu'une mère pour moi à cette époque.

A mes amis, j'avoue que la liste est longue mais je vais quand même prendre le risque de citer quelques-uns au risque d'offenser ceux qui ne trouverons pas leur noms marqués (croyez-moi je vous porte tous dans mon coeur) : Romual (tu as été plus qu'un soutien pour moi jeune homme depuis nos moments au lycée), Benjamin, Roméo, Roméo, Fabrice, Brillant, Mado, Sophie [Jennifer], Césaire, Louis, Aboubakar, Gérard, Nicaire, Massongo, Gaetan, Patrice, Bruno, Kalina.

Tous les compagnons du 124 avec qui j'ai partagé de beaux moments autour du poulet du dimanche chez notre parrain Jean François Tessier (JFT). Oui mes dimanches n'auraient pas été pareils sans vous, sans nos fous rires. Mes remerciements vont en particulier à JFT pour son accueil et sa générosité envers les étrangers. Cet ami m'a réellement marqué pendant mon séjour à Bordeaux. Merci pour la relecture de ma thèse tu es très génial.

A ma famille auprès de qui je témoigne également ma gratitude, Désolé de n'avoir pas toujours été là depuis maintenant 34 ans. Mon cher frère, ami et père Bernard, le tout puissant t'a appelé à lui très tôt. J'aurai aimé fêter ces moments avec toi mais le destin n'a pas voulu. Que la terre de nos aïeux te soit légère mon bon ami. A mes frères aînés Alain et Grégoire, vous avez été comme des pères pour moi et je vous en suis reconnaissant. A tous mes aînés Joséphine, Rogers, Fernand, Hortensine, Barthélémy, Edmond, Jean, Joichim, Chancelle, et mes cadets Sidonie, Miguel, Ulrich, Jordan, Yann. Aux plus petits, puisse le tout puissant continuer de vous bénir et de frayer pour vous le meilleur chemin vers la réussite. Merci à Alain et Miguel pour la relecture de ma thèse.

A mes beaux-parents Rosalie et Richard, merci pour vos encouragements et surtout pour la confiance que vous avez mise en moi. Grâce à vous, je me réveille chaque matin avec beaucoup de forces pour affronter les difficultés de la vie. A mes beaux-frères Syntich, Boris, Descartes et Duplex, merci pour votre confiance et désolé de n'avoir pas toujours été là.

Oui !!!*on garde toujours le meilleur pour la fin (ce qui n'a rien d'exclusif)*. Mon amour, ma moitié, ces quelques mots pour te remercier de m'avoir supporté tout ce temps. Dire que depuis notre union nous n'avons pas réellement profité de notre vie de couple car il fallait couronner 5 années d'études pour voir arriver ce moment, et tu as su me supporter pendant ce temps. Oui mes mauvais humeurs, surtout lorsque mes modèles ne convergeaient pas et quand je devais avoir réunion le lendemain!!! Merci pour ton soutien indéfectible Idene, merci pour les filles. Ce présent marque la fin du premier round dans le parcours du combattant. J'espère que nous pourrons enfin profiter de la vie.

Dédicaces

A mes parents maman Rose et Papa Gilles,
à mes filles Tiphaine Andrea et Aurore Alexandra

Valorisation scientifique

Articles

Publications issues de la thèse

- C. L. Sofeu, T. Emura, and V. Rondeau. A joint frailty-copula model for meta-analytic validation of failure time surrogate endpoints in clinical trials. *Submitted*.

- C. L. Sofeu and V. Rondeau. How to use frailtypack for validating failure-time surrogate endpoints using individual patient data from meta-analyses of randomized controlled trials. *Under review*.

- C. L. Sofeu, T. Emura, and V. Rondeau. 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, 2019. doi: 10.1002/sim.8162. URL-https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8162.

Autres publications durant la thèse

- Sofeu CL, Tejiokem MC, Penda CI, Protopopescu C, Ateba Ndongo F, Tetang Ndiang S, Guemkam G, Warszawski J, Faye A, Giorgi R; ANRS-PEDIACAM study group. "Early treated HIV-infected children remain at riskof growth retardation during the first five years of live: results from the ANRS-PEDIACAM cohort inCameroon." *PLoS One*, 2019 July, 14(7): e0219960.

- Penda CI, Tejiokem MC, Sofeu CL, Ndiang ST, Ateba Ndongo F, Kfutwah A, Guemkam G, Warsza-wski J, Faye A. "Low rate of early vertical transmission of HIV supports the feasibility of effective implementation of the national PMTCT guidelines in routine practice of referral hospitals in Cameroon." *Paediatr Int Child Health.* 2019;39(3)208-215.

- Sofeu CL, Broban A, Njifou Njimah A et al. "Improving systematic rabies surveillance in Cameroon: A pilot initiative and results for 2014-2016." *PLoS Negl Trop Dis.* 2018 Sep 6; 12(9):e0006597.

Communications orales et conférences

Conférences issues de la thèse

- Sofeu CL, Emura T and Rondeau V. "A joint frailty-copula model for meta-analytic validation of failure time surrogate endpoints in clinical trials." *40th Annual Conference*

of the International Biostatistics (ISCB), Leuven, Belgium, 2019 July 14-18, oral section, OC 3-3

- Rondeau V, Sofeu CL and Emura T. "One-step validation method for surrogate endpoints using data from multiple randomized cancer clinical trials with failure time endpoints." *7th CHANNEL Network Conference*, Rothamsted, 2019 July 10-12, oral section.

- Sofeu CL, Emura T and Rondeau V. "One-step validation method for surrogate endpoints in multiple randomized cancer clinical trials with failure time endpoints." *GDR «Statistiques et santé» et la société Française de biométrie*, Nantes, France, 2018 September 27-28, oral section.

- Sofeu CL, Emura T and Rondeau V. "One-step validation method for surrogate endpoints in multiple randomized cancer clinical trials with failure time endpoints." *XXIX International Biometric Conference (IBC)*,Barcelona, 2018 July 8-13, oral section.

Autres conférences

- Sofeu CL, Tejiokem MC, Protopopescu C, Penda CI, Ateba Ndongo F, Tetang Ndiang S, Guemkam G, Warszawski J, Faye A, Giorgi R; ANRS-PEDIACAM study group. "Early treated HIV-infected children remain at risk of growth retardation during the first five years of live: Results from the ANRS-Pediacam cohort in Cameroon." *Revue d'Épidémiologie et de Santé Publique, EPI-CLIN*, Saint-Etienne, 2017 May 17-19, oral section.

Séminaire invité

Sofeu CL, Emura T and Rondeau V. "One-step validation method for surrogate endpoints using data from multiple randomized cancer clinical trials with failure time endpoints", *Institute of Statistical Science, Academia Sinica*, Taiwan, 2019 March 22, Oral section.

package `R`

Implémentation des fonctions `JointSurroPenal()`, `JointSurroCopPenal()`, `jointSurrSimul()`, `jointSurrCopSimul()`, `jointSurroPenalSimul()`, `jointSurroTKendall()`, `ste()`, `loocv()`, `plot()`, `predict()`, `summary()` dans le package `R frailtypack`

Missions complémentaires

Mission d'enseignement

- **Etablissement :** Institut de Santé Publique d'Epidémiologie et de Développement
- **Année académique :** 2017/2018 et 2018/2019
- **Nombre d'heures :** 64 heures.
- **Enseignements 1 :** Analyses descriptives, tests statistiques, modèles de régression avec le logiciel Rcommander (package `Rcmdr`); M1 SP-ISPED
- **Enseignements 2 :** Informatique; UFR Médicales,

Mission d'expertise

- **Entreprise :** Capionis
- **Année académique :** 2016/2017
- **Nombre de jours :** 18,66
- **Missions (3) :** Revue de la littérature, Analyse des données, Rédaction des parties Méthodes et Résultats d'un article scientifique.

Notations

S	Variable aléatoire associée au critère de substitution (Surrogate endpoint)
T	Variable aléatoire associée au vrai critère de jugement (True endpoint)
Z	Variable aléatoire binaire représentant le traitement
i et j	Utilisés pour indexer les essais / centres (i) et les individus (j)
G	Nombre d'essais cliniques ou de clusters ou de centres dans en cas d'essai multicentriques
N	Nombre de sujets dans de l'étude ou taille de l'échantillon
n_i	Nombre de sujets dans l'essai i
$f(X)$	Densité de probabilité de la variable aléatoire X
$f(X\|Z)$	Densité de probabilité de X conditionnelle a Z
α, β, mu, γ,	$\cdots$ et autre lettres grecques minuscules: paramètres.
Σ, D	Matrices de variance-covariance ou de dispersion des effets aléatoires corrélés
Φ	Signifie en général le vecteur de l'ensemble des paramètres.
$\perp$	Indépendance entre deux variables
$V^{'}$	Transposée du vecteur (ou de la matrice V)
$Cov(X,Y)$	Covariance de X et Y
$Var(X,Y)$	Variance de X et Y
MNV	Distribution normale multivariée
C_θ	Fonction de copule bivariée de paramètres θ
$f^{''}(t)$	Dérivée seconde de $f(t)$ par rapport à t
$R^2_{trial(f)}$	R^2_{trial} obtenu à partir du modèle complet, f pour `full` en anglais
$R^2_{trial(r)}$	R^2_{trial} obtenu à partir du modèle réduit, r pour `reduce` en anglais

Liste des symboles

aLCV	Approximate likelihood cross-validation criterion
CIRC	Centre International de Recherche sur le Cancer
CRAN	Comprehensive `R` Archive Network
DFS	Disease-free survival
FDA	Food and Drug Administration
GIST	Gastro-intestinales stromales
IDH	Indice de développement humain
loocv	Leave-one-out cross-validation
LPFS	Local progression-free survival
MFS	Metastatic-free survival
MPFS	Metastatic progression-free survival
MPI	Message Passing Interface
MSE	Erreurs quadratiques moyennes (mean squared errors, en anglais)
OMS	Organisation Mondiale de la Santé
OpenMP	Open Multi-Processing
OS	Overall survival ou survie globale
PFS	Progression-free survival
STE	Effet minimum d'un critère de substitution ou surrogate threshold effect
TTE	Critère de jugement à temps d'évènement

Table des matières

1 Introduction **16**

1.1 Épidémiologie du cancer 16

1.2 Traitements du cancer et point sur les essais cliniques 17

1.2.1 Traitements existants 17

1.2.2 Différentes phases d'un essai clinique 18

1.3 Points sur les critères de jugement 19

1.3.1 Vrai critère de jugement [True endpoint] 19

1.3.2 Les critères de substitution [Surrogate endpoints] 20

1.4 Approches statistiques pour la validation des critères de substitution 21

1.4.1 Validation à partir de données issues d'un seul essai clinique 21

1.4.2 Validation à partir des données issues des méta-analyses 22

1.5 Objectifs de la thèse 24

1.5.1 Objectif général 24

1.5.2 Plus spécifiquement 24

2 État de l'art **27**

A Méthodes statistiques utilisées pour valider les critères de substitution . 27

2.1 Définition et Critères de Prentice 28

2.1.1 Définition 28

2.1.2 Critères de Prentice 28

2.1.3 Vérification des critères de Prentice 29

2.2 Approche de validation basée sur un essai clinique 30

2.2.1 Proportion de l'effet du traitement expliqué (PE, Proportion explained) . 30

2.2.2 Effet relatif et association ajustée [Relatif affect and adjusted association] 31

2.3 Approche de validation en contexte de méta-analyse 33

2.3.1 Cas de deux variables de Gauss 33

2.3.2 Cas de deux temps d'évènements 39

2.4 Autres approches de validation 44

B Modèles à fragilités 45
2.5 Modèles à fragilités partagées 45
2.5.1 Type de données 45
2.5.2 Définition du modèle 46
2.5.3 Estimation des paramètres et vraisemblance pénalisée 47
2.6 Modèles conjoints à fragilités 48
2.6.1 Définition du modèle 49
2.6.2 Modèle conjoint pour temps jusqu'à la progression et le décès 49
2.6.3 Calcul de la vraisemblance 51
2.6.4 Mesure d'association au niveau individuel 52
2.7 Modèle conjoint à fragilités et à copules 52
2.7.1 Définition du modèle 53
2.7.2 Mesure d'association au niveau individuel 53

3 Validation en une étape des critères de substitution **55**
3.1 Article 55
3.2 Annexes article 73
3.3 Discussion 80
3.4 Annexes supplémentaires 80
3.4.1 Formulation de la log-vraisemblance marginale avec intégration par Monte-Carlo (MC) 80
3.4.2 Formulation de la log-vraisemblance marginale avec intégration par MC au niveau essai et quadrature de GH au niveau individuel 81

4 Validation des critères de substitution avec `frailtypack` **82**
4.1 Article 82
4.2 Annexes article 109

5 Modèle conjoint à fragilités et à copules pour critères de substitution **123**
5.1 Article 123
5.2 Code R pour l'application de ce nouveau modèle 150

6 Discussion générale **153**
6.1 Conclusion sur le travail de thèse 153
6.2 Autres avantages et limites 158
6.3 Discussions supplémentaires et perspectives 159
6.4 Démarche pour valider un critère de substitution 161

6.5 Conclusion générale 161

Bibliographie **162**

Annexes **170**

Annexe Démonstration de la présence des biais d'estimations sur R^2_{trial} non ajusté . . 171

Chapter 1

Introduction

1.1 Épidémiologie du cancer

Le cancer représente la deuxième cause de mortalité dans le monde, après les maladies cardiovasculaires (CIRC 2018; Global Burden of Disease Cancer Collaboration 2017). Le Centre International de Recherche sur le Cancer (CIRC) qui est une agence de l'Organisation Mondiale de la Santé (OMS) spécialisée sur le cancer a estimé à 18,1 millions le nombre de nouveaux cas de cancer dans le monde en 2018, et à 9,6 millions le nombre de décès par cancer au cours de la même année, estimation faite dans 185 pays. Ces chiffres présentent une croissance du fardeau mondial du cancer comparée aux estimations faites sur 195 pays par Global Burden of Disease Cancer Collaboration (2017) qui étaient de 17,5 millions de nouveaux cas et 8,7 millions de décès liés à la maladie en 2015. On note également que l'incidence et le taux de décès liés au cancer diffèrent suivant le sexe, les hommes étant les plus affectés avec un homme sur cinq contre une femme sur six dans le monde qui développeraient un cancer au cours de leur vie suivant les estimations du CIRC en 2018. De même, un homme sur huit et une femme sur 11 meurent de cette maladie.

Bien que le monde entier soit concerné par ce fléau, les taux d'incidence globaux pour de nombreux cancers dans les pays à fort indice de développement humain (IDH) sont généralement deux à trois fois plus élevés que dans les pays à faible ou moyen IDH. Cette situation s'expliquerait par le fait que les pays à IDH plus faible présentent une fréquence plus élevée de certains types de cancers associés à une moins bonne survie, mais également parce que l'accès à un diagnostic opportun et à un traitement efficace y sont moins fréquents. Avec une population totale qui représente 9% de la population mondiale, en 2018, l'Europe comptabilisait 23.4% de la charge mondiale de cancer et 20.3% des décès dus au cancer (CIRC 2018). Plus de détails sur le taux de mortalité par cancer ainsi que le taux d'incidence du cancer en Europe et dans

les pays membres de l'Union Européenne peuvent être retrouvés dans l'article de Ferlay et al. (2013)

Le CIRC présente le cancer du poumon comme le cancer le plus fréquemment diagnostiqué chez les hommes en 2018 (14,5 contre 8,4 % chez les femmes) et la principale cause de décès par cancer chez les hommes (22,0 %, soit environ un décès sur cinq), suivi pour l'incidence par le cancer de la prostate (13,5 %) et le cancer colorectal (10,9 %) chez les hommes et, pour la mortalité, par le cancer du foie (10,2 %) et le cancer de l'estomac (9,5 %). Le Global Burden of Disease Cancer Collaboration (2017) présentait le cancer de la prostate comme celui le plus incident chez les hommes en 2015. Chez la femme, le cancer du sein est le cancer le plus fréquent (24,2 %, soit environ un sur quatre des nouveaux cas de cancer diagnostiqués dans le monde), suivi par le cancer du poumon et le cancer colorectal. Le cancer du sein est le plus fréquent dans 154 des 185 pays couverts par GLOBOCAN 2018. Il est la principale cause de décès par cancer chez les femmes (15,0 %), suivi par le cancer du poumon (13,8 %) et le cancer colorectal (9,5 %). Le cancer du col de l'utérus se classe quant à lui au quatrième rang pour l'incidence (6,6 %) et la mortalité (7,5 %).

1.2 Traitements du cancer et point sur les essais cliniques

1.2.1 Traitements existants

Pour les patients chez lesquels a été diagnostiqué un cancer, le traitement dépend du type de cancer, de son étendue, de son histoire et du parcours du malade. Dans certains cas, un seul type de traitement peut s'avérer suffisant alors que pour d'autres, plusieurs traitements complémentaires peuvent être envisagés. Parmi les traitements les plus courants, on rencontre la chirurgie, la radiothérapie, la chimiothérapie et l'hormonothérapie. Une nouvelle discipline dans la recherche des traitements concerne l'immunothérapie qui vise à mobiliser les défenses immunitaires du patient contre sa maladie. C'est pourquoi de nombreux essais cliniques d'immunothérapie sont en cours actuellement. Un autre type de traitement concerne la thérapie ciblée. Il s'agit d'un traitement médicamenteux qui cible les anomalies spécifiques du cancer, comme des gènes, des protéines ou des modifications de l'environnement tissulaire qui contribuent à la croissance du cancer. Ce type de traitement cherche à éliminer les cellules cancéreuses ou à bloquer leur croissance et leur propagation tout en limitant les dommages aux cellules normales. Les thérapies ciblées ont pour avantage d'avoir moins d'effets secondaires que les autres médicaments anti-cancéreux selon la Fondation contre le cancer (2019).

Nous faisons face à deux catégories de traitements dont l'une est à visée curative et l'autre à visée palliative. Contrairement aux traitements curatifs à partir desquels on espère une

guérison, les traitements palliatifs permettent de ralentir ou même stopper pendant un certain temps l'évolution de la maladie, sans toutefois donner lieu à une guérison définitive. Dans cette démarche de recherche permanente des solutions visant à éradiquer cette pandémie, il est nécessaire de développer de nouveaux traitements, qui permettraient de limiter la prolifération des cellules cancéreuses ou de les éradiquer pour les cancers métastatiques. De tels traitements devraient avoir moins d'effets indésirables et garantir une bonne qualité de vie aux patients. D'où la nécessité de suivre de façon minutieuse les différentes phases qui régissent le développement d'un nouveau médicament.

1.2.2 Différentes phases d'un essai clinique

L'évaluation de l'efficacité et de la tolérance d'un médicament passe par la mise en place d'un essai clinique qui se déroule le plus souvent en 4 phases précédées d'une phase pré-clinique . La phase pré-clinique consiste en l'étude de la molécule, sa structure, son effet sur les cellules, son effet sur l'animal au niveau comportemental et biologique, ainsi que l'étude des organes-cibles. Elle se réalise exclusivement sur des modèles animaux à savoir des rongeurs (souris, rats et gerbilles) et des non-rongeurs : chiens (de moins en moins utilisé), porcs pour la « proximité biologique » avec l'homme ou les primates (lorsqu'une molécule a démontré son intérêt). A l'issue de ces études, la dose maximale tolérée par l'animal de laboratoire est convertie en dose maximale sécuritaire à utiliser chez l'homme dans les prochaines phases de l'essai clinique.

La phase I consiste à évaluer la tolérance et l'absence d'effets indésirables chez des sujets le plus souvent volontaires sains. Parfois ces essais peuvent être proposés à des patients en impasse thérapeutique, pour lesquels le traitement étudié représente la seule chance de survie. C'est au cours de cette phase que la dose maximale tolérée est également définie. Les groupes étudiés sont le plus souvent de petites tailles (20 à 80 participants). Certains médicaments dont on sait par nature qu'ils sont toxiques (par exemple les anticancéreux) peuvent ne pas faire l'objet d'une phase I et entrer directement en phase II. Dans la phase II, on s'intéresse à la dose optimale du médicament et à ses éventuels effets indésirables chez un groupe plus conséquent de sujets malades. L'objectif est de confirmer l'activité clinique préliminaire et/ou pharmacologique du médicament à la dose recommandée à l'issue de la phase I. Bien souvent, cette phase est comparative : la nouvelle molécule est administrée à un groupe de patients, tandis que l'autre reçoit un placebo. La phase III quant à elle est réservée à l'étude comparative de l'efficacité du nouveau médicament face au placebo ou à un médicament de référence, s'il existe. Les groupes sont de tailles importantes, souvent plusieurs milliers de participants. Ces essais, très souvent multicentriques, se conduisent en double aveugle (ce qui est rarement le cas en oncologie), ce qui permet d'écarter tout préjugé ou jugement faussé de l'une ou l'autre

partie sur son efficacité ou ses effets indésirables.

En fin dans la phase IV les médicaments dont l'efficacité a été démontrée dans la phase III et ayant obtenu une autorisation de mise sur le marché continuent de faire l'objet d'un suivi strict et long. L'objectif est d'identifier tout effet secondaire grave et/ou inattendu dû à son administration et de préciser les conditions d'utilisation pour certains groupes de patients à risque. Cette phase permet d'analyser les interactions médicamenteuses et favorise la mise au point de nouvelles formes galéniques ainsi que des extensions d'indications thérapeutiques.

Un traitement dont l'efficacité a été démontrée à l'issue de la phase III mériterait d'être mis sur le marché assez rapidement afin de permettre à un grand nombre de patients d'en tirer des bénéfices. Toutefois, il convient de noter que la durée et le succès de la phase III dépendent du critère de jugement qui sera utilisé pour évaluer l'efficacité du traitement.

1.3 Points sur les critères de jugement

1.3.1 Vrai critère de jugement [True endpoint]

L'un des facteurs les plus importants qui influent sur la durée et la complexité du processus de développement de nouveaux traitements, est le choix de l'événement d'intérêt qui sera utilisé pour évaluer l'efficacité du traitement. Fleming et al. (1994) ont proposé deux critères principaux pour sélectionner le bon événement d'intérêt, et sa capacité à bien détecter les effets du traitement ainsi que sa pertinence clinique pour répondre aux objectifs de l'étude. Cependant, il semble que le critère d'évaluation clinique le plus sensible et pertinent, habituellement appelé le "vrai" événement d'intérêt, est difficile à utiliser dans un essai clinique. Par exemple, dans des essais cliniques sur le cancer, la survie globale (OS) définie comme le temps depuis la randomisation jusqu'au décès (toute cause) est un critère clinique fréquemment utilisé pour évaluer l'effet de nouveaux traitements. Cependant, face aux progrès thérapeutiques et au développement de nouveaux types de cytostatique, on assiste à une baisse significative de la mortalité liée au cancer. Par conséquent, un essai clinique dont le critère de jugement principal est la survie globale demanderait des temps de suivi plus longs pour atteindre le nombre d'événements nécessaires pour mettre en évidence une différence significative (si elle existe) de l'effet du traitement entre les patients traités et le groupe contrôle. Ce phénomène induit une augmentation des coûts de développement des essais cliniques. Par ailleurs, les échecs thérapeutiques peuvent conduire à la multiplicité des lignes de traitement chez certains patients, diluant ainsi l'effet recherché du traitement sur le critère de jugement considéré (Matulonis et al. 2015).

Suite à ce constat, ces dernières années, plusieurs travaux de recherche clinique se sont intéressés au remplacement de ce critère de jugement principal par des critères dits de substi-

tution ou critères intermédiaires (Fleming et al. 1994; Prentice 1989; Fiteni et al. 2014). Ces critères de substitution sont des mesures qui peuvent être recueillies plus tôt et/ou en utilisant moins de sujets que les critères de jugement principaux. Ils sont essentiellement plus fréquents et/ou plus facile à utiliser. L'utilisation d'un autre critère de jugement observable, comme la progression du cancer, permettrait donc de raccourcir la durée d'un essai clinique tout en fournissant le plus tôt possible des conclusions sur le critère de jugement principal.

1.3.2 Les critères de substitution [Surrogate endpoints]

Définition et identification des critères de substitution

Un critère de substitution est un critère intermédiaire pouvant être observé plus tôt que le vrai critère de jugement et dont l'utilisation permettrait donc de prédire l'effet du traitement sur le vrai critère de jugement. Dans la suite de ce chapitre, nous nous intéressons exclusivement aux critères de jugement qui sont des temps d'évènement. L'utilisation des critères de substitution passe par la nécessité de définir de façon précise et objective chaque critère de jugement à temps d'évènement (TTE) comme le préconisent les recommandations internationales (Bellera et al. 2014). Cette définition passe par des procédures pouvant faciliter la comparaison entre les essais cliniques et favoriser la reproductibilité (Therasse et al. 2000; Eisenhauer et al. 2009; Kassaï et al. 2007).

Suivant le type de cancer, on distingue globalement des approches de définition basées sur des opinions d'experts (Bellera et al. 2014, 2013) et d'autres approches plus objectives qui font recours à des procédures standardisées (WHO 1979; Therasse et al. 2000; Eisenhauer et al. 2009). Bellera et al. (2014) ont proposé des recommandations basées sur des opinions d'experts pour la définition des TTE à utiliser pour l'évaluation du traitement dans des essais cliniques randomisés sur les sarcomes et les tumeurs gastro-intestinales stromales (Gastro-intestinales stromales). Les TTE retenus étaient des critères de jugement composites incluant différents évènements pouvant survenir au cours du suivi. Par exemple, la survie sans récidive (Disease-free survival, DFS) définie comme le temps depuis la randomisation jusqu'à la survenue du décès toutes causes ou la survenue d'un évènement local, régional ou métastatique/avancé. Parmi les critères intermédiaires proposés, la DFS et la survie sans métastases (MFS) sont propres aux cancers adjuvants, tandis que pour des cancers métastatiques on aurait plutôt recours à la survie sans progression (PFS), la survie sans progression locale (LPFS) ou la survie sans progression métastatique (MPFS).

Eisenhauer et al. (2009) ont proposé des critères standardisés permettant de définir la PFS, et donc d'évaluer la réponse tumorale pour différents types de tumeurs solides. Ces critères viennent en complément de ceux proposés par Therasse et al. (2000) et l'OMS (WHO 1979)

afin de trouver un langage commun pour la définition de la progression lors de la mise en place des essais cliniques. Par ailleurs, à l'exception de l'OS qui est considérée comme gold standard et comme le critère clinique le plus pertinent par la Food and Drug Administration (FDA), les conclusions sur l'efficacité d'un nouveau médicament dans les essais de phase III en s'appuyant uniquement sur des critères intermédiaires semblent être dangereuses si ces dernières n'ont pas été soigneusement validées. Kassaï et al. (2007) discutent des limites quant à l'utilisation des critères intermédiaires dans des essais cliniques, et particulièrement des approches purement descriptives pour la validation des critères de substitution. De telles méthodes, surtout celles basées sur de simples corrélations entre le critère de substitution et le vrai critère de jugement ont conduit les agences de réglementation à être méfiantes vis-à-vis de l'utilisation des critères de substitution dans des essais cliniques de phase III.

1.4 Approches statistiques pour la validation des critères de substitution

Avant qu'un critère de substitution soit proposé il doit être validé de manière rigoureuse, à l'aide de méthodes statistiques appropriées. En s'appuyant sur des exemples observés sur le cancer de la prostate, de l'estomac ou du sein, Paoletti et al. (2016) reviennent sur les deux principaux cadres utilisés pour valider les critères de substitution, à savoir la validation basée sur un seul essai clinique et celle basée sur des méta-analyses. L'utilisation d'une simple corrélation entre les deux critères (de substitution et le principal) ne suffit pas pour valider un critère de substitution (Baker and Kramer 2003; Korn et al. 2005). Plusieurs auteurs (Burzykowski et al. 2005; Freedman et al. 1992; Prentice 1989; Baker 2018) ont proposé différentes méthodes pour mener à bien cette validation.

1.4.1 Validation à partir de données issues d'un seul essai clinique

Prentice (1989) a proposé quatre critères opérationnels pour vérifier si un triplé (T, S, Z) (vrai critère de jugement, critère de substitution et traitement) permet de répondre à la définition d'un critère de substitution. Le quatrième critère, souvent considéré comme le critère de Prentice exige que la totalité de l'effet du traitement sur le vrai critère de jugement soit capturé par le critère de substitution. Bien que considéré comme intéressant, ce dernier critère de Prentice a été largement critiqué et est présenté comme irréaliste ou "conceptuellement difficile" dans le sens où il exige que le test statistique de l'effet du traitement sur le vrai critère de jugement soit non significatif après ajustement sur le critère de substitution (Freedman et al. 1992). Ainsi ce dernier critère pourrait être utile pour rejeter un faible critère de substitution (quand le

test statistique de l'effet du traitement sur le vrai critère reste statistiquement significatif après ajustement sur le critère de substitution), mais il est inadéquat pour valider un bon critère de substitution. En effet, ne pas rejeter l'hypothèse nulle peut être dû simplement à un manque de puissance statistique.

Ce constat justifie l'utilisation d'un grand nombre d'observations pour valider des critères de substitution (Buyse and Molenberghs 1998), et donc le recours aux données provenant des méta-analyses. De plus comme discuté par Paoletti et al. (2016), une autre limite liée à la validation basée sur les critères de Prentice (et donc sur un seul essai clinique) viendrait de l'ajustement sur le critère de substitution qui est observé au cours du suivi. Par conséquent, il devient difficile d'interpréter l'effet du traitement en raison de l'ajustement sur une variable dépendante du temps. En revanche, les approches basées sur des méta-analyses non seulement prennent en compte l'association au niveau individuel, mais également l'association au niveau essai. Des travaux clés ont proposé de valider des critères de substitution (tels que la survie sans événement, le temps jusqu'à la progression ou la survenue d'une récidive) dans un contexte de méta-analyses d'essais cliniques ou d'études multicentriques (Burzykowski et al. 2005; Michiels et al. 2009).

1.4.2 Validation à partir des données issues des méta-analyses

Lorsqu'on considère le cas où à la fois le critère de substitution et le vrai critère de jugement sont des temps d'évènement, la validation du critère de substitution est plus complexe en raison de plusieurs paramètres, comme la présence de la censure et de risques compétitifs. Pour toutes ces raisons, différents auteurs ont tenté de développer des méthodes de validation des critères de substitution dans ce contexte d'analyse de survie. Par exemple, Burzykowski et al. (2001) ont proposé une méthode basée sur une extension de l'approche développée par Buyse et al. (2000) en s'appuyant sur des données issues des méta-analyses. Une revue de littérature récente décrit dans le contexte des essais cliniques en oncologie les différentes approches possibles pour valider un critère de substitution (Buyse et al. 2016).

Bien que les critères de substitution soient toujours un sujet de recherche très présent dans la littérature, le consensus actuel est de baser la validation sur une approche de "corrélation" et une stratégie de modélisation en deux étapes (Burzykowski et al. 2005; Paoletti et al. 2016). Dans une première étape, afin de valider la qualité du critère de substitution au niveau individuel, Burzykowski et al. (2001) ont utilisé une mesure d'association entre le critère de substitution et le vrai critère de jugement en utilisant un modèle à copules. Dans une deuxième étape, le critère de substitution est considéré comme valide au niveau essai s'il est capable de prédire l'effet du traitement sur le vrai critère de jugement en s'appuyant sur l'effet du traitement observé sur

le critère de substitution. Afin d'évaluer la validité au niveau essai, Burzykowski et al. (2001) ont utilisé un modèle bivarié à effets aléatoires basé sur les effets du traitement estimés dans le modèle de première étape. Les auteurs ont proposé d'ajuster ce dernier modèle sur les erreurs d'estimations des effets du traitement. La qualité du critère de substitution au niveau essai a été évaluée par le coefficient de détermination ajusté (R^2_{ajust}) obtenu à partir de ce modèle. Cependant, dans beaucoup d'études considérées, des problèmes de convergence ou d'estimation des modèles sont rencontrés et ce coefficient R^2_{ajust} n'est pas toujours disponible (Burzykowski et al. 2005; Renfro et al. 2012; Shi et al. 2011; Rotolo et al. 2019). Renfro et al. (2012) ont souligné que ces problèmes de convergence étaient le plus souvent rencontrés dans l'étape 1 (au niveau individuel) de la validation. Mais même lorsque cette étape 1 fournit des estimateurs, l'étape 2 (au niveau essai) ne donne pas toujours une estimation du coefficient de détermination ajusté. Ces problèmes numériques sont fréquemment rencontrés et sont influencés par le niveau d'association entre le critère de substitution et le vrai critère de jugement, le nombre et la taille des groupes (ou essais) ainsi que les hypothèses faites sur les risques de base d'un essai à l'autre.

Plusieurs articles ont été publiés sur la validation des critères de substitution, mais les méthodes proposées ne sont pas satisfaisantes notamment en termes de biais sur la mesure d'association considérée au niveau essai (Shi et al. 2011). Il ressort des revues de la littérature récentes (Branchoux et al. 2019; Savina et al. 2018) que plusieurs études s'appuyaient sur une validation au niveau essai par de simples corrélations (ou encore sur la régression linéaire pondérée) entre les effets du traitement observés sur le critère de substitution et le vrai critère de jugement. Il a été montré par Shi et al. (2011) à partir des études de simulation que le coefficient de détermination obtenu sur la base de ces approches est dans le meilleur des cas comparable au R^2 non ajusté obtenu dans l'approche de (Burzykowski et al. 2001). Or le R^2 non ajusté tel que nous le montrerons dans le chapitre suivant est biaisé.

Dans ce contexte, il semble nécessaire d'améliorer les méthodes proposées pour évaluer des critères de substitution. Pour résoudre ces problèmes, nous nous sommes proposé dans cette thèse de développer de nouvelles méthodologies statistiques basées sur des modèles conjoints pour valider des critères de substitution. Les modèles utilisés pour valider un critère de substitution associé à un critère de jugement clinique dépendent du type de variables observées. Nous nous concentrerons principalement sur le cas où à la fois le critère de substitution et le critère de jugement principal sont des temps d'évènement.

1.5 Objectifs de la thèse

1.5.1 Objectif général

L'objectif de cette thèse était de proposer une nouvelle approche statistique de validation d'un critère de substitution pour la survie globale après un premier cancer. Cette approche prolonge la méthode en deux étapes décrite par Burzykowski et al. (2001), avec la résolution de certains problèmes de convergence et d'optimisation. Une approche globale et non en deux étapes a été développée en s'appuyant sur une modélisation conjointe d'un critère de substitution et d'un critère de jugement principal. Un logiciel statistique associé à ces développements a été proposé.

1.5.2 Plus spécifiquement

Dans la première partie de cette thèse, nous avons travaillé sur un nouveau modèle conjoint à fragilités partagées qui lie la fonction de risque associée au critère de substitution à la fonction de risque associée au critère de jugement principal. Nous avons proposé un modèle à effets aléatoires avec une structure de corrélation permettant de répondre à la définition d'un critère de substitution dans les méta-analyses. Contrairement à l'approche standard qui effectue une modélisation en deux étapes (Burzykowski et al. 2001), nous avons proposé une approche globale en une seule étape grâce à un modèle conjoint. Nous nous sommes appuyé sur des modèles conjoints à fragilités existants (Rondeau et al. 2007, 2015, 2008). Notre modèle traite par exemple la dépendance entre les temps de progression tumorale et la survie globale en utilisant des effets aléatoires individuels pour valider le critère de substitution au niveau individuel. Dans ce même modèle, des effets aléatoires spécifiques au groupe (ou à l'essai) et en interaction avec le traitement ont été introduits pour valider le critère de substitution au niveau essai.

L'analyse combinée de plusieurs essais exige également de tenir compte de l'hétérogénéité entre les essais (Rondeau et al. 2015). A cet effet, nous avons inclus dans notre modèle des effets aléatoires partagés associés aux fonctions de risque de base pour la prise en compte de l'hétérogénéité au niveau essai. Tous les effets aléatoires sont supposés gaussiens. La difficulté avec cette approche réside dans l'approximation des intégrales présentes dans la vraisemblance du modèle et donc l'estimation des paramètres. Nous avons donc travaillé sur différentes méthodes d'intégration numérique. Les performances du modèle ont été validées par des études de simulation et ont été comparées aux approches existantes. En parallèle de l'estimation du modèle développé, nous avons travaillé sur des mesures pour quantifier et évaluer la qualité des critères de substitution au niveau individuel et au niveau essai. Par conséquent, nous avons proposé une nouvelle définition du Tau de Kendall (τ) et un nouveau coefficient de détermina-

tion.

Dans la deuxième partie de ce projet, nous nous sommes fixé comme objectif de développer un logiciel convivial et complet, utile pour biostatisticiens, épidémiologistes, cliniciens et d'autres personnes impliquées dans l'épidémiologie et la recherche clinique. L'idée de ce travail était de vulgariser le modèle auprès de la communauté scientifique, de favoriser la reproductibilité des résultats présentés dans cette thèse mais également de permettre à toute personne désireuse d'étendre ces modèles d'avoir une base de travail conséquente. Nous avons ainsi étendu le package `R frailtypack` (Król et al. 2017) à la validation des critères de substitution. `frailtypack` est un logiciel open source destiné à l'analyse des données corrélées à temps d'évènement (Rondeau et al. 2007). Ce package implémente plusieurs types de modèles à fragilités. La première extension avec les modèles conjoints permettait d'estimer les paramètres d'un modèle conjoint pour des données récurrentes et un évènement terminal. Nous y avons ainsi implémenté de nouveaux outils associés à la validation des critères de substitution, avec une documentation conséquente. Pour faciliter l'utilisation du package à travers les options proposées, nous avons préparé un tutoriel incluant des exemples d'applications et d'interprétation des sorties du logiciel.

Enfin, nous avons examiné un autre type de modèle conjoint mais cette fois-ci en combinant des effets aléatoires et des copules. L'objectif était d'étendre le modèle conjoint à fragilités et à copules de Clayton proposé par Emura et al. (2017) dans un contexte de méta-analyses à un modèle de validation d'un critère de substitution. Dans cette nouvelle approche la dépendance au niveau individuel entre, par exemple, les temps de progression tumorale et la survie globale a été prise en compte par un paramètre de copule, alors que l'hétérogénéité de l'effet du traitement au niveau essai est prise en compte par des effets aléatoires corrélés. Par simplicité, nous avons considéré les copules de Clayton et les copules de Gumbel. La flexibilité et les avantages mathématiques de ces copules permettent d'estimer les paramètres des modèles conjoints complexes pour des données de méta-analyses avec un temps de calcul réduit. Les performances de cette méthode ont également été évaluées par des études de simulations et ont été comparées à l'approche précédente incluant les effets aléatoires au niveau individuel.

Ces développements sont motivés par l'analyse de différentes bases de données et méta-analyses dans le domaine du cancer. Nous avons ainsi appliqué nos modèles à des données individuelles de méta-analyses issues de la recherche de nouveaux traitements pour le cancer gastrique et le cancer ovarien.

Dans la suite de cette thèse, nous présentons dans le Chapitre 2 un état de l'art sur quelques méthodes statistiques existantes, et permettant la validation des critères de substitution. La deuxième partie de ce chapitre est consacrée à la présentation des modèles conjoints à fragilités qui ont orienté la spécification des modèles développés dans cette thèse. Les Chapitres 3, 4

et 5 sont consacrés à la présentation des articles associés aux objectifs spécifiques introduits précédemment. Enfin dans le Chapitre 6, nous présentons une discussion générale de nos travaux, suivie par des perspectives et une conclusion générale.

Chapter 2

État de l'art

Dans ce chapitre, nous présenterons les méthodes statistiques de validation des critères de substitution depuis les critères de Prentice (1989). Ainsi, nous évoluerons progressivement des approches de validation s'appuyant sur un essai clinique randomisé vers des approches plus englobantes, basées sur des données de méta-analyses d'essais cliniques randomisés (Burzykowski et al. 2005). Étant donné dans le cadre de cette thèse notre intérêt pour les approches de validation en une étape, nous ferons dans la suite de ce chapitre une revue sur les modèles conjoints à fragilités et les méthodes d'estimations afférentes.

A Méthodes statistiques utilisées pour valider les critères de substitution

Soient S et T deux variables aléatoires représentant le critère de substitution (surrogate endpoint) et le vrai critère de jugement (True endpoint); soit Z une variable aléatoire binaire représentant le traitement.

La validation d'un critère de substitution basée sur des approches statistiques demande de s'appuyer sur une démarche appropriée. Sur la Figure 2.1 ci-contre, on peut voir qu'une simple corrélation entre les critères de jugement ne suffit pas pour prédire l'effet du traitement sur le vrai critère de jugement lors de la mise en place d'un essai clinique. Dans le premier cas de figure (a), on observe une forte corrélation entre S et T dans chacun des bras de traitement (Z_0 et Z_1). Par contre, une augmentation de l'espérance de S sous l'effet du traitement n'entraîne pas une augmentation de l'espérance de T, et donc l'effet du traitement sur S n'est pas associé à l'effet du traitement sur T. Dans le deuxième cas de figure (b), on n'observe pas d'association entre S et T dans les différents groupes de traitement. En revanche, une augmentation de l'espérance de S sous l'effet du traitement entraîne une augmentation de l'espérance de T, et

donc les effets du traitement sont corrélés.

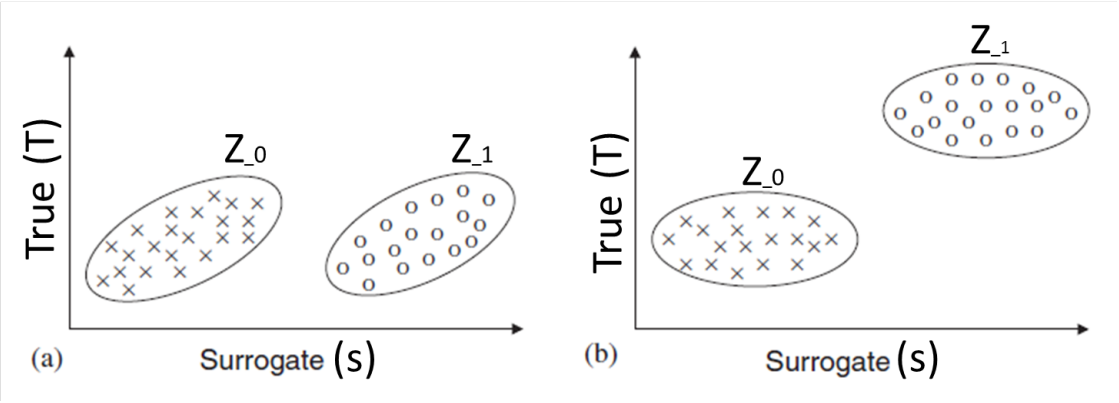


Figure 2.1: Relation hypothétique entre le critère de substitution (S) et le vrai critère de jugement (T) pour deux bras de traitement (Z_0 et Z_1). (a) présente une association forte au niveau individuel et faible au niveau essai. (b) présente une association faible au niveau individuel et forte au niveau essai. Inspiré de Korn et al. (2005).

Pour qu'un critère de substitution soit considéré comme valide, il doit être valide au niveau individuel et au niveau essai. S'appuyant sur des données issues d'un essai clinique randomisé, Prentice (1989) a formulé une définition pour un critère de substitution ainsi que quatre critères opérationnels permettant de vérifier si un triplé (T, S, Z) obéit à la dite définition.

2.1 Définition et Critères de Prentice

2.1.1 Définition

Un critère de substitution selon Prentice est une variable réponse pour laquelle le test d'hypothèse nulle d'absence d'effet du traitement est également un test d'hypothèse valide d'absence d'effet du traitement sur le vrai critère de jugement. Cette définition peut se résumer comme suit:

$$f(S|Z) = f(S) \leftrightarrow f(T|Z) = f(T) \tag{2.1}$$

où $f(X)$ est la densité de probabilité de la variable aléatoire X, $f(X|Z)$ est la densité de probabilité de X conditionnellement à la valeur de Z et $\leftrightarrow$ symbolise la relation d'équivalence.

2.1.2 Critères de Prentice

Afin de vérifier si un triplé (T, S, Z) obéit à la définition précédente, les critères opérationnels qui suivent doivent être vérifiés:

$$f(S|Z) \neq f(S) \tag{2.2}$$
$$f(T|Z) \neq f(T) \tag{2.3}$$
$$f(T|S) \neq f(T) \tag{2.4}$$
$$f(T|S,Z) = f(T|S) \tag{2.5}$$

Ces critères de Prentice stipulent que: l'effet du traitement sur le critère de substitution soit statistiquement significatif (2.2), l'effet du traitement sur le vrai critère de jugement soit statistiquement significatif (2.3), l'effet du critère de substitution sur le vrai critère de jugement soit statistiquement significatif (2.4) et que l'effet du traitement sur le vrai critère de jugement soit totalement capturé par le critère de substitution (2.5).

2.1.3 Vérification des critères de Prentice

Dans la suite de ce chapitre nous supposons que le critère de substitution et le vrai critère de jugement sont deux variables gaussienne. Les deux premiers critères de Prentice (2.2) et (2.3) peuvent être vérifiés en testant si les paramètres α et β dans le modèle suivant sont significativement différents de 0 :

$$S_j = \mu_S + \alpha Z_j + \epsilon_{Sj} \tag{2.6}$$
$$T_j = \mu_T + \beta Z_j + \epsilon_{Tj} \tag{2.7}$$

où $(\epsilon_{Sj}, \epsilon_{Tj})'$ sont des effets aléatoires corrélés distribués selon une loi normale multivariée de moyenne nulle et de matrice de variance-covariance

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ & \sigma_{TT} \end{pmatrix} \tag{2.8}$$

avec $\epsilon_{Sj} \perp \epsilon_{Sk}$ et $\epsilon_{Tj} \perp \epsilon_{Tk}$ pour $j \neq k$ (le symbole $\perp$ représente l'indépendance entre deux variables); α et β sont les effets fixes du traitement sur S et T respectivement; μ_S et μ_T représentent les intercepts pour S et T respectivement dans le groupe contrôle. Le troisième critère peut être vérifié en testant si le paramètre γ est significativement différent de 0 dans le modèle décrivant la relation entre S et T :

$$T_j = \mu + \gamma S_j + \epsilon_j. \tag{2.9}$$

Le quatrième critère de Prentice peut être vérifié à partir de la distribution de T conditionnellement à S et Z. A partir des équations (2.6) et (2.7), et d'après les propriétés de la loi normale multidimensionnelle, on obtient :

$$T_j = \tilde{\mu}_T + \beta_S Z_j + \gamma_z S_j + \tilde{\epsilon}_{Tj} \tag{2.10}$$

où

$$\beta_S = \beta - \sigma_{TS}\sigma_{SS}^{-1}\alpha, \tag{2.11}$$

$$\gamma_z = \sigma_{TS}\sigma_{SS}^{-1}, \tag{2.12}$$

et la variance des effets aléatoires $\tilde{\epsilon}_T$ est donnée par

$$\sigma_{TT} - \sigma_{TS}^2\sigma_{SS}^{-1}$$

.

Vérifier le quatrième critère de Prentice revient à montrer que l'effet du traitement dans le modèle (2.10) est nul ($\beta_S \equiv 0$). En revanche ne pas rejeter l'hypothèse nulle ($\beta_S = 0$) dans le modèle (2.10) pourrait être dû à un manque de puissance statistique. C'est dans ce contexte que Freedman et al. (1992); Buyse and Molenberghs (1998) ont proposé de nouvelles métriques qui permettraient de valider (2.5), en s'appuyant sur des estimations plutôt que sur un test statistique.

2.2 Approche de validation basée sur un essai clinique

2.2.1 Proportion de l'effet du traitement expliqué (PE, Proportion explained)

La validation du critère (2.5) pose une difficulté conceptuelle telle que discutée par Freedman et al. (1992) en ce sens où elle exige que le test statistique de l'effet du traitement sur T (β_S) soit non significatif lorsqu'on ajuste sur S. Ainsi, le critère (2.5) pourrait permettre de rejeter un faible critère de substitution, lorsque le test sur l'effet du traitement après ajustement sur S reste significatif. En revanche le critère (2.5) reste inadéquat pour valider un bon critère de jugement en ce sens où il est difficile de prouver que l'hypothèse nulle (H_0: $\beta_S = 0$) est vraie, car le non rejet de H_0 pourrait simplement être dû à un problème de puissance. Cette observation justifie donc l'utilisation des tailles d'échantillons élevées afin de mettre en évidence un effet significatif s'il existe. Par ailleurs, comme discuté par Burzykowski et al. (2005), même en

l'absence du problème de puissance, la signification statistique des tests non ajustés et ajustés ne permet pas de quantifier l'impact du critère de substitution sur l'analyse du vrai critère de jugement. En effet, la signification statistique ne peut pas prouver que S permet de capturer l'effet intégral du traitement sur T. De ces observations, Freedman et al. (1992) ont suggéré de se focaliser sur la proportion de l'effet du traitement expliquée (PE) par S. Ainsi, un bon critère de substitution permettra d'expliquer une large proportion d'effet. Étant donné le triplé (T, S, Z), PE est défini comme suit:

$$PE(T, S, Z) = \frac{\beta - \beta_S}{\beta} = 1 - \frac{\beta_S}{\beta} \tag{2.13}$$

où β et β_S sont les estimations des effets du traitement Z sans et avec ajustement sur S. Par exemple, β et β_S peuvent être obtenus à partir des modèles (2.7) et (2.10) dans le cas de deux variables de Gauss. Un critère de substitution valide exige que PE soit égale à 1. En pratique, un critère de substitution sera acceptable si la borne inférieure de l'intervalle de confiance de PE est proche de 1. Buyse et al. (2016) ont discuté des difficultés émanant de l'utilisation de PE. Notamment, PE tend à être instable lorsque β est proche de 0, une situation qui peut arriver en pratique. Par ailleurs, l'intervalle de confiance de PE aura tendance à être large et parfois avec des bornes situées au-delà de la limite $[0, 1]$, sauf en cas de grande taille d'échantillon (ce qui est possible en faisant recours à des données de méta-analyse) ou d'un grand effet du traitement sur T (ce qui n'est pas courant en pratique). Un autre problème viendrait de la mauvaise spécification du modèle (2.10). En présence d'une interaction significative dans ce modèle entre S et T (voir Equation 2.14), la définition même de PE deviendrait problématique.

$$T_j = \tilde{\mu}_T + \tilde{\beta}_S Z_j + \tilde{\rho}_Z S_j + \gamma Z_j S_j + \tilde{\epsilon}_T j \tag{2.14}$$

Ces difficultés fondamentales issues de la définition du PE ont donc amené Buyse and Molenberghs (1998) à proposer deux autres métriques permettant la validation des critères de substitution.

2.2.2 Effet relatif et association ajustée [Relatif affect and adjusted association]

Buyse and Molenberghs (1998) ont proposé de remplacer PE par deux nouvelles mesures d'association. La première étant l'effet relatif (RE) qui est le rapport des effets du traitement sur le vrai critère de jugement et le critère de substitution:

$$RE(T, S, Z) = \frac{\beta}{\alpha} \tag{2.15}$$

où β, α et les paramètres de variances associés sont obtenus à partir des modèles (2.6) à (2.8). Si on peut faire l'hypothèse d'une relation multiplicative entre β et α et si RE était connu, il pouvait être utilisé pour prédire l'effet du traitement sur T à partir de l'effet observé sur S. En pratique, RE est plutôt estimé, et la précision de l'estimation doit être pertinente pour une meilleure précision de la prédiction. RE ainsi défini permet de mesurer la qualité du critère de substitution au niveau de la population, et vaut 1 pour une validité parfaite comme ont suggéré Buyse and Molenberghs (1998).

La deuxième quantité proposée par Buyse and Molenberghs (1998) permet de prendre en compte l'association au niveau individuel entre les deux critères de jugement, après prise en compte de l'effet du traitement. Il s'agit de l'association ajustée (ρ_z). Si les variables S et T sont issues d'une distribution normale, l'association ajustée est définie ainsi :

$$\rho_z = \frac{\sigma_{ST}}{\sqrt{\sigma_{SS}\sigma_{TT}}} \tag{2.16}$$

où σ_{ST}, σ_{SS} et σ_{TT} sont les éléments de la matrice de variances-covariances Σ définie dans (2.8). Tout comme précédemment, on aura une validité parfaite au niveau individuel pour le critère de substitution si ρ_z vaut 1. Toutefois, il convient en pratique de juger du niveau d'association acceptable pour considérer un critère de substitution comme valide.

Au-delà des problèmes cités plus haut émanant de la définition du RE, Molenberghs et al. (2002) ont montré que PE pouvait s'écrire comme une fonction de RE et et ρ_z. En effet, si on définit $\lambda^2 = \sigma_{TT}\sigma_{SS}^{-1}$, alors $\lambda\rho_z = \sigma_{ST}\sigma_{SS}^{-1}$ et à partir de (2.11) et (2.13), on obtient:

$$\beta_S = \beta - \rho_z\lambda\alpha$$

$$PE = \lambda\rho_z\frac{\alpha}{\beta} = \lambda\rho_z\frac{1}{RE}. \tag{2.17}$$

Cette nouvelle formulation du PE permet de voir cette quantité comme une mesure prenant en compte l'association au niveau individuel et au niveau essai. Seulement, à partir de (2.17), PE peut prendre n'importe quelle valeur, et donc des valeurs au delà de 1, rendant son interprétation difficile en tant que proportion.

Par ailleurs, la difficulté dans l'utilisation du RE réside au niveau de son intervalle de confiance qui peut être très large et de l'impossibilité de vérifier l'hypothèse de la relation multiplicative entre les effets du traitement (voir équation 2.15) en s'appuyant sur des données issues d'un seul essai clinique. Face à ce constat, Buyse et al. (2000) ont suggéré de généraliser le concept de validation au niveau individuel et au niveau de la population dans un contexte de méta-analyse ou d'études multicentriques. L'usage des données issues de méta-analyse d'essais cliniques randomisés permet d'améliorer la précision des estimations dans le processus de vali-

dation des critères de substitution, de limiter les problèmes de puissance statistique, et surtout de prendre en compte l'hétérogénéité au niveau essai. Ainsi, un critère intermédiaire validé comme critère de substitution à partir des données de méta-analyse pourra être utilisé dans tout autre essai pour prédire l'effet du traitement sur le critère de jugement final.

2.3 Approche de validation en contexte de méta-analyse

Dans la suite de ce chapitre, nous supposons disposer des données individuelles issues d'une méta-analyse d'essais cliniques randomisés incluant $i = 1 \cdots G$ essais, et que chaque essai inclue $j = 1 \cdots n_i$ sujets. Le nombre total de sujets dans cette méta-analyse vaut N. Soit T_{ij} et S_{ij} deux variables aléatoires décrivant les observations du sujet j de l'essai i sur le critère de jugement principal et sur le critère de substitution, et Z_{ij} l'indicatrice de traitement pour le même sujet.

2.3.1 Cas de deux variables de Gauss

Lorsque $(S_{ij}, Tij)'$ suivent une distribution normale multivariée, Buyse et al. (2000) ont proposé deux approches distinctes de modélisation pour la validation des critères de substitution. La première approche s'appuie sur un modèle hiérarchique en deux étapes et la seconde s'appuie sur un modèle linéaire à effets mixtes multivarié.

Modèle en deux étapes

Dans le modèle en deux étapes, la première étape est basée sur un modèle à effets fixes et défini comme suit:

$$S_{ij}|Z_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \epsilon_{Sij}, \tag{2.18}$$

$$T_{ij}|Z_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \epsilon_{Tij}, \tag{2.19}$$

où μ_{Si} et μ_{Ti} représentent les intercepts spécifiques au traitement; α_i et β_i représentent les effets fixes du traitement sur les critères de jugement spécifiques aux essais; ϵ_{Sij} et ϵ_{Tij} sont deux effets aléatoires corrélés distribués selon une loi normale multivariée de moyenne nulle et de matrice de variance-covariance :

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ & \sigma_{TT}, \end{pmatrix} \tag{2.20}$$

avec $\epsilon_{Sij} \perp \epsilon_{Sik}$ et $\epsilon_{Tij} \perp \epsilon_{Tik}$ pour $j \neq k$.

A la deuxième étape, Buyse et al. (2000) ont supposé que

$$\begin{pmatrix} \mu_{Si} \\ \mu_{Ti} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{Si} \\ m_{Ti} \\ a_i \\ b_i \end{pmatrix} \tag{2.21}$$

où le deuxième terme à gauche du modèle (2.21) représente des effet fixes et le terme à droite des effets aléatoires supposés gaussiens de moyenne nulle et de matrice de dispersion

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix} \tag{2.22}$$

Modèle linéaire à effets mixtes multivarié

Il est également possible de combiner les deux étapes du modèle précédent au sein d'un seul modèle linéaire à effets aléatoires comme suit:

$$S_{ij}|Z_{ij} = \mu_S + m_{Si} + \alpha Z_{ij} + a_i Z_{ij} + \epsilon_{Sij}, \tag{2.23}$$

$$T_{ij}|Z_{ij} = \mu_T + m_{Ti} + \beta Z_{ij} + b_i Z_{ij} + \epsilon_{Tij}, \tag{2.24}$$

où μ_S et μ_T sont des intercepts fixes, α et β sont les effets fixes du traitement Z sur S et T, m_{Si} et m_{Ti} sont des intercepts aléatoires, et a_i et b_i sont des effets aléatoires en interaction avec le traitement dans l'essai i. Dans ce modèle, le vecteur des effets aléatoires $(m_{Si}, m_{Ti}, a_i, b_i)$ est supposé gaussien de moyenne 0 et de matrice de dispersion D définie en (2.22). Les termes d'erreurs ϵ_{Si} et ϵ_{Ti} suivent la même hypothèse que celle présentée dans le modèle à effets fixes (2.18) − (2.19), avec pour matrice de variance-covariance Σ définie dans (2.20).

Sur la base des paramètres estimés à l'issue des modèles précédents, Buyse et al. (2000) ont proposé de nouvelles mesures d'association pour la validation au niveau individuel et au niveau essai d'un critère de substitution. Nous décrivons ces nouvelles quantités dans la suite de cette section.

Validation au niveau essai

La validation des critères de substitution est motivée par la capacité pour le critère de substitution de prédire l'effet du traitement sur le critère de jugement principal à partir de l'effet du traitement observé sur le critère de substitution. Sur cette base, Buyse et al. (2000) ont

proposé d'étudier la qualité de la prédiction de l'effet de Z sur T dans un nouvel essai $i = 0$ à partir : (a) des informations issues du processus de validation en utilisant les essais $i = 1 \cdots G$ ainsi que les modèles (2.18) - (2.22) ou (2.23) - (2.24) et (b) de l'estimation de l'effet de Z sur S dans l'essai $i = 0$

En ajustant les modèles hiérarchiques (2.18) - (2.22) ou le modèle à effets mixtes (2.23) - (2.24) aux données issues de méta-analyses, on obtient les paramètres du modèle et les composantes de variances. Supposons à présent que pour l'essai $i = 0$, on dispose des données sur S mais pas sur T. On peut ajuster le modèle linéaire suivant sur les données de l'essai 0 afin d'obtenir les estimations pour μ_{S0} et α_0:

$$S_{0j} = \mu_{S0} + \alpha_0 Z_{0j} + \epsilon_{S0j}. \tag{2.25}$$

Par ailleurs, partant des modèles (2.21) et (2.25) on a les estimations pour m_{S0} et a_0 :

$$\begin{aligned} \hat{m}_{S0} &= \hat{\mu}_{S0} - \hat{\mu}_S \\ \hat{a}_0 &= \hat{\alpha}_0 - \hat{\alpha}. \end{aligned}$$

Les auteurs se sont par la suite intéressés à l'estimation de l'effet de Z sur T sachant l'effet de Z sur S, c'est-à-dire $(\beta + b_0 | m_{S0}, a_0)$. On peut montrer que $(\beta + b_0 | m_{S0}, a_0)$ suit une distribution normale dont la moyenne et la variance sont données comme suit:

$$E(\beta + b_0 | m_{S0}, a_0) = \beta + \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}' \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{S0} - \mu_S \\ \alpha_0 - \alpha \end{pmatrix} \tag{2.26}$$

$$Var(\beta + b_0 | m_{S0}, a_0) = d_{bb} - \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}' \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_S \\ d_{ab} \end{pmatrix}, \tag{2.27}$$

où V' représente la transposée du vecteur (ou de la matrice) V.

Preuve:

Sachant que, $b_0 \sim N(0, d_{bb})$ et que

$$\begin{pmatrix} m_{S0} \\ a_0 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_{Sa} = \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix} \right];$$

alors,

$$\begin{bmatrix} (m_{S0}, a_0) \\ b_0 \end{bmatrix} \sim MVN\left[\begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{Sa} & \mathbf{B} \\ \mathbf{B} & d_{bb} \end{pmatrix} \right],$$

avec

$$\begin{aligned}\mathbf{B} &= Cov[b_0, (m_{S0}, a_0)] \\ &= [Cov(b_0, m_{S0}), Cov(b_0, a_0)] \\ &= (d_{bS}, d_{ba});\end{aligned}$$

où $Cov(X, Y)$ et $Var(X, Y)$ représentent la covariance et la variance des variables aléatoires X et Y. D'après les propriétés de la loi normale multivariée,

$$E(b_0|m_{S0}, a_0) = E(b_0) + Cov(b_0, (m_{S0}, a_0))Var(m_{S0}, a_0)^{-1} [(m_{S0}, a_0)' - E(m_{S0}, a_0)']$$

et

$$Var(b_0|m_{S0}, a_0) = Var(b_0) - Cov(b_0, (m_{S0}, a_0))Var(m_{S0}, a_0)^{-1} Cov((m_{S0}, a_0), b_0).$$

Par conséquent,

$$\begin{aligned}E(\beta + b_0|m_{S0}, a_0) &= \beta + E(b_0|m_{S0}, a_0) \\ &= \beta + (d_{bS}, d_{ba})\Sigma_{Sa}^{-1}(m_{S0}, a_0)' \\ &= \beta + \begin{bmatrix} d_{Sb} \\ d_{ab} \end{bmatrix}' \begin{bmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{bmatrix}^{-1} \begin{bmatrix} \mu_{S0} - \mu_S \\ \alpha_0 - \alpha \end{bmatrix}\end{aligned}$$

et

$$\begin{aligned}Var(\beta + b_0|m_{S0}, a_0) &= Var(\beta) + Var(b_0) - \mathbf{B}Var(m_{S0}, a_0)^{-1}\mathbf{B}' \\ &= 0 + Var(b_0) - \mathbf{B}Var(m_{S0}, a_0)^{-1}\mathbf{B}' \\ &= d_{bb} - \begin{bmatrix} d_{Sb} \\ d_{ab} \end{bmatrix}' \begin{bmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{bmatrix}^{-1} \begin{bmatrix} d_{Sb} \\ d_{ab} \end{bmatrix}\end{aligned}$$

Comme remarqué par Buyse et al. (2000), un critère de substitution sera considéré comme parfait au niveau essai si la variance conditionnelle de l'effet estimé du traitement Z sur T donnée dans l'équation (2.27) est égale à zéro. Ce qui équivaut à :

$$\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}' \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_S \\ d_{ab} \end{pmatrix} = d_{bb}$$

Ainsi, S est un critère de substitution parfait pour T au niveau essai si

$$\frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}' \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_S \\ d_{ab} \end{pmatrix}}{d_{bb}} = 1 \tag{2.28}$$

L'expression à gauche de (2.28) n'est rien d'autre que le coefficient de détermination du modèle de régression entre les effets de Z sur T et S $(b_i|m_{Si}, a_i)$. Ainsi, la mesure permettant d'évaluer la qualité d'un critère de substitution au niveau essai est le coefficient de détermination

$$R^2_{trial(f)} = R^2_{b_i|m_{Si},a_i} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}' \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_S \\ d_{ab} \end{pmatrix}}{d_{bb}}. \tag{2.29}$$

Buyse et al. (2000) suggèrent un gain en intuition en considérant le cas où la prédiction de b_0 se fait indépendamment de l'intercept aléatoire m_{S0}. Les expressions (2.26) et (2.27) se réduisent dans ce cas à

$$\begin{aligned} E(\beta + b_0|a_0) &= \beta + \frac{d_{ab}}{d_{aa}}(\alpha_0 - \alpha) \\ Var(\beta + b_0|a_0) &= d_{bb} - \frac{d^2_{ab}}{d_{aa}} \end{aligned}$$

et

$$R^2_{trial(r)} = R^2_{b_i|a_i} = \frac{d^2_{ab}}{d_{aa}d_{bb}} \tag{2.30}$$

$R^2_{trial(r)}$ ainsi obtenu sera considéré comme issu du modèle à effets aléatoires réduit, tandis que $R^2_{trial(f)}$ dans (2.29) est considéré comme issu du modèle complet.

Remarque : r fait référence au terme anglais `reduce`, pour modèle réduit et f au terme anglais `full`, pour modèle complet.

Validation au niveau individuel

Afin de valider le critère de substitution au niveau individuel, Buyse et al. (2000) suggèrent de considérer tout comme Buyse and Molenberghs (1998) l'association entre les deux critères de jugement après ajustement sur les effets du traitement. Pour cela, il est nécessaire de construire la distribution conditionnelle de T sachant Z et S. A partir des modèles marginaux (2.18) et

(2.19), on peut montrer que la distribution conditionnelle de T_{ij}, sachant S_{ij} et Z_{ij} est :

$$\begin{aligned} T_{ij}|Z_{ij}, S_{ij} \sim & N\{\mu_{Ti} - \sigma_{TS}\sigma_{SS}^{-1}\mu_{Si} + (\beta_i - \sigma_{TS}\sigma_{SS}^{-1}\alpha_i)Z_{ij} + \sigma_{TS}\sigma_{SS}^{-1}S_{ij}; \\ & \sigma_{TT} - \sigma_{TS}^2\sigma_{SS}^{-1}\}. \end{aligned} \tag{2.31}$$

En effet à partir de (2.18) et (2.19),

$$\begin{aligned} S_{ij}|Z_{ij} &\sim N[(\mu_{Si} + \alpha_i Z_{ij}), \sigma_{SS}] \\ T_{ij}|Z_{ij} &\sim N[(\mu_{Ti} + \beta_i Z_{ij}), \sigma_{TT}] \end{aligned}$$

$$\Rightarrow \begin{pmatrix} S_{ij} \\ T_{ij} \end{pmatrix} \sim MNV \left[\mathbf{M} = \begin{pmatrix} \mu_{Si} + \alpha_i Z_{ij} \\ \mu_{Ti} + \beta_i Z_{ij} \end{pmatrix}, \mathbf{V} = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \sigma_{ST} & \sigma_{TT} \end{pmatrix} \right].$$

Suivant les propriétés de la loi normale multivariée,

$$T_{ij}|S_{ij}, Z_{ij} \sim N[E(T_{ij}|S_{ij}, Z_{ij}), Var(T_{ij}|S_{ij}, Z_{ij})],$$

avec

$$\begin{aligned} E(T_{ij}|S_{ij}, Z_{ij}) = E(T_{ij}) + Cov(T_{ij}, (S_{ij}, Z_{ij}))Var(S_{ij}, Z_{ij})^{-1} \\ [(S_{ij}, Z_{ij})' - E(S_{ij}, Z_{ij})'] \end{aligned}$$

et

$$\begin{aligned} Var(T_{ij}|S_{ij}, Z_{ij}) = Var(T_{ij}) - Cov(T_{ij}, (S_{ij}, Z_{ij}))Var(S_{ij}, Z_{ij})^{-1} \\ Cov(T_{ij}, (S_{ij}, Z_{ij})). \end{aligned}$$

Par conséquent,

$$\begin{aligned} E(T_{ij}|S_{ij}, Z_{ij}) &= \mu_{Ti} + \beta_i Z_{ij} + (\sigma_{TS}, 0)\sigma_{SS}^{-1}(S_{ij} - \mu_{Si} - \alpha_i Z_{ij}, 0) \\ &= \mu_{Ti} + \beta_i Z_{ij} + \sigma_{TS}\sigma_{SS}^{-1}(S_{ij} - \mu_{Si} - \alpha_i Z_{ij}) \\ &= \mu_{Ti} - \sigma_{TS}\sigma_{SS}^{-1}\mu_{Si} + (\beta_i - \sigma_{TS}\sigma_{SS}^{-1}\alpha_i)Z_{ij} + \sigma_{TS}\sigma_{SS}^{-1}S_{ij} \end{aligned}$$

et

$$\begin{aligned} Var(T_{ij}|S_{ij}, Z_{ij}) &= \sigma_{TT} - \sigma_{TS}\sigma_{SS}^{-1}\sigma_{TS} \\ &= \sigma_{TT} - \sigma_{TS}^2\sigma_{SS}^{-1}. \end{aligned}$$

De même, partant du modèle linéaire à effets mixtes, (2.23)−(2.24) on a cette autre distribution:

$$\begin{aligned} T_{ij}|Z_{ij}, S_{ij} \sim & N\{\mu_T + m_{Ti} - \sigma_{TS}\sigma_{SS}^{-1}(\mu_S + m_{Si}) + [\beta + b_i - \sigma_{TS}\sigma_{SS}^{-1}(\alpha + a_i)]Z_{ij} + \\ & \sigma_{TS}\sigma_{SS}^{-1}S_{ij}; \quad \sigma_{TT} - \sigma_{TS}^2\sigma_{SS}^{-1}\}, \end{aligned} \tag{2.32}$$

où on conditionne aussi sur les effets aléatoires. Il s'en suit que S est un parfait critère de substitution pour T si $Var(T_{ij}|S_{ij}, Z_{ij}) = 0$. L'association entre T et S après ajustement sur les effets de Z est capturée aussi bien dans (2.31) que dans (2.32) par

$$R^2_{indiv} = R^2_{\epsilon_{Ti}|\epsilon_{Si}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}}, \tag{2.33}$$

le carré du coefficient de corrélation entre les variables ajustées $S_{ij} - (\mu_{Si} + \alpha_i Z_{ij})$ et $T_{ij} - (\mu_{Ti} + \beta_i Z_{ij})$.

Sur la base des développements précédents, Buyse et al. (2000) ont suggéré de considérer S comme valide au niveau essai si $R^2_{trial(f)}$ ou $R^2_{trial(r)}$ sont suffisamment proches de 1, de même S sera considéré comme valide au niveau individuel si R^2_{indiv} est suffisamment proche de 1. S sera considéré comme valide s'il l'est au niveau individuel et au niveau essai. Il est important de noter que dans cette nouvelle formulation des recommandations pour la validation des critères de substitution, il n'est plus nécessaire d'avoir des effets significatifs du traitement sur S ou sur T pour avoir un critère de substitution valide. En particulier, bien que très rare en pratique (Buyse et al. 2000), il est possible d'observer $\alpha \equiv 0$ et d'avoir un critère de substitution parfait. En effet, même si le traitement n'a aucun effet sur le critère de substitution dans son ensemble, les fluctuations autour de zéro dans les différents essais peuvent être très fortement prédictives de l'effet du traitement sur le critère principal.

Pour être utile en pratique, un critère de substitution valide doit être capable de prédire l'effet du traitement sur le critère principal avec une précision suffisante pour permettre une distinction en toute sécurité entre les effets cliniquement intéressants et les effets qui ne le sont pas. Cela nécessite que l'estimation de $\beta + b_0$ soit suffisamment grande et que son intervalle de prédiction soit suffisamment étroit.

2.3.2 Cas de deux temps d'évènements

Dans la suite de ce document nous considérons que S et T sont des temps d'évènements. Ainsi, toutes les méthodes décrites par la suite se situeront dans ce contexte.

Description de l'approche de copule en deux étapes

Burzykowski et al. (2001) ont proposé d'étendre le modèle hiérarchique à deux niveaux de Buyse et al. (2000) au cas où S_{ij} et T_{ij} sont des temps d'évènements. Il s'agit de la méthode standard et celle la plus utilisée dans la littérature dans ce contexte (Savina et al. 2018; Branchoux et al. 2019). Dans la première étape, ils ont remplacé le modèle conjoint (2.18) - (2.19) par un modèle pour deux variables corrélées à temps d'évènements. Burzykowski et al. (2001) se sont appuyés sur un modèle de copule bivarié pour définir la fonction de survie conjointe de (S_{ij}, T_{ij}) :

$$\begin{aligned} F(s,t) &= P(S_{ij} \geq s, T_{ij} \geq t) \\ &= C_\theta(F_{Sij}(s), F_{Tij}(t)), \qquad s,t \geq 0 \end{aligned} \tag{2.34}$$

où $(F_{Sij}(s), F_{Tij}(t))$ représentent les fonctions de survie marginales et C_θ est une fonction de distribution bivariée sur $[0,1]^2$, avec $\theta \in R$. C_θ est une fonction de copule permettant de décrire l'association entre S_{ij} et T_{ij}. Parmi les fonctions de copules les plus usuelles, Burzykowski et al. (2001) ont considéré par simplicité les copules de Clayton, de Gumbel-Hougaard et de Plackett. Dans le modèle de Clayton, la fonction de copule a la forme

$$C_\theta(u,v) = (u^{1-\theta} + v^{1-\theta} - 1)^{\frac{1}{1-\theta}}, \qquad \theta > 1$$

Cela implique une association positive; la force de l'association diminue avec la diminution de θ et atteint l'indépendance lorsque $\theta \to 1$. Dans le modèle de Hougaard, la fonction de copule a la forme

$$C_\theta(u,v) = \exp[-\{(-logu)^{\frac{1}{\theta}}\} + (-logv)^{\frac{1}{\theta}}\}^\theta], \qquad 0 < \theta < 1$$

Elle induit une association positive entre les temps d'évènements; la force de l'association diminue avec l'augmentation de θ et atteint l'indépendance lorsque $\theta \to 1$. Dans le modèle de Plackett par contre, la fonction de copule est définie comme suit :

$$C_\theta(u,v) = \begin{cases} \frac{1+(u+v)(\theta-1)-H_\theta(u,v)}{2(\theta-1)} & \text{si } \theta \neq 1 \\ uv & sinon \end{cases}$$

où $H_\theta(u,v) = \sqrt{[1+(\theta-1)(u+v)]^2 + \theta(1+\theta)uv}$ et $\theta \in [0,\infty]$. On a une indépendance entre S et T lorsque $\theta = 1$.

Afin de modéliser les effets du traitement sur les distributions marginales de S_{ij} et T_{ij} dans

l'équation (2.34), les modèles à risque proportionnel ont été utilisés :

$$F_{Sij}(s) = \exp\{-\int_0^s \lambda_{Si}(x)\exp(\alpha_i Z_{ij}dx\} \tag{2.35}$$

$$F_{Tij}(t) = \exp\{-\int_0^{'} \lambda_{Ti}(x)\exp(\beta_i Z_{ij}dx\}, \tag{2.36}$$

où λ_{Si} et $\lambda_{Ti}(x)$ sont des fonctions de risques de base spécifiques aux essais. Dans leur approche Burzykowski et al. (2001) font l'hypothèse que la distribution des temps de survie appartient à une famille paramétrique, ce qui permet de spécifier les fonctions de risque de base de façon paramétrique. La distribution de Weibull est celle utilisée dans ce cas.

Dans la deuxième étape, pour valider au niveau essai les critères de substitution, Burzykowski et al. (2001) ont proposé d'utiliser le modèle à effets aléatoires réduit :

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} a_i \\ b_i \end{pmatrix} \tag{2.37}$$

où le deuxième terme à droite de l'équation (2.37) représente des effets aléatoires supposés gaussiens de moyenne nulle et de matrice de dispersion

$$D = \begin{pmatrix} d_{aa} & d_{ab} \\ d_{ab} & d_{bb} \end{pmatrix} \tag{2.38}$$

Validation de la surrogacy au niveau individuel

Pour un modèle de copule particulier, la force de l'association entre S_{ij} et T_{ij}, après ajustement de leurs distributions marginales sur l'essai et les effets du traitement, dépend de θ. Ainsi, θ peut être considéré comme un candidat naturel pour la mesure d'association recherchée. Compte tenu de la difficulté à interpréter ou à comparer directement θ pour différents modèles, Burzykowski et al. (2001) ont proposé de travailler avec une transformation de θ et donc du tau de Kendall (τ). Pour le modèle de copule (2.34) On peut établir le lien suivant entre le τ et θ (Duchateau and Janssen 2008) :

$$\tau = 4\int_0^1 \int_0^1 C_\theta(u,v)C_\theta(du,dv) - 1. \tag{2.39}$$

Le tau de Kendall est la différence entre la probabilité de concordance et la probabilité de discordance de deux réalisations de (S_{ij}, T_{ij}). Il est compris dans l'intervalle [-1, 1] et vaut 0 en cas d'indépendance entre S_{ij} et T_{ij}.

La relation entre τ et θ est particulièrement simple dans les modèles de copule de Clayton

et de Gumbel. Pour les modèles de Clayton, $\tau = (\theta - 1)/(\theta + 1)$, alors que pour le modèle de Gumbel, $\tau = 1-\theta$. Comme nous le verrons au Chapitre 5, à partir d'un changement de variable, on peut avoir une relation différente entre τ et θ pour les fonctions de copules considérées. Ces relations permettent de construire une estimation $\hat{\tau}$ de τ, à partir de l'estimation $\hat{\theta}$ de θ, où $\hat{\tau}$ et $\hat{\theta}$ sont obtenus par maximisation de la vraisemblance du modèle (2.34). Une mesure de dépendance alternative entre S_{ij} et T_{ij} est le coefficient de corrélation de Spearman noté ρ. Pour une fonction de copule bivariée $C(u, v)$, on montre que ρ peut s'écrire comme (Fredricks and Nelsen 2007) :

$$\rho = 12 \int_0^1 \int_0^1 C_\theta(u, v)(du, dv) - 3$$

Toutefois, la mesure d'association que nous considérons dans le cadre de cette thèse est le τ de Kendall.

Validation de la surrogacy au niveau essai

Avec l'utilisation du modèle (2.37) dans la deuxième étape de l'approche de Burzykowski et al. (2001), la qualité du critère de substitution S au niveau essai est évaluée à partir du coefficient de détermination $\mathrm{R}^2_{trial(r)}$ défini plus haut (2.30). Toutefois, l'usage en pratique pose un problème de biais que nous illustrons de façon analytique par la suite.

Biais dans l'estimation du $\mathbf{R}^2_{trial(r)}$

En pratique, seules les estimations $\hat{\alpha}_i$ et $\hat{\beta}_i$, obtenues à partir du modèle de copule à la première étape (2.34) - (2.36), sont disponibles. On sait qu'en ignorant les erreurs de mesure lors de l'ajustement des modèles de régression, les coefficients estimés des modèles peuvent être biaisés comme il a été remarqué par Burzykowski et al. (2005). De plus, traiter $R^2_{trial(r)}$ comme si les estimations $\hat{\alpha}_i$ et $\hat{\beta}_i$ étaient égales aux vrais effets du traitement non observés, induirait un biais dans son estimation. En effet, partant des estimations du modèle de première étape (2.34)-(2.36), on peut écrire le modèle suivant pour tenir compte des erreurs d'estimations :

$$\begin{pmatrix} \hat{\alpha}_i \\ \hat{\beta}_i \end{pmatrix} = \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} + \begin{pmatrix} \epsilon_{ai} \\ \epsilon_{bi} \end{pmatrix} \tag{2.40}$$

où $(\alpha_i, \beta_i)'$ viennent de (2.35)-(2.36) et les erreurs d'estimations ϵ_{ai} et ϵ_{bi} sont normalement distribuées de moyenne nulle et de matrice de covariance:

$$\Omega_i = \begin{pmatrix} \sigma_{aa,i} & \sigma_{ab,i} \\ \sigma_{ab,i} & \sigma_{bb,i} \end{pmatrix}, \tag{2.41}$$

et $(\alpha_i, \beta_i)'$ sont définis comme dans le modèle (2.37) avec pour matrice de dispersion D donnée par (2.22). Par conséquent, $(\hat{\alpha}_i, \hat{\beta}_i)'$ suivent une distribution normale de moyenne $(\alpha, \beta)'$ et de matrice de dispersion $D + \Omega_i$. Toujours pour montrer la nécessité de prendre en compte les erreurs d'estimation, Burzykowski et al. (2005) ont considéré le cas où $\Omega_i = \Omega$, avec

$$\Omega = \begin{pmatrix} \sigma_{aa} & \sigma_{ab} \\ \sigma_{ab} & \sigma_{bb} \end{pmatrix} \tag{2.42}$$

et noté par ρ la corrélation qui découle de Ω, c'est-à-dire $\rho = Corr(\epsilon_a, \epsilon_b)$. Nous montrons dans l'Annexe que la corrélation entre $\hat{\alpha}_i$ et $\hat{\beta}_i$ peut prendre la forme :

$$Corr(\hat{\alpha}_i, \hat{\beta}_i) = \frac{Corr(\alpha_i, \beta_i)}{\sqrt{(1+k_a)(1+k_b)}} + \frac{\rho}{\sqrt{(1+k_a^{-1})(1+k_b^{-1})}}$$

où $k_a = \sigma_{aa}/d_{aa}$ et $k_b = \sigma_{bb}/d_{bb}$ désignent les rapports de fiabilité pour $\hat{\alpha}_i$ et $\hat{\beta}_i$. Il s'en suit que lorsque les erreurs d'estimation sont indépendantes ($\rho = 0$) $R^2_{trial(r)}$ est sous-estimé, tandis que $\rho \neq 0$ impliquerait une sous-estimation ou une surestimation de $R^2_{trial(r)}$.

Prise en compte des erreurs d'estimation et définition du $R^2_{trial(r)}$ ajusté (adjR^2_{trial})

Afin de prendre en compte les erreurs d'estimation sur $\hat{\alpha}_i$ et $\hat{\beta}_i$, dans l'estimation du $R^2_{trial(r)}$, Burzykowski et al. (2001), ont considéré une approche basée sur les développements de van Houwelingen et al. (2002). Plus précisément, la matrice de dispersion D, définie par 2.37 peut être obtenue en implémentant les modèles (2.40)-(2.41) et (2.37)-(2.38) sur les paires estimées $(\hat{\alpha}_i, \hat{\beta}_i)$. Afin d'ajuster le modèle, les matrices de covariance Ω_i sont supposées connues et égales à leurs estimations obtenues à partir du modèle de copule bivarié (2.34). Par conséquent, une estimation $\hat{R}^2_{trial(r)}$ de $R^2_{trial(r)}$ trial peut être obtenue à partir de l'estimation résultante $\hat{D}$ de D au moyen du coefficient de détermination ajusté:

$$adjR^2_{trial} = \hat{R}^2_{trial(r)} = \frac{\hat{d}^2_{ab}}{\hat{d}_{aa}\hat{d}_{bb}}. \tag{2.43}$$

Comme souligné par Burzykowski et al. (2005), la convergence du modèle permettant l'estimation de $\hat{R}^2_{trial(r)}$ dans (2.43) n'est pas toujours garantie.

Une alternative a été proposée pour tenir compte des biais dans l'estimation de $R^2_{trial(r)}$ et est discutée par Burzykowski et al. (2005) dans la section 11.2.1. Il ressort de cette approche une difficulté supplémentaire dans l'utilisation de la méthode en deux étapes de Burzykowski

et al. (2001) pour la validation des critères de substitution à temps d'évènement. En effet, l'approche alternative décrite ne garantit pas la positivité de l'estimation de la variance des erreurs résiduelles dans le modèle de régression linéaire de β_i sur α_i ($\beta_i = \gamma_0 + \gamma_1\alpha_i + \epsilon_i$), mais également la variance de $\hat{\alpha}_i$ après ajustement sur les erreurs d'estimation ($\hat{d}_{aa}$).

C'est partant de ces difficultés liées à la convergence du modèle ainsi que des problèmes d'estimation du $R^2_{trial(r)}$ ajusté que nous proposons dans cette thèse une nouvelle approche de validation des critères de substitution en une étape, basée sur des modèles conjoints à fragilités et à copules. Dans la deuxième section de ce chapitre nous présentons une revue sommaire des modèles conjoints à fragilités ayant motivé la spécification de nos modèles.

2.4 Autres approches de validation

Parallèlement aux approches méta-analytiques développées pour la validation des critères de substitution un autre paradigme de validation s'appuyant sur des méthodes d'inférence causale, à l'instar de l'analyse de médiation est en pleine expansion dans la littérature (Alonso et al. 2019, 2016; Vandenberghe et al. 2018; Buyse et al. 2016). La particularité de cette nouvelle approche réside dans la possibilité d'avoir une interprétation causale de la relation entre le critère de substitution et le critère de jugement principale. L'analyse de médiation vise à étudier l'effet d'un traitement ou d'une exposition sur un critère de jugement lorsqu'une partie de cet effet passe par l'activation d'un facteur de médiation (Pearl 2001). L'effet total de ce traitement peut ainsi se décomposer en deux effets : un effet direct indépendant du facteur de médiation et un effet indirect lié à l'effet du traitement sur ce facteur de médiation (Taylor et al. 2005) tel qu'illusté dans la Figure 2.2.

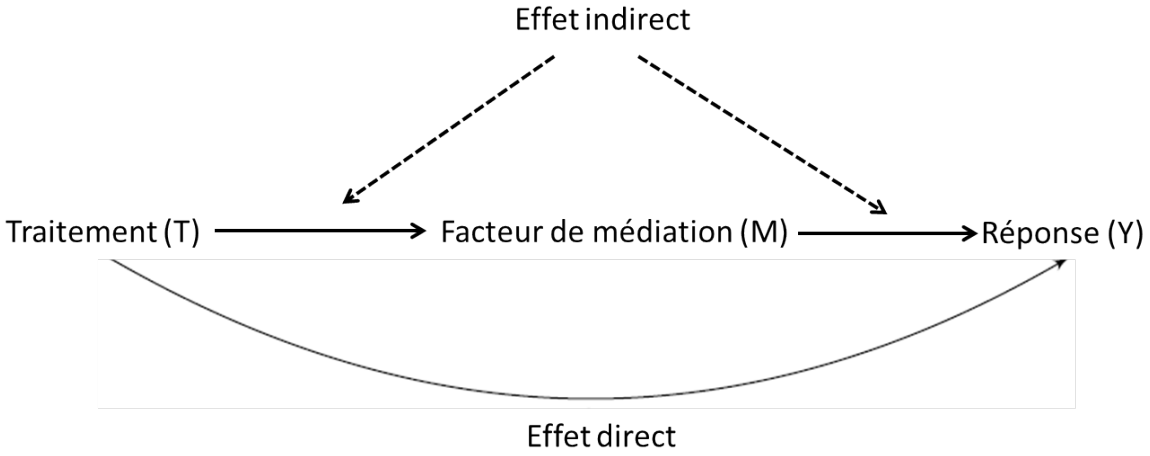


Figure 2.2: Décomposition de l'effet total du traitement sur la réponse en un effet indirect passant par le facteur de médiation et un effet direct indépendant de M.

Avec une telle décomposition, si nous considérons le facteur de médiation comme état le critère de substitution, plus la proportion de l'effet total du traitement passant par le critère de substitution sera importante meilleure sera la qualité du critère de substitution. En générale un critère de substitution sera considéré comme valide si le rapport `effet indirect / effet total` est proche de `1`. Le lecteur pourra se référer aux travaux de (Holland 1986) pour la méthode permettant de déterminer les effets directs et indirects dans les données d'essais cliniques.

Alonso et al. (2019) dans la continuité de ce qui est présenté ci-dessus ont proposé une définition alternative de la mesure d'association basée sur les effets causaux du traitement au niveau individuel. Cet autre concept permet d'apporter une interprétation causale à la mesure d'association au niveau individuel. Nous n'entrons pas en profondeur dans la description des approches causales compte tenu du cadre de développement des méthodes dans cette thèse. Toutefois, le lecteur pourra suivre les références suggérées dans cette section pour aller plus loin dans cette approche.

B Modèles à fragilités

2.5 Modèles à fragilités partagées

2.5.1 Type de données

La corrélation des temps de survie peut intervenir lorsque des individus appartiennent à des groupes, tels que des familles, des centres, des essais cliniques ou des zones géographiques. Alternativement, une corrélation peut être associée à des événements récurrents, lorsqu'un sujet subit un même évènement plusieurs fois, telles que des réhospitalisations ou des attaques cardiaques. Dans ces deux situations, on pourrait ajuster directement sur les variables de regroupement en effet fixes (exemple: zone géographique quand données groupées ou sujet lui-même si données récurrentes). Cependant, comme discuté dans Commenges and Jacqmin-Gadda (2015), les inconvénients à l'utilisation de cette approche sont de plusieurs ordres: le nombre de paramètres peut être très élevé lorsque le nombre de groupes est grand, ce qui peut conduire à des problèmes numériques; cette approche nécessite d'avoir dans un modèle à effets fixes, au moins un évènement par groupe de sujets. Il s'agit d'une approche adaptée si la taille de l'échantillon est élevée et avec des tailles de groupes élevées et peu de groupes.

Les données de survies environnementales conduisent très souvent à des analyses statistiques particulières. En effet, ces données sont souvent regroupées en zones géographiques (commune, ville, état, pays) et il est fréquent que les sujets d'une même zone partagent des facteurs de

risque non identifiés (génétiques, environnementaux). Les facteurs de risque non observés ou non mesurés et partagés par un groupe vont créer une dépendance des évènements étudiés dans chaque groupe. Dans les analyses de survie, si cette hétérogénéité non observée est ignorée, elle peut créer un biais important dans l'estimation de la variance des paramètres de régression et sur l'estimation de la fonction de risque. Une description plus complète concernant les données récurrentes peut être consultée dans Commenges and Jacqmin-Gadda (2015).

2.5.2 Définition du modèle

Le modèle à fragilités est une extension du modèle àrisque proportionnel dans lequel on rajoute une variable de fragilité u_i (ou effet aléatoire) spécifique à chaque groupe i ou à chaque individu i dans le cadre des données récurrentes. Dans le contexte de données groupées, les observations pour chaque sujet j, $j = 1, \cdots, n_i$ du groupe i ($i = 1, \cdots, G$) sont recueillies; dans le contexte des données récurrentes, chaque sujet i a n_i temps de récurrence, et n_i est aléatoire. On observe le temps $\tilde{T}_{ij}$, qui est le minimum entre un temps de censure C_{ij} et un temps d'évènements T_{ij}, et δ_{ij}, un indicateur de censure qui vaut 1 si $T_{ij} \leq C_{ij}$. Les T_{ij} et C_{ij} sont supposés indépendants. La fonction de risque conditionnelle aux fragilités u_i (identiquement et indépendamment distribuée) pour le sujet j du groupe i s'exprime alors par

$$\lambda_{ij}(t|X_{ij}, u_i) = u_i \lambda_0(t) \exp(\beta' X_{ij}) \tag{2.44}$$

où X_{ij} représente le vecteur de variables explicatives, β le vecteur de dimension p des effets fixes, $\lambda_0(t)$ est la fonction de risque de base et u_i est la variable de fragilité spécifique à chaque groupe. u_i prend en compte l'ensemble des variables explicatives non observées dans le jeu de données, et communes à un même groupe. Dans ce modèle les temps de survie sont supposés indépendants entre les groupes. De même conditionnellement aux u_i, on a une indépendance des temps de survie dans chaque groupe. Le modèle (2.44) est appelé modèle à fragilités partagées puisque tous les individus d'un même groupe partagent la même fragilité. Il peut se présenter sous la forme :

$$\lambda_{ij}(t|X_{ij}, \omega_i) = \lambda_0(t) \exp(\beta' X_{ij} + \omega_i) \tag{2.45}$$

où ω_i est un effet aléatoire, contrairement à $u_i = \exp(\omega_i)$ qui est définie comme une variable de fragilité (Duchateau and Janssen 2008). Tandis que les u_i sont supposées suivre une distribution gamma dans le modèle (2.44), on a plutôt une distribution log-normale des u_i avec le modèle (2.45). La distribution gamma permet des simplifications importantes au moment de la construction de la log-vraisemblance du modèle. Il est à noter que d'autres formes de distributions peuvent être considérées.

2.5.3 Estimation des paramètres et vraisemblance pénalisée

Supposons les données de survie censurées à droite et considérons le modèle (2.44). La contribution marginale à la vraisemblance du groupe i est donnée par:

$$L_i(\lambda_0(.),\beta,\theta) = \int_{u_i} L_i(\lambda_0(.),\beta,\theta|u_i)f(u_i)du_i \tag{2.46}$$

où la densité de probabilité des fragilités u_i (en supposant que $u_i \sim \Gamma(1/\theta,\theta)$) est donnée par

$$f(u_i) = \frac{u_i^{(1/\theta-1)}\exp(-u_i/\theta)}{\Gamma(1/\theta)\theta^{1/\theta}}, \quad u_i > 0, \quad \theta > 0 \tag{2.47}$$

et

$$L_i(\lambda_0(.),\beta,\theta|u_i) = \prod_{j=1}^{n_i} \lambda_{ij}(\tilde{T}_{ij}|u_i)^{\delta_{ij}}\exp(-\Lambda_{ij}(\tilde{T}_{ij}|u_i)) \tag{2.48}$$

représente la contribution conditionnelle aux fragilités u_i des individus du groupe i à la vraisemblance marginale.

A partir des expressions (2.46) à (2.48), on obtient l'expression de la log-vraisemblance marginale qui est donnée par:

$$\begin{aligned} l(\lambda_0(.),\beta,\theta|u_i) = \sum_{i=1}^{G}\Bigg\{ & \sum_{j=1}^{n_i}\delta_{ij}\{\beta' X_{ij} + \ln(\lambda_0(\tilde{T}_{ij}))\} \\ & -(1/\theta + m_i)\ln\left[1+\theta\sum_{j=1}^{n_i}\Lambda_0(\tilde{T}_{ij})\exp(\beta' X_{ij})\right] \\ & + I_{\{m_i\neq 0\}}\sum_{k=1}^{m_i}[\ln(1+\theta(m_i-k))]\Bigg\} \end{aligned} \tag{2.49}$$

avec

$$I_{\{m_i\neq 0\}}\sum_{k=1}^{m_i}[\ln(1+\theta(m_i-k))] = m_i\ln(\theta) + \ln(\Gamma(1/\theta+m_i) - \ln(\Gamma(1/\theta)))$$

et $m_i = \sum_{j=1}^{n_i} I_{\{\delta_{ij}=1\}}$ le nombre d'évènements observés dans le i^{ime} groupe.

Afin d'obtenir des fonctions de risques lisses dans des modèles à fragilités partagées, Rondeau et al. (2003) ont proposé une méthode semi-paramétrique par vraisemblance pénalisée pour l'estimation des paramètres et des fonctions de risque de base des modèles (2.44) et (2.45). Cette approche par vraisemblance pénalisée permet de contraindre l'estimateur de la fonction

de risque de base à être continue et à avoir de faibles variations locales. La log-vraisemblance pénalisée est définie comme suit:

$$l_{pl}(\lambda_0(.), \beta, \theta) = l(\lambda_0(.), \beta, \theta) - \kappa \int_0^\infty \lambda_0''^2(t)dt \tag{2.50}$$

où $\kappa \geq 0$ est un paramètre de lissage, qui peut être obtenu par validation croisée (O'Sullivan 1988; Joly et al. 1998). Une autre approche consiste à fixer le nombre de degrés de liberté du modèle et à en déduire le paramètre de lissage en utilisant la relation qui lie ces deux valeurs. $\lambda_0''^2(t)$ représente la dérivée seconde de la fonction de risque de base $\lambda_0(t)$.

La maximisation de (2.50) définit les estimateurs du maximum de vraisemblance pénalisée (MPnLE), $\hat{\lambda}_0(.)$, $\hat{\beta}$, $\hat{\theta}$. L'estimateur de la variance des paramètres peut être obtenu directement par H_{pl}^{-1}, où H_{pl} est la hessienne de la log-vraisemblance pénalisée. Les estimateurs des fonctions de risque de base $\hat{\lambda}_0(.)$ ne peuvent pas être calculés explicitement et sont approchés sur une base de M-splines cubiques (Ramsay 1988). Des I-splines (Integrated splines) sont utilisés pour estimer les fonctions de risque cumulé. La maximisation de la log-vraisemblance pénalisée est faite grâce à l'algorithme de Marquardt (Marquardt 1963). Par ailleurs, la méthode des différences finies est utilisée pour le calcul numérique des dérivées premières et secondes de la log-vraisemblance (2.50) ou (2.49), à partir desquelles les hessiennes et donc les matrices de variances-covariances asymptotiques sont déduites.

Une autre approche d'estimation par vraisemblance partielle pénalisée dans laquelle le terme de pénalisation porte sur la distribution des effets aléatoires est décrite dans Commenges and Jacqmin-Gadda (2015). On y retrouve également une description de l'estimation des effets aléatoires.

2.6 Modèles conjoints à fragilités

Le modèle à fragilités présenté dans la section précédente peut être étendu au cas où on a une censure informative de l'évènement d'intérêt par un évènement terminal comme le décès. Dans ce cas, l'hypothèse de censure non informative de l'évènement d'intérêt par le décès n'est plus valide et il convient donc de la prendre en compte. Pour ce faire, il est recommandé de modéliser conjointement les deux temps d'évènements à l'aide d'autres approches, connexes à celles proposées dans la suite de ce chapitre.

2.6.1 Définition du modèle

2.6.2 Modèle conjoint pour temps jusqu'à la progression et le décès

Le modèle décrit dans cette section a été proposé par Rondeau et al. (2015) pour deux temps d'évènements dans un contexte de méta-analyse. Soit une étude avec G groupes indépendants ($i = 1, \ldots, G$). Notons X_{ij} le j^{ime} temps d'évènements (ou TTP) pour le sujet j ($j = 1, \ldots, N_i$) du groupe i, C_{ij} le temps de censure (différent du décès) correspondant et D_{ij} le temps de décès. Chaque temps de suivi correspond à $T_{ij} = min(X_{ij}, C_{ij}, D_{ij})$ et δ_{ij} son indicateur binaire de censure, qui vaut 0 si le sujet est décédé ou si l'observation est censurée, et 1 si le temps X_{ij} est observé ($\delta_{ij} = I_{(T_{ij}=X_{ij})}$, où $I_{()}$ représente la fonction indicatrice). De la même manière, on note T^*_{ij} le dernier temps de suivi pour le sujet j, qui est soit un temps de censure soit un temps de décès ($T^*_{ij} = min(C_{ij}, D_{ij})$ et $\delta^*_{ij} = I(T^*_{ij} = D_{ij})$. Ce qu'on observe réellement est $(T_{ij}, T^*_{ij}, \delta_{ij}, \delta^*_{ij})$. Les auteurs supposent que le décès et la progression ne peuvent pas se produire en même temps. Par conséquent, le décès arrive en premier dans un petit intervalle $[t, t + dt[$. Ainsi, pour un sujet qui subit une progression le même jour que le décès, ils ne comptent que pour un événement terminal, pas comme une progression. Rondeau et al. (2015) ont proposé deux formulations de la modélisation conjointe du TTP et de la survie globale (OS).

Dans le premier modèle, ils supposent que l'association entre TTP et OS est simplement le résultat des associations individuelles ou des facteurs individuels non mesurés. Ils considèrent des effets aléatoires individuels non observés ω_{ij} pour la prise en compte de cette hétérogénéité entre les sujets. Suivant le modèle conjoint pour données récurrentes et à temps d'évènements précédemment proposé par Rondeau et al. (2007), ils ont défini le modèle conjoint pour les fonctions de risque de progression ($r_{ij}(.)$) et de décès ($\lambda_{ij}(.)$) comme suit:

$$\begin{cases} r_{ij}(t|\omega_{ij}, \mathbf{Z_{ij}}) = \omega_{ij} r_0(t) \exp(\alpha Z_{ij0} + \sum_{k=1}^{p} \gamma_{1k} Z_{ijk}(t)) = \omega_{ij} r_{ij}(t) \\ \lambda_{ij}(t|\omega_{ij}, \mathbf{Z_{ij}}) = \omega_{ij}^{\zeta} \lambda_0(t) \exp(\beta Z_{ij0} + \sum_{k=1}^{p} \gamma_{2k} Z_{ijk}(t)) = \omega_{ij}^{\zeta} \lambda_{ij}(t) \end{cases} \tag{2.51}$$

dans lequel, Z_{ij0} est une variable binaire représentant le bras de traitement dans lequel le patient a été randomisé et $Z_{ijk}(t)$ ($k = 1, \ldots, p$) le vecteur des facteurs pronostiques, supposés le même pour les deux critères de jugement. Les effets aléatoires ω_{ij} (terme de fragilité) sont supposés indépendants, et suivent une distribution gamma de moyenne 1 et de variance η. La dépendance entre T^*_{ij} et T_{ij} sachant $Z_{ijk}(t)$ ($k = 0, \ldots, p$) est ici liée au fait que les effets aléatoires non observés (ω_{ij}) affectent en même temps les temps de progression et de décès. Les fragilités partagées ω_{ij} permettent de prendre en compte l'hétérogénéité dans les données associée aux variables non observées. Selon la valeur de ζ, on peut avoir le même effet de la fragilité sur

les deux critères de jugement ($\zeta = 1$), une association positive entre T_{ij}^* et T_{ij} ($\zeta > 0$) ou une association négative si $\zeta < 0$. En revanche, $\zeta = 0$ implique que TTP et OS ne sont pas associés, et par conséquent une censure non informative de la progression par le décès. Dans ce modèle, les auteurs supposent une indépendance entre les sujets du même groupe après prise en compte des facteurs pronostiques et après ajustement sur les effets aléatoires spécifiques aux sujets.

Pour une distribution log normale des effets aléatoires ω_{ij}^*, l'équation (2.51) s'écrierait plutôt comme suit:

$$\begin{cases} r_{ij}(t|\omega_{ij}^*, \mathbf{Z_{ij}}) = r_0(t)\exp(\omega_{ij}^* + \alpha Z_{ij0} + \sum_{k=1}^{p}\gamma_{1k}Z_{ijk}(t)) \\ \lambda_{ij}(t|\omega_{ij}^*, \mathbf{Z_{ij}}) = \lambda_0(t)\exp(\zeta\omega_{ij}^* + \beta Z_{ij0} + \sum_{k=1}^{p}\gamma_{2k}Z_{ijk}(t)) \end{cases}$$

Dans lequel le facteur de puissance ζ devient un facteur multiplicatif. Afin d'uniformiser les notations nous ne faisons pas de distinction entre le facteur multiplicatif et le facteur de puissance dans la suite de cette thèse. De plus, indépendamment de la distribution des effets aléatoires, ce paramètre s'interprète de la même façon. En effet, on passe l'expression ω_{ij}^{ζ} dans l'exponentielle afin de garantir que le terme de fragilité $\exp(\omega_{ij}^*) > 0$. $\exp(\zeta\omega_{ij}^*) = [\exp(\omega_{ij}^*)]^{\zeta}$, et donc ζ reste un paramètre de puissance.

Dans le deuxième modèle, Rondeau et al. (2015) considèrent que l'association entre TTP et OS est la résultante d'une association entre les groupes. Le modèle est par conséquent défini comme suit :

$$\begin{cases} r_{ij}(t|u_i, \mathbf{Z_{ij}}) = u_i r_0(t)\exp(\alpha Z_{ij0} + \sum_{k=1}^{p}\gamma_{1k}Z_{ijk}(t)) = u_i r_{ij}(t) \\ \lambda_{ij}(t|u_i, \mathbf{Z_{ij}}) = u_i^{\alpha}\lambda_0(t)\exp(\beta Z_{ij0} + \sum_{k=1}^{p}\gamma_{2k}Z_{ijk}(t)) = u_i^{\alpha}\lambda_{ij}(t) \end{cases} \tag{2.52}$$

où comme précédemment, les effets aléatoires u_i (terme de fragilités partagées) sont supposés indépendants et distribués suivant une gamma de moyenne 1 et de variance θ. Toutes les autres hypothèses sont identiques à celles du modèle (2.51). Toutefois, il est important de noter que le terme de fragilité (u_i) est partagé par tous les sujets du même groupe i. Si la variance des effets aléatoires u_i est différente de 0 et α est également différent de 0, la composante de variance représente en plus de l'association entre les groupes, la dépendance entre le temps de progression et l'évènement terminal.

Une autre façon de prendre en compte l'hétérogénéité entre les essais qui était un inconvénient dans le modèle (2.51) est de stratifier les fonctions de risque de base sur les groupes. Dans ce cas, le modèle conjoint (2.51) devient:

$$\begin{cases} r_{ij}(t|\omega_{ij}, \mathbf{Z_{ij}}) = \omega_{ij} r_{0,i}(t)\exp(\alpha Z_{ij0} + \sum_{k=1}^{p}\gamma_{1k}Z_{ijk}(t)) \\ \lambda_{ij}(t|\omega_{ij}, \mathbf{Z_{ij}}) = \omega_{ij}^{\zeta}\lambda_{0,i}(t)\exp(\beta Z_{ij0} + \sum_{k=1}^{p}\gamma_{2k}Z_{ijk}(t)) \end{cases} \tag{2.53}$$

L'utilisation des fonctions de risque stratifiées permet d'éviter les fragilités supplémentaires spécifiques aux groupes. Une fois de plus, les hypothèses retenues sur le modèle sont les mêmes que précédemment, mais les fonctions de risque de base $(r_{0,i}, \lambda_{0,i})$ dépendent des groupes. Toutefois, cette approche présente l'inconvénient d'accroître le nombre de paramètres du modèle lorsqu'on s'intéresse à l'estimation des fonctions de risque de base. En effet, il faut dans ce cas estimer autant de paramètres associés à ces fonctions pour chaque essai. Une autre conséquence à cette approche serait d'imposer des contraintes supplémentaires sur le nombre de sujets par essai, le nombre de sujet par bras de traitement, et le nombre de sujets présentant un évènement lié au critère de jugement principale et au critère de substitution, afin de garantir l'identifiabilité du modèle.

2.6.3 Calcul de la vraisemblance

Notons $T_i = (T_{i1}, \cdots, T_{in_i})$ les temps d'observation et $T_i^* = (T_{i1}^*, \cdots, T_{in_i}^*)$ les derniers temps de suivi pour les sujets du groupe i, $\Phi = (r_0(.), \lambda_0(.), \beta, \alpha, \theta)$ le vecteur des paramètres du modèle (2.52). La contribution marginale à la log-vraisemblance des sujets du groupe i en considérant la censure à droite est donnée par:

$$L_i(T_i, T_i^*, \Phi) = \int_{u_i} L_i(T_i, T_i^*, \Phi | u_i) f(u_i) du_i, \tag{2.54}$$

où la densité de probabilité des u_i est donnée par l'expression (2.47), et la distribution conditionnelle des temps de suivi est donnée par

$$L_i(T_i, T_i^*, \Phi | u_i) = \prod_{j=1}^{n_i} r_{ij}(T_{ij}|u_i)^{\delta_{ij}} \exp(-R_{ij}(T_{ij}|u_i)) \prod_{j=1}^{n_i} \lambda_{ij}(T_{ij}^*|u_i)^{\delta_{ij}^*} \exp(-\Lambda_{ij}(T_{ij}^*|u_i)) \tag{2.55}$$

A partir des expressions (2.55) et (2.54), on peut déduire l'expression de la i^{ime} contribution à la log-vraisemblance et donc la formulation de la log-vraisemblance marginale qui est donnée par :

$$\begin{aligned} l(\Phi) \quad = \quad & \sum_{i=1}^{G} \left\{ \sum_{j=1}^{n_i} \left[\delta_{ij} \log r_{ij}(T_{ij}) + \delta_{ij}^* \log \lambda_{ij}(T_{ij}^*) - \log \Gamma(1/\theta) - \frac{1}{\theta} \log \theta \right. \right. \\ & \left. \left. + \log \int_0^{+\infty} u_i^{(m_i + \alpha m_i^* + 1/\theta - 1)} \exp \left(- u_i/\theta - u_i \sum_{j=1}^{n_i} R_{ij}(T_{ij}) - u_i^{\alpha} \sum_{j=1}^{n_i} \Lambda_{ij}(T_{ij}^*) \right) du_i \right] \right\} \end{aligned} \tag{2.56}$$

Comme dans le cas du modèle à fragilités partagées (2.44), l'estimation des paramètres du modèle 2.52 s'appuie sur la maximisation de la log-vraisemblance marginale pénalisée, définie

dans le cas conjoint par

$$l_{pl}(\Phi) = l(\Phi) - \kappa_1 \int_0^\infty r_0''^2(t)dt - \kappa_2 \int_0^\infty \lambda_0''^2(t)dt$$

où κ_1 et κ_2 sont des paramètres de lissage qui peuvent s'obtenir par validation croisée à partir des modèles réduits de la forme de (2.44) et $l(\Phi)$ l'expression (2.56).

2.6.4 Mesure d'association au niveau individuel

Afin de mesurer l'association au niveau individuel entre TTP et OS, Rondeau et al. (2015) ont proposé d'utiliser le tau de Kendall. Il est compris dans l'intervalle [-1,1] et prend la valeur nulle lorsque T_{ij} et T_{ij}^* sont indépendants. A partir du modèle conjoint (2.52), on montre (Duchateau and Janssen 2008; Rondeau et al. 2015) que le τ de Kendall est donné par :

$$\begin{aligned}\tau = 2 \int_0^\infty \int_0^\infty \frac{(u_i^{\alpha+1} + u_{i'}^{\alpha+1})}{(u_i + u_{i'})(u_i^\alpha + u_{i'}^\alpha)} \frac{u_{i'}^{(1/\theta-1)} \exp(-u_{i'}/\theta)}{\Gamma(1/\theta)\theta^{1/\theta}} du_{i'} \\ \times \frac{u_i^{(1/\theta-1)} \exp(-u_i/\theta)}{\Gamma(1/\theta)\theta^{1/\theta}} du_i - 1\end{aligned} \tag{2.57}$$

Les paramètres associés aux effets aléatoires u_i (θ et α) pouvaient être utilisés pour mesurer l'association entre les deux critères de jugement. Seulement, étant donné leur appartenance à l'ensemble des réels, il est difficile de dire dans un contexte d'évaluation du TTP comme critère de substitution pour l'OS par exemple si la TTP est un critère de substitution valide ou non au niveau individuel.

2.7 Modèle conjoint à fragilités et à copules

Dans le modèle (2.52) proposé par Rondeau et al. (2015), l'hétérogénéité sur les fonctions de risque de base est prise en compte par un terme de fragilité u_i. En revanche, il reste encore une dépendance résiduelle due au fait que dans les méta-analyses, tous les facteurs pronostiques des évènements d'intérêts ne sont pas observés. Fort de ce constat, Emura et al. (2017) ont proposé un modèle conjoint à fragilités et à copules qui étend le modèle conjoint à fragilités (2.52) en introduisant une dépendance intra-sujet (niveau individuel) entre les temps de progression et de décès, à l'aide des copules (Nelsen 2006). L'avantage de considérer les copules plutôt que des effets aléatoires au niveau individuel réside principalement dans la réduction du nombre de paramètres (on n'estime que le paramètres de copule) et l'absence des intégrales dans le log-vraisemblance du modèle.

2.7.1 Définition du modèle

Le modèle proposé est le suivant:

$$\begin{cases} r_{ij}(t|u_i, \mathbf{Z_{1,ij}}) = u_i r_0(t) \exp(\beta_{\mathbf{1}}' \mathbf{Z_{1,ij}}) \\ \lambda_{ij}(t|u_i, \mathbf{Z_{2,ij}}) = u_i^{\alpha} \lambda_0(t) \exp(\beta_{\mathbf{2}}' \mathbf{Z_{2,ij}}), \\ Pr(T_{ij} > x, T_{ij}^* > y|u_i) = C_\theta[S_{T_{ij}}(x|u_i), S_{T_{ij}^*}(y|u_i)] \end{cases} \tag{2.58}$$

où les fonctions de survie et les fonctions de risque sont liées par:

$$\begin{cases} S_{T_{ij}}(x|u_i) = \exp\left\{ - u_i R_0(x) \exp(\beta_{\mathbf{1}}' \mathbf{Z_{1,ij}}) \right\}, & R_0(x) = \int_0^x r_0(t)dt, \\ S_{T_{ij}^*}(y|u_i) = \exp\left\{ - u_i^{\alpha} \Lambda_0(y) \exp(\beta_{\mathbf{2}}' \mathbf{Z_{2,ij}}) \right\}, & \Lambda_0(y) = \int_0^y \lambda_0(t)dt. \end{cases}$$

Emura et al. (2017) se sont focalisés sur la modélisation des dépendances positives entre T_{ij} et T_{ij}^* en utilisant les copules de Clayton et de Gumbel. Ils ont proposé une spécification des fonctions de copule légèrement différente de celle définie précédemment dans le modèle (2.34). Pour les copules de Clayton :

$$C_\theta(v, w) = (v^{-\theta} + w^{-\theta} - 1)^{-1/\theta}, \quad \theta > 0, \tag{2.59}$$

et pour les copules de Gumbel :

$$C_\theta(v, w) = \exp\left[- \{(-\log v)^{\theta+1} + (-\log w)^{\theta+1}\}^{\frac{1}{\theta+1}} \right], \quad \theta \geq 0. \tag{2.60}$$

2.7.2 Mesure d'association au niveau individuel

Comme vu dans le cas des modèles de copules en deux étapes pour la validation des critères de substitution (2.34), la dépendance entre T_{ij} et T_{ij}^* au niveau individuel peut être mesurée à l'aide du τ de Kendall. A partir des modèles (2.59)-(2.60) τ peut s'exprimer simplement comme une fonction du paramètre de copule : $\tau = \theta/(\theta + 2)$ pour la fonction de copule de Clayton et $\tau = \theta/(\theta + 1)$ pour la fonction de copule de Gumbel.

Qu'on se situe dans le cas du modèle conjoint à fragilités de Rondeau et al. (2007) et de ses extensions (Rondeau et al. 2015), ou alors dans le cas du modèle conjoint à fragilités et à copules (Emura et al. 2017), aucun n'inclut les effets aléatoires au niveau essai en interaction avec le traitement. Or il s'agit d'une spécification du modèle nécessaire à la validation des critères de substitution au niveau essai. Par conséquent, dans la suite de cette thèse et principalement dans les Chapitres 3 et 5, nous nous proposons d'étendre ces modèles pour permettre la validation en une étape d'un critère de substitution. Bien évidemment, l'idée est de rester dans la même

philosophie que la méthode standard (Burzykowski et al. 2001) lorsqu'on fait face à deux temps d'évènements.

Chapter 3

Méthode de validation en une étape des critères de substitution utilisant les données issues de plusieurs essais cliniques randomisés sur le cancer avec des critères de jugement à temps d'évènements

---

3.1 Article

Le choix du critère de jugement est une étape fondamentale dans le processus de développement d'un nouveau traitement. Bien que ce soit pertinent, recourir à un critère de jugement difficile à observer comme par exemple le temps jusqu'au décès risquerait de compromettre les bénéfices cliniques du nouveau traitement, de retarder sa durée de mise sur le marché s'il s'avère qu'il est efficace, et d'exiger une taille d'échantillon suffisamment grande pour atteindre la puissance statistique nécessaire à la mise en évidence d'une différence significative si elle existe. L'utilisation des critères de substitution semble être une option importante pour palier à ces problèmes. Il s'agit d'un champ de recherche assez présent dans la littérature depuis près de trois décennies. Toutefois, l'utilisation d'un critère de substitution doit faire suite à une démarche rigoureuse de validation pour ainsi éviter de tirer des conclusions erronées sur la validité ou non du médicament en cours de développement. Plusieurs approches statistiques ont été développées pour valider un critère de substitution (Buyse et al. 2016). Cependant, lorsque le critère de jugement principal et le critère de substitution sont des temps d'évènements, le consensus actuel est d'utiliser des données de méta-analyses (ou des études multicentriques) et

de baser la validation sur l'approche en deux étapes de Burzykowski et al. (2001), qui a été présentée dans le Chapitre 2. Conscient des difficultés présentes dans la mise en œuvre de cette méthode, nous proposons dans ce chapitre un nouveau concept de validation des critères de substitution, basé sur une approche en une étape.

En effet, nous proposons de modéliser conjointement le risque de survenue du critère de substitution, qui peut être une progression, et le risque d'observer l'évènement terminal. Nous utilisons pour cela un modèle conjoint à fragilités partagées dans lequel nous introduisons deux niveaux d'effets aléatoires: (a) des effets aléatoires au niveau individuel, ω_{ij}, communs aux critères de substitution et au critère de jugement principal et permettant de prendre en compte l'hétérogénéité au niveau individuel due à l'absence d'autres variables explicatives pertinentes dans le jeu de données. La variance de ω_{ij} permet de mesurer l'association entre S et T au niveau individuel. (b) Au niveau essai, nous considérons des effets aléatoires corrélés en interaction avec le traitement, dont les paramètres de variance nous permettront de définir la mesure d'association au niveau essai entre S et T. Nous proposons à l'issue de ce modèle une nouvelle définition du τ de Kendall et du coefficient de détermination R^2_{trial} pour la validation au niveau individuel et au niveau essai d'un critère de substitution. Nous considérons dans ce modèle des fonctions de risque de base flexibles en utilisant des splines (Rondeau et al. 2003). Par conséquent, les paramètres du modèle sont estimés en maximisant la log-vraisemblance marginale pénalisée, à l'aide de l'algorithme de `Marquardt`. Plusieurs méthodes d'intégration numériques ont été considérées pour approcher les intégrales présentes dans la log-vraisemblance marginale.

Les performances du modèle ont été étudiées en simulation et nous avons observé des résultats satisfaisants en ce qui concerne les mesures d'association pour la validation du critère de substitution. Le modèle était assez robuste à la mauvaise spécification et à la variation des caractéristiques des données comparées aux approches existantes. Nous avons également observé une réduction considérable des problèmes de convergence, comparée à l'approche en deux étapes. Dans une analyse comparative avec les approches existantes, nous avons appliqué le modèle à un jeu de données réelles sur le cancer gastrique pour évaluer si la DFS était un critère de substitution valide pour l'OS dans l'étude du bénéfice apporté par un traitement adjuvant. Les résultats pour le R^2_{trial} étaient comparables entre les approches. Toutefois l'écart-type du R^2_{trial} était mieux estimé avec le nouveau modèle. En ce qui concerne le τ de Kendall, bien que par définition on ne pouvait pas les comparer directement, ils traduisaient une validation acceptable au niveau individuel. Pour conclure, nous avons proposé dans ce travail, une approche supplémentaire en une étape pour valider des critères de substitution en cancer, utilisant des données individuelles des patients randomisés dans des méta-analyses d'essais cliniques.

Ce travail a été publié dans *Statistics in Medicine* (Sofeu et al. 2019) et la méthode a été

implémentée dans le package `R frailtypack` (Voir Chapitre 4).

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

One-step validation method for surrogate endpoints using data from multiple randomized cancer clinical trials with failure-time endpoints

Casimir Ledoux SOFEU*[1] | Takeshi Emura[2] | Virginie Rondeau[1]

[1]INSERM U1219 (Biostatistic), Université Bordeaux Segalen, 146 rue Léo Saignat, 33076 Bordeaux Cedex, France.
[2]Graduate Institute of Statistics, National Central University, Jhongda Road, Jhongli City, Taoyuan 32001, Taiwan

Correspondence
*Casimir Ledoux SOFEU, INSERM U1219 (Biostatistic), Université Bordeaux Segalen, 146 rue Léo Saignat, 33076 Bordeaux Cedex, France. Email: casimir-ledoux.sofeu@inserm.fr, scl.ledoux@gmail.com

Summary

A surrogate endpoint can be used instead of the most relevant clinical endpoint to assess the efficiency of a new treatment. Before being used, a surrogate endpoint must be validated based on appropriate methods. For two failure-time endpoints, two association measurements are usually used, Kendall's tau at the individual-level and the adjusted coefficient of determination ($R^2_{trial,adj}$) at the trial-level. However, $R^2_{trial,adj}$ is not always available due to model estimation constraints. We propose a one-step validation approach based on a joint frailty model, including both individual-level and trial-level random effects. Parameters have been estimated using a semi-parametric penalized marginal log-likelihood method, and various numerical integration approaches were considered. Both individual and trial-level surrogacy was evaluated using a new definition of Kendall's tau and the coefficient of determination. Estimators' performances were evaluated using simulation studies and satisfactory results were found. The model was applied to individual patient data meta-analyses in gastric cancer to assess disease-free survival (DFS) as a surrogate for overall survival (OS), as part of the evaluation of adjuvant therapy.

KEYWORDS:
Cancers clinical trials, Joint frailty models, Meta analysis, Numerical integration, One step Validation method, Surrogate endpoint.

1 | INTRODUCTION

One of the most important factors in setting up clinical trials is the choice of the endpoint to be used to assess the efficacy of the new treatment. The choice is often focused on the most sensitive and relevant criterion[1], usually referred to as the "true" endpoint. The use of the true endpoint is quite difficult in clinical trials. For example, in cancer clinical trials, overall survival is a common clinical endpoint used to evaluate the effect of new treatments. However, its use requires a sufficiently long follow-up time and a sufficiently high sample size to show a significant difference in the treatment effect; other composite criteria such as quality of life are often difficult to measure. Faced with this difficulty, in recent years, several clinical research projects have focused on the replacement of this primary endpoint with so-called surrogate endpoints or intermediate criteria. These are criteria that can be observed earlier, more conveniently, or more frequently; and whose use would predict the effect of treatment on the primary endpoint[2]. The use of another observable endpoint, such as cancer progression, would therefore shorten the duration of a clinical trial while providing earlier conclusions on the primary endpoint.

However, before a potential surrogate is proposed, it must be validated or evaluated using appropriate methods. The use of a simple correlation between the surrogate endpoint and the true endpoint does not guarantee a correct inference based on a potential surrogate endpoint[3,4]. Several authors[5,6,7] have proposed different approaches to carry out this validation in the context of a clinical trial. Finally, two measurements have been recommended to this end, namely the proportion of treatment effect explained (ρ_z) and the relative effect (RE)[7]. However, some practical problems persist, such as the confidence interval of RE, which can be wide, and the need to take into account the heterogeneity of treatment effects both on the surrogate and the true endpoints in order to use RE for prediction. To overcome these problems it is possible to use data with a sufficiently large sample size, or data from multiple randomized clinical trials (or meta-analysis), as suggested by Buyse and Molenberghs[7].

In this meta-analysis framework, Buyse *et al.*[8] proposed an approach based on both trial-level and individual-level associations for two Gaussian distributed endpoints. When both the surrogate endpoint and the true endpoint are survival times, the validation of the surrogate is more complex due to several parameters, such as the presence of censorship and competing risks. For these reasons, different validation approaches in the context of failure-time endpoint have been developed in recent years to evaluate surrogate endpoints, and include the likelihood reduction factor[9], the Bayesian approach[10,11], the causal inference[12] and the one-step Poisson approach[13]. A literature review describes some of the different possible validation approaches in the context of clinical trials in oncology[14].

Although surrogate endpoints are still a very important research topic, the current consensus is to base validation on a "correlation" approach based on a two-step model in which Burzykowski *et al.*[15,16] proposed an extension of the model developed by Buyse *et al.*[8] to event-time endpoints. In a first step, in order to validate the quality of the potential surrogate at the individual-level, the authors proposed to use a measure of association between the surrogate and the true endpoints using a copula model. In a second step, the potential surrogate is designated as a "good" criterion at the "trial-level" if it is able to predict the treatment effect on the true endpoint based on the treatment effect observed on the surrogate. To evaluate this, Burzykowski *et al.*[15] proposed to use a random effects model and the quality of the surrogate at the trial-level was evaluated through of a coefficient of determination (R^2_{trial}). In the latter model, due to the fact that the treatment effects used are estimated in the first-step model, the authors proposed to account for the estimation errors which induce bias in the estimation of R^2_{trial}. Therefore, they proposed an adjusted R^2_{trial} ($R^2_{trial,adj}$) in this context.

However, $R^2_{trial,adj}$ is not always available, mainly due to convergence problems or estimations with adjusted models in clinical trials[16,15,10,13]. Renfro *et al.*[10] reported that these convergence problems were most often encountered in the first stage model (at the individual-level) of validation. But even when the first stage provided estimators, the second stage (at the trial-level) did not always give an estimate of the adjusted coefficient of determination due to further numerical problems. These concerns are frequently encountered and are influenced by the number and size of the groups (or trials) as well as the assumptions made about the basic risks from one trial to another. In their application, Burzykowski *et al.*[15] made a strong assumption of homogeneity between trials in order to obtain $R^2_{trial,adj}$. Simulation studies proposed by Rotolo *et al.*[13] showed an observed bias on Kendall' τ and $R^2_{trial,adj}$ especially when faced with a high trial-level or individual-level association, as well as real convergence issues with copula models.

Emura *et al.*[17] recently proposed a joint frailty-copula model between tumor progression and death for meta-analysis. This approach extends the joint model for the dependence between clustered times to tumor progression and deaths proposed by Rondeau *et al.*[18], which only takes into account the trial-level or the individual-level heterogeneity. Emura *et al.*[17] deals with heterogeneity between trials using a joint model with a trial random effect, and the residual dependence between endpoints was considered in a copula model. None of these models or other joint models on survival endpoints[19,20,21] has accounted for both individual and trial-level heterogeneity, and the random effect treatment-by-trial interaction.

Taking inspiration from these models, in this paper we propose a surrogate endpoint one-step validation method based on the joint frailty model, including respectively a shared random effect accounting for heterogeneity at the individual-level, two correlated random effects treatment-by-trial interaction and a shared random effect associated with baseline risks accounting for the heterogeneity between trials. New definition of Kendall's τ have been proposed in this context. Distinct numerical integration methods were used to estimate the integrals in the marginal log-likelihood, and results were compared. The performances of the estimators were evaluated through simulations. This approach extends the two-stage validation method[15]. The particularity of our approach compared to the recent one-step Poisson validation method[13] lies in the level of the model definition, the estimation approach and the individual-level validation (Kendall's τ).

We propose to organize this paper as follows. In section 2 we define the joint surrogate model. The full penalized log-likelihood construction and the estimation methods are also described within this section. The surrogacy evaluation criteria are defined in

Section 3. In Section 4 we present the simulation studies and in section 5 an application to individual patient data meta-analyses in gastric cancer is proposed. Finally, in section 6 we present a concluding discussion.

2 | METHODS

2.1 | Joint frailty model for surrogate (S) and true (T) endpoints

In the spirit of the random effects models defined in the context of two normally distributed endpoints [8], the joint frailty model needed for failure-time endpoints included both the individual and trial-level random effects. The proposed model, called "joint surrogate model", will, in one step lead to define the validation criteria of a candidate surrogate endpoint. The joint surrogate model is therefore defined as follows:

$$\begin{cases} \lambda_{S,ij}(t|\omega_{ij}, u_i, \upsilon_{S_i}, Z_{ijk}) = \lambda_{0S}(t)\exp(\omega_{ij} + u_i + \upsilon_{S_i} Z_{ij1} + \sum_{k=1}^{p} \beta_{Sk} Z_{ijk}) \\ \lambda_{T,ij}(t|\omega_{ij}, u_i, \upsilon_{T_i}, Z_{ijk}) = \lambda_{0T}(t)\exp(\zeta\omega_{ij} + \alpha u_i + \upsilon_{T_i} Z_{ij1} + \sum_{k=1}^{p} \beta_{Tk} Z_{ijk}) \end{cases} \tag{1}$$

with,

$$\omega_{ij} \sim N(0, \theta), \quad u_i \sim N(0, \gamma) \tag{2}$$

and

$$\begin{pmatrix} \upsilon_{S_i} \\ \upsilon_{T_i} \end{pmatrix} \sim MVN\left(\mathbf{0}, \Sigma_\upsilon\right), \Sigma_\upsilon = \begin{pmatrix} \sigma^2_{\upsilon_S} & \sigma_{\upsilon_{ST}} \\ \sigma_{\upsilon_{ST}} & \sigma^2_{\upsilon_T} \end{pmatrix}. \tag{3}$$

In model (1), $\lambda_{S,ij}(.)$, $\lambda_{0S}(t)$ and $\beta_{Sk}(k = 1, \cdots, p)$ are respectively the hazard function of failure-time for the j^{th} patient in trial i, the baseline hazard function and the fixed effects (or log-hazard ratio) corresponding to the covariates X_{ijk} associated with the surrogate endpoint; $\lambda_{T,ij}(.)$, $\lambda_{0T}(t)$ and $\beta_{Tk}(k = 1, \cdots, p)$ are defined as above and are associated with the true endpoint. ω_{ij} is a shared individual-level frailty that will take into account heterogeneity at the individual-levels; ω_{ij} will serve to answer the question, are the surrogate and the true endpoints correlated at the individual-level? ; u_i is a shared frailty effect associated with the baseline hazard function that will serve to take into account the heterogeneity between trials of the baseline hazard function, associated with the fact that we have several trials in this meta-analytical design. The power parameters ζ and α distinguish both individual and trial-level heterogeneities between the surrogate and the true endpoint. υ_{S_i} and υ_{T_i} are two correlated random effects treatment-by-trial interactions (trial-level frailties in interaction with the treatment) which aim to answer the question, does the effect of treatment on the surrogate endpoint reliably predict the effect of treatment on the true endpoint in each trial? Z_{ij1} represents the treatment arm to which the patient has been randomized. We assume that ω_{ij}, u_i and υ_{S_i} are mutually independent, but also that ω_{ij}, u_i and υ_{T_i} are mutually independent.

2.2 | Inference in the Gaussian random effects joint model for S and T endpoints

2.2.1 | Log-likelihood construction

Let S_{ij} and T_{ij} denote the follow-up times associated respectively with surrogate and true endpoints for subject j ($j = 1 \cdots n_i$) belonging in trial i ($i = 1 \cdots G$), where n_i is the size of trial$i$ and G is the total number of considered trials. Similarly, δ_{ij} and δ^*_{ij} denote the progression and death indicators. Let $\Phi = (\sigma^2_{\upsilon_S}, \sigma^2_{\upsilon_T}, \sigma_{\upsilon_{ST}}, \theta, \gamma, \lambda_{0T}(.), \lambda_{0S}(.), \beta_{Sk}, \beta_{Tk})$, the vector containing all the unknown parameters of (1).The full marginal log-likelihood is given by (4). Detail on construction is given in appendix A.

$$
\begin{aligned}
l(\Phi) = \sum_{i=1}^{G} \Bigg[\log \Bigg\{ & \int_{u_i} \int_{v_{S_i}} \int_{v_{T_i}} \\
& \exp\left\{ -\frac{1}{2}\left(2\log(2\pi) + log|\Sigma_v| + \log(2\pi\gamma) + \frac{u_i^2}{\gamma} \right) - \frac{1}{2}(v_{S_i}, v_{T_i})\Sigma_{v_i}{}^{-1}(v_{S_i}, v_{T_i})' \right\} \\
& \times \Bigg[\prod_{j=1}^{ni} \int_{\omega_{ij}} \exp\left\{ \delta_{ij}\left(\log(\lambda_{0S}(T_{ij})) + \sum_{k=1}^{p} \beta_{Sk} Z_{ijk} \right) + \delta_{ij}^{*}\left(\log(\lambda_{0T}(D_{ij})) + \sum_{k=1}^{p} \beta_{Tk} Z_{ijk} \right) \right. \\
& + u_i(\delta_{ij} + \delta_{ij}^{*}\alpha) + (v_{S_i}\delta_{ij} + v_{T_i}\delta_{ij}^{*}) Z_{ij1} + \omega_{ij}(\delta_{ij} + \delta_{ij}^{*}\zeta) - \frac{1}{2}\log(2\pi\theta) - \frac{\omega_{ij}^2}{2\theta} \\
& - \Lambda_{0S}(T_{ij}) \exp\Big(\sum_{k=1}^{p} \beta_{Sk} Z_{ijk}\Big) \exp(\omega_{ij} + u_i + v_{S_i} Z_{ij1}) \\
& \left. - \Lambda_{0T}(D_{ij}) \exp\Big(\sum_{k=1}^{p} \beta_{Tk} Z_{ijk}\Big) \exp(\zeta\omega_{ij} + \alpha u_i + v_{T_i} Z_{ij1}) \right\} d\omega_{ij} \Bigg] dv_{T_i} dv_{S_i} du_i \Bigg\} \Bigg]. \qquad (4)
\end{aligned}
$$

2.2.2 | The estimation methods

In order to estimate the integrals over the random effects present in the likelihood, we considered four numerical integration strategies including the full Monte-Carlo integration (MC), the full Pseudo-adaptive Gaussian-Hermite quadrature (PGH), a combination of Monte-Carlo for trial-level random effects integration and the non-adaptive (MC-GH) and pseudo-adaptive (MC-PGH) Gaussian-Hermite quadrature method for individual-level random effect integration.

2.2.3 | The semi-parametric penalized likelihood approach

As introduced by Rondeau *et al.*[19,18], in the context of frailties models we used a semi parametric penalized likelihood approach to estimate the different parameters $\sigma^2_{v_S}$, $\sigma^2_{v_T}$, $\sigma_{v_{ST}}$, θ, γ, β_{Sk}, β_{Tk} and the baseline hazard functions $\lambda_{0S}(t)$ for surrogate endpoint or $\lambda_{0T}(t)$ for death times. We thereby obtained a smoothed estimate of the baseline hazard function by approximating it using splines. We penalized the likelihood by a term which has large values for rough functions[22,23]. The result is this definition of the penalized log-likelihood:

$$
pl(\Phi) = l(\Phi) - k_1 \int_0^{\infty} \lambda''^2_{0S}(t)dt - k_2 \int_0^{\infty} \lambda''^2_{0T}(t)dt, \qquad (5)
$$

where $l(\Phi)$ is the full log likelihood defined in (4), k_1 and k_2 the positive smoothing parameters which control the trade-off between the data fit and the smoothness of the functions. Maximization of (5) defines the maximum penalized likelihood estimators (MPnLEs) $\hat{\sigma}^2_{v_S}$, $\hat{\sigma}^2_{v_T}$, $\hat{\sigma}_{v_{ST}}$, $\hat{\theta}$, $\hat{\gamma}$, $\hat{\alpha}$, $\hat{\zeta}$, $\hat{\beta}_{Sk}$, $\hat{\beta}_{Tk}$, $\hat{\lambda}_{0S}(t)$ and $\hat{\lambda}_{0T}(t)$. We directly use H^{-1} as a variance estimator, where H is minus the converged hessian of the penalized log-likelihood.

The estimators $\hat{\lambda}_{0S}(.)$ and $\hat{\lambda}_{0T}(.)$ were approximated using cubic M-splines, which are a variant of cubic B-splines[24]. M-splines are non-negative and easy to integrate or differentiate. As we use a cubic spline (or of order 4), the second derivative of λ_{0S} and λ_{oT} is approximated by a linear combination of piecewise polynomial approximation of order 2. This approximation allows flexible shapes of the hazard functions while reducing the number of parameters. If we denote $\tilde{\lambda}_{0S}$ as an approximation to the MPnLE $\hat{\lambda}_{0S}$, the approximation error can be made as small as required by increasing the number of knots, as shown by Rondeau *et al.*[25]. Therefore, to obtain a good estimation of the theoretical hazard function, it is necessary to use as many knots as possible to get a MPnlE close to the true hazard function. In our approach, although there are two different hazard functions (for surrogate and true endpoints), we use the same basis of splines for each function. However, the spline coefficients are different for the distinct functions.

The smoothing parameters were chosen using a maximizing likelihood cross-validation criterion described in Joly *et al.*[23]. Two separate Cox proportional hazard models with no covariates were used to obtain k_1 and k_2.

2.2.4 | Computational procedure

Our new approach for the validation of surrogate endpoints was implemented using a `Fortran` program. The estimated parameters were obtained by the robust Marquardt algorithm[26], which is a combination of Newton-Raphson and steepest descent algorithms. This variant is more stable than the Newton-Raphson algorithm while preserving its fast convergence property near the maximum. We imposed a positivity constraint on the variance parameters and the spline coefficients. Following this change of variable, the standard errors of variance-covariance parameters were computed using the delta method. This re-parametrization was very convenient numerically without any adverse effects on the approximation.

3 | CRITERIA TO EVALUATE SURROGATE ENDPOINTS

3.1 | Individual-level surrogacy

To evaluate the individual-level surrogacy, Buyse and Molenberghs[7] suggested using the association between the surrogate (S) and the true (T) endpoints after adjustment for the treatment effects. For two failure-time endpoints, Burzykowski *et al.*[15] proposed to use Kendall's τ based on a copula model to measure the strength of association between S_{ij} and T_{ij} after adjusting their marginal distributions for the trial and the treatment effects. So, as in the original Burzykowski *et al.*[15] model, we proposed to use Kendall's τ to evaluate the individual-level surrogacy. Kendall's τ is the difference between the probability of concordance and the probability of discordance of two realizations of S_{ij} and T_{ij}[27,28]. It belongs to the interval [-1,1] and assumes a zero value when S_{ij} and T_{ij} are independent.

We assume the joint conditional distributions of S_{ij} and T_{ij} after adjusting for both the trial and the treatment effects. It can be shown as described in appendix B, that for the proposed joint surrogate model (1), Kendall's τ is given by:

$$\tau = 2 \int_{u_i} \int_{\omega_{ij}} \int_{u_{i'}} \int_{\omega_{i'j'}} \frac{\exp(\omega_{ij} + u_i + \zeta\omega_{ij} + \alpha u_i) + \exp(\omega_{i'j'} + u_{i'} + \zeta\omega_{i'j'} + \alpha u_{i'})}{(\exp(\omega_{i'j'} + u_{i'}) + \exp(\omega_{ij} + u_i))(\exp(\zeta\omega_{i'j'} + \alpha u_{i'}) + \exp(\zeta\omega_{ij} + \alpha u_i))} \frac{1}{\sqrt{2\pi\theta}} \exp\left[-\frac{1}{2}\frac{\omega_{i'j'}^2}{\theta}\right] \frac{1}{\sqrt{2\pi\gamma}} \exp\left[-\frac{1}{2}\frac{u_{i'}^2}{\gamma}\right] d\omega_{i'j'} du_{i'} \frac{1}{\sqrt{2\pi\theta}} \exp\left[-\frac{1}{2}\frac{\omega_{ij}^2}{\theta}\right] \frac{1}{\sqrt{2\pi\gamma}} \exp\left[-\frac{1}{2}\frac{u_i^2}{\gamma}\right] d\omega_{ij} du_i - 1.$$

We estimated Kendall's τ using a Monte-Carlo integration procedure.

3.2 | Trial-level surrogacy

The key motivation for validating a surrogate endpoint is to be able to predict the effect of treatment on the true endpoint, based on the observed effect of treatment on the surrogate endpoint. As shown by Buyse *et al.*[8], a study of the quality of the prediction can be used to base the surrogacy evaluation at the trial-level on the coefficient of determination obtained from the covariance matrix Σ_v defined in (3). Therefore,

$$R^2_{trial} = \frac{\sigma^2_{v_{ST}}}{\sigma^2_{v_S}\sigma^2_{v_T}}. \tag{6}$$

The SEs of R^2_{trial} was calculated using the delta method, and the 95% CI of Kendall's τ using the parametric bootstrap. The use of delta method can lead to confidence limits violating the [0,1], as noted by Burzykowski *et al.*[15]. However, using other methods would not materially change the conclusions of the article. On the basis of previous developments, we suggest terming a surrogate endpoint as valid if R^2_{trial} and Kendall's τ are sufficiently close to 1.

4 | SIMULATIONS

4.1 | Simulation design

A simulation study of the joint surrogate and true endpoints model was performed to evaluate the estimators. We considered two sample sizes with a variable number of subjects (N=600 and 1000 resp·) and a variable number of trials per sample (n=10 and 30 resp·). Five hundred simulated data sets were used in each case. For each simulation run, the proposed joint frailty model (1) was used, and the data was generated with the following algorithm:

1. For each subject j ($j = 1, \dots, n_i$) from cluster i ($i = 1, \dots, G$), we generated a Gaussian random variable $\omega_{ij} \sim N(0, \theta)$, with θ fixed following the desired value of Kendall's τ.

2. For each trial i, we generated respectively a Gaussian random variable $u_i \sim N(0, \gamma)$, with γ defined as θ, and a couple of Gaussian random variables $(\upsilon_{S_i}, \upsilon_{T_i}) \sim MVN(\mathbf{0}, \Sigma_\upsilon)$, with Σ_υ defined as in (3), where $\sigma^2_{\upsilon_S}$=$\sigma^2_{\upsilon_T}$=0.7 and $\sigma_{\upsilon_{ST}} = 0.42$ or 0.63

3. We generated the binary treatment variable Z_{ij1} from a Bernoulli distribution with P(Z=1)=0.5.

4. A fixed right-censoring variable was considered with $C_{ij} = 549.24$ and $C_{ij} = 225.24$ for respectively $\sim 40\%$ and 70% censorship.

5. We generated a uniform random variable $u \sim U(0, 1)$ and therefore event times D_{ij} and P_{ij} associated with the true and the surrogate endpoints following a Weibull distribution and using the proposed model (1). The scale and shape parameters were respectively set to $\gamma_T = 3.0$ and $\rho_T = 0.0025$ for the true endpoint; $\gamma_S = 1.8$ and $\rho_S = 0.0045$ for the surrogate endpoint. The corresponding observed death times were $T_{ij} = min(D_{ij}, C_{ij})$ and $\delta^*_{ij} = 1$ if $T_{ij} = D_{ij}$. In the same way, the observed progression times were $S_{ij} = min(P_{ij}, T_{ij})$ and $\delta_{ij} = 1$ if $S_{ij} = P_{ij}$ and $P_{ij} \neq T_{ij}$. When progression and death occurred the same days, we just considered the death event.

For all simulation designs, the fixed-treatment effects β_S and β_T were set to -1.25. In order to investigate the effect of the degree of correlation both at the individual and the trial-levels on the estimators' performance we considered for all the scenarios a strong and weak correlation by varying R^2_{trial} (between 0.36 and 0.81) and Kendall's τ (between 0.38 and 0.61). We eliminated the rare cases (less than 3%, except with high censoring where we observed up to 10% according to the trial number) where convergences or numerical problems occurred in estimating parameters.

We extend our simulation study to investigate the impact of model misspecification on the surrogacy evaluation, by generating new trials of data from a Clayton copula model. We considered two scenarios with Kendall's $\tau = 0.61$ and $R^2_{trial} = 0.36$ with respectively 30 and 10 trials.

4.2 | Evaluation criteria

Convergences were considered when the difference between two consecutive log likelihoods was small ($< \epsilon_b$), the estimated coefficients were stable (consecutive values ($< \epsilon_a$) and the gradient was small enough ($< \epsilon_d$). The default values were $\epsilon_a = \epsilon_b = \epsilon_d = 10^{-4}$. We considered a maximum number of 35 iterations. In appendix C we describe an algorithm used to manage the non-convergence cases. We are mainly interested in the trial-level and the individual-level surrogacy. For each parameter, we report the mean, the empirical standard errors (SD), i.e. the standard error of estimates (with the exception of Kendall's τ), the mean of the estimated standard errors (SE) and the coverage percentage of the 95% confidence interval estimates (CP in %).

4.3 | Results

In this section we assess the proposed joint surrogate model's performance and, based on a simulation, compare the results with those obtained from the existing approaches.

4.3.1 | Integration methods assessment

When comparing the full Monte-Carlo integration with an integration method combining both Monte-Carlo and the Gaussian-Hermite quadrature in cases of high trial-level association (R^2_{trial}=0.81), weak individual-level association (Kendall's τ=0.378)

and a high trial number (G=30), we observed the same estimation performance across all integration strategies for most parameters (see table 1). However, parameters associated with the individual-level random effects (θ and ζ) presented lower biases and smaller standard errors (SE and SD) when using MC than with MC-GH and MC-PGH integrations. Surrogacy measurements (Kendall's τ, R^2_{trial} and fixed treatment effects (β_S and β_T) were not very different between these integration methods. With a low trial number (G=10) with a consistent number of subjects per trial, individual-level association (τ) and some associated parameters (θ and ζ) were better estimated using MC-GH compared to MC-PGH in terms of bias and CP.

We compared the previous integration methods with the full pseudo-adaptive Gaussian-Hermite quadrature (PGH). The results described in Table D1 of appendix D show that the variance of the shared trial random effect associated with the baseline hazard risk (γ) and the corresponding power parameter (α), were better estimated using the PGH method in terms of bias. However, $\sigma_{v_{ST}}$ was slightly underestimated when using PGH, leading to a decline in the R^2_{trial} coverage rate. Nevertheless, by increasing the number of quadrature points to 15 or 20 underlined bias and coverage could be considerably improved.

In cases of high individual and trial-level association, combining integration methods presented satisfactory and comparable results (see table D2 in appendix D). However, in the scenario with τ=0.64, a high trial number (G=30) and a small trial size, MC gave the worst estimation performance on θ, R^2_{trial} and Kendall's τ in terms of bias and CP (data not shown). These observed results underline the need to use the Gaussian Hermite quadrature instead of Monte-Carlo for integration over the individual-level random effects with only one random effect, especially in the event of a small trial size.

4.3.2 | Proposed one-step joint surrogate model assessment

In table 2, we assess the effect of increasing clinical trial size and the censoring rate on the estimates in both high individual and trial-level association. Overall, we found low bias on R^2_{trial} and Kendall's τ as well as on other estimated parameters. In addition, when increasing the trial number but keeping the same sample size, we observed an improvement in the CP of almost all estimated parameters except θ and ζ in which we observed an increase in standard errors. By increasing the trial size, we improved the precision of most estimations and the associated variabilities (SE and SD) and obtained 95% CP. The 95% CP for R^2_{trial} was the most affected by this improvement. Another simulation design based on 3000 subjects and 30 trials showed a 95% CP of 82%. Similar results were observed with weak trial-level association (see the third part of Table D3 in appendix D). The proposed model was quite robust to changes in the censoring level, especially as regards the surrogacy assessment. However, in the scenario with a high censoring rate (70%), we encountered slightly more numerical problems (around 10% rejection rate compared to 1% with 40% of censoring).

In all the presented scenarios, with a small trial number, we encountered some estimation problems with the standard error of the variance of the trial random effect associated with the baseline risk. Moreover, with a lower trial-level association, we found a slight overestimation of R^2_{trial} regardless of the trial number and the individual-level association (see the two first parts of Table D3 in appendix D). We also observed an improvement in parameter estimation in terms of biases and coverage rate with an increase in the number of generated samples for the Monte-Carlo integration method (data not shown).

In all models estimated during the simulations, we initialized parameters with their true values. However, by using arbitrary initial values, the results were quite close to those previously observed (data not shown).

4.3.3 | Comparison between the proposed one-step joint surrogate approach and the existing two-step Copula and one-step Poisson approaches

In table 3, we compare the performance of the proposed joint surrogate approach to those of the existing approaches. We therefore used an approach based on the copula model[16] with two distinct copula families, Clayton's and Plackett's; and the one-step Poisson approach[13] which hypothesizes that the individual-level (respectively the-trial) random effects are the same regardless of the endpoint. The copula-based models were estimated using the *BFGS* (Quasi-Newton optimization method) algorithm. Concerning the one-step Poisson approach, we used the *Bobyqa*(Optimization by quadratic approximation routines) algorithm. All these optimization methods are available in the R package *Optimx*.

Numerical problems were encountered when performing simulations using the copula two-step approaches, especially with a high trial number (with small trial sizes). The rejection rates range between 47% and 61% for Clayton's and Plackett's copula when G=30 and between 3 and 11% when G=10. With Hougaard's copula, we always encountered a 100% rejection rate (results not shown). Using the Poisson-based model, the rejection rates were less than 4%. Conversely, as shown in table 3, the proposed joint model always had less than 1% of rejections.

Concerning surrogacy evaluation, at the individual-level, we observed that the existing approaches underestimated Kendall's τ. The proposed joint model estimated Kendall's τ well, with bias and MSE around 10-3, irrespective of the trial number and both the individual and trial-level association. For trial-level surrogacy evaluation, the proposed joint surrogate model better estimated the R^2_{trial} and MSE across the scenarios. In the scenario with a high individual and trial-level association, both the Poisson and the joint approaches showed a comparable MSE. Irrespective of the scenario, the existing approaches overestimated the R^2_{trial}. With a weak trial-level association, regardless of the individual-level association, we observed in all cases a strong overestimation of R^2_{trial}, but quite a moderate one with the new approach.

Same observations were again noted with the two-step approach using the *Bobyqa* optimization algorithm, as described in table D4 in appendix D. However, Kendall's τ in this case was the most underestimated and the percentage of rejection was higher than with the quasi-newton *BFGS* algorithm.

In oder to confirm the robustness of the proposed joint surrogate model in term of the surrogacy evaluation, we performed other simulation scenarios by generating the data with a Clayton copula model. As described in table 4, we observed good estimates for R^2_{trial} with the proposed joint surrogate model in term of bias and MSE, regardless of the number of trials or trials size. Also, comparable results were found with the existing approaches. Regarding the Kendall's τ, as previously described, the Poisson approach tends to underestimate the real value of τ. However, convergence issues persist when estimating with the two-step copula approaches, with more than 80% rejection rate in scenario design with 30 trials.

5 | APPLICATION TO THE ADJUVANT CHEMOTHERAPY AND RESECTABLE GASTRIC CANCER META-ANALYSES DATA

The meta-analysis of the GASTRIC (Global Advanced/Adjuvant Stomach Tumor Research International Collaboration) Group[29], available at the publisher's web site[14], was used to illustrate the proposed joint surrogate model. This meta-analysis was carried out using individual data on patients with curatively resected gastric cancer.

5.1 | Data collection and endpoints

Data from all published randomized trials, with a patient recruitment end date before 2004, and comparing adjuvant chemotherapy with surgery alone for resectable gastric cancers, were searched for electronically using the strategy described in the GASTRIC Group's article[29]. Two endpoints were considered in this study. The primary one was overall survival (OS) defined as the time from randomization to death from any cause or to the last follow-up that was used as a date of censoring. The second endpoint, called disease-free survival (DFS), was the time to relapse, second cancer, or death from any cause, whichever came first.

5.2 | Data description

Overall, the complete available data concerned 3288 patients from 14 randomized clinical trials. From this, 49.7% (1634/3288) of subjects were randomized in the treatment arm. We observed 54% (1763/3288) of DFS related events with a median follow-up time of 495 days (Inter quartile range (IQR): 229.0-994.5 days). More than fifty percent of the patients died during follow-up (1705/3288) with a median follow-up time of 708.0 (IQR: 384.0-1279.0) days. The median follow-up time observed in all patients (in days) was 1675.5 (IQR: 658.8-2485.0) for OS vs 1522.0 [443.0-2429.0] for DFS.

More recently, Oba *et al.*[30] investigated whether DFS is a valid surrogate for OS in these trials of adjuvant chemotherapy for gastric cancer, based on a two-step approach. Individual-level surrogacy was assessed using the Spearman rank correlation coefficient obtained from a bivariate Plackett's copula model combined with trial-specific Weibull models for DFS and OS. They encountered numerical problems with the evaluation at trial-level of the estimated value of R^2_{trial} after adjusting for the estimation error. The adjusted R^2_{trial} was equal to 1.000 (95% CI=0.999-1.000). Given that the adjusted R^2_{trial} value was very close to the upper limit of 1 and given the range of the CI, the authors advised interpreting the obtained numerical results with caution, as they may be easily influenced by numerical errors.

5.3 | Surrogacy evaluation using the proposed joint surrogate model

The results of the one-step joint surrogate model with both ζ and α fixed are described. We observed a fixed significant treatment effect of -0.45 (CI: -0.68 ; -0.21) and -0.26 (CI: -0.49 ; -0.02) on the surrogate and true endpoints. As described by The GASTRIC Group[29], this confirms the benefit of adding chemotherapy to adjuvant fluorouracil-based chemotherapy in terms of both OS and DFS. The high variance parameters observed in the individual random effect, θ=7.52 (SE=0.25), suggested a strong heterogeneity and a high positive association between endpoints at the individual-level. Furthermore, the variances (γ=3.17 (SE=0.44)) observed in the trial random effect also showed a significant heterogeneity at trial-level on the baseline hazard function.

The individual-level surrogacy between DFS and OS was evaluated with a Kendall's τ=0.67 (95%CI : 0.65-0.70). At the trial-level, we found a high value of the coefficient of determination, equal to R^2_{trial}=0.990 (95%CI : 0.843-1.137). These results were close to those obtained using both the two-step copula model and the one-step Poisson approach for the trial-level association. However, compared to existing approaches[13,14], we observed a lower confidence interval in R^2_{trial}, as described in Table 5. Moreover, the proposed estimated Kendall's τ based on the one-step joint surrogate model was lower than those obtained using previous models, except for Hougaard's copula model. However, it remained high. These results suggest that DFS is valid as a surrogate endpoint for OS both at the individual-level and at the trial-level.

We noticed that the estimation of the joint surrogate model on the resectable gastric cancers data took 9.13 minutes, using a processor `Intel(R) Xeon(R) CPU E5-2690 v2 @ 3.00GHz` including 40 cores and a read Only Memory (RAM) of 378 Go.

6 | DISCUSSION

In this work we propose a novel one-step validation joint model to evaluate surrogate endpoints in multiple randomized cancer clinical trials with failure-time endpoints. This model extends that of Burzykowski *et al.*[15], which is the reference in this context, with the advantage of taking into account the heterogeneities of both the individual and the trial-levels in one step. The proposed joint surrogate model presents some particularities compared to the Poisson-based approach. We used power parameters (α and ζ) that distinguish both individual and trial-level heterogeneities between the surrogate and the true endpoint. We then proposed a more flexible model to deal with negatively correlated endpoints, even though generally in oncology surrogate endpoints are more often positively correlated to the true endpoint or overall survival. Moreover, instead of considering a parametric form of the baseline risk function, as in the original model[15], or a piecewise constant[13,31], we assume a more flexible semi-parametric approach. A continuous shape of the hazard function using spline approximation was used with the aim of obtaining smoothed estimates, which represent incidence and mortality rates in epidemiology. Consequently, we used the penalized maximum likelihood method and the Marquard optimization algorithm which is known to be more stable than the Newton-Raphson algorithm in complex problems, and allows a direct variance estimator of random effects as the parameters of the baseline risk functions[32]. Although the heterogeneity between trials of the baseline hazard function could be taken into account using a trial-specific baseline hazards function (stratification), that require less assumptions, we chose to use a shared random effect, given the high number of parameters in this model, with the multiplicity of included trials.

Several integration methods are considered in this work and the compared results are described. It follows that the proposed joint surrogate model was robust enough with different integral approximations. We drew the same conclusions in terms of surrogacy evaluation regardless of integration method, although computational time was high with full integration approaches (MC, PGH). The use of MC to integrate over trial random effects serves to reduce computational time independently of the random-effects dimension, in contrast to the Gaussian Hermite quadrature which requires exponential computational time with the increase in the number of points. We found satisfactory results by considering just 300 Monte-Carlo samples, although a gain in variability could be achieved by increasing the number of samples. Using full PGH with only 12 points, the results obtained were close to those of other integration methods in terms of surrogacy evaluation, although we found a drastic drop in the coverage rate. However, by increasing the number of quadrature points to 15, this deficiency could be overcome. Rizopoulos[33] obtained satisfactory results with computing gains by adopting the pseudo-adaptive rule with 3 points rather than the standard Gauss-Hermite rule with 15 points. More recently, Ferrer *et al.*[34] made a similar observation using two-step pseudo-adaptive approaches with 9 quadrature points, in a context of a high random-effects dimension. Using PGH to integrate over individual-level random effects, comparable results with standard GH were found. This could be due to the fact that we just considered one observation per subject. Rizopoulos[33] suggested better results by increasing the observation numbers per subject.

We also observed a robustness of the joint surrogate model when the censoring rate was increased, although with high censoring, the proportion of non-convergent cases was close to 10%. These results were different and more satisfactory in terms of bias and percentage of coverage than those previously found with a two-step copula-based approach [10,35]. In addition, the proposed model described in this work performed surrogacy evaluation better than existing approaches in terms of biases and MSE. The existing approaches globally underestimated the individual-level and overestimated the trial-level evaluation. Based on study simulation, the one-step Poisson-based model was the most concerned by this problem, as described in previous works [13]. Following this observation Rotolo *et al.* [13] advised using their model without accounting for heterogeneity in the baseline risk, although this is known to be an unrealistic assumption. These results suggest that more vigilance is needed regarding the conclusions drawn on the quality of the surrogate endpoint from previous models, especially when the observed R^2_{trial} is moderated. Moreover, when data were generated using a Clayton copula model, observed results suggested a robustness of the joint surrogate model to a potential misspecification. The two-step approaches with the last simulation design show, as expected, good estimation properties, but, always faced convergence issues. Thereby, regardless of the underlying model to the data, one can expected stable results compared to existing approaches that revealed moderate properties on R^2_{trial} in case of data generation based on joint surrogate model.

In the comparative analysis, as expected, we observed convergence problem with the two-step copula based approach, mainly in case of few subjects per trial. This problem often arrives during the adjustment of the model at the second stage, on the estimation errors of the trial-specific treatment effect, estimated from the model in the first step. In adjusted model, variance estimators are no guaranteed to be positive as shown in Burzykowski *et al.* [16], which leads to a non-defined estimation of adjusted R^2_{trial}. The proposed joint surrogate model solves this issue by avoiding the adjustment on R^2_{trial}. All the model parameters are estimated in one-step. In addition, the robust *Marquardt* [26] algorithm used to estimate the model parameters is known to be enough stable in complex problems, and therefore lead to minimizing convergence issues. The robustness of the *Marquardt* algorithm is guaranteed, through an update of the Hessian matrix during iterations, to ensure a positive definite matrix. We observed during simulation studies that, using *BFGS* algorithm with the two-step approach, convergence issues were reduced compared to those obtained using the *Bobyqa* algorithm. However, the non-convergence rate was still high, meaning that the main problem was in the level of the adjusted model.

To conclude, we have proposed a more robust approach to validate surrogate endpoints in multiple randomized cancer clinical trials with failure-time endpoints. The proposed integration methods present comparable performances with regard to the surrogacy evaluation criteria. However, the approach based on the combined integration method, which requires less computational time for the estimation of the parameters, is to be preferred over the full Monte-Carlo or full pseudo-adaptive integration. In the event of strong individual-level association, full Monte Carlo integration is to be avoided, mainly when data include small trial sizes. The model presents a slight difficulty in estimating the variance of the trial random effects associated with the baseline risk, especially in the event of a small number of trials. Particular attention should be paid to the estimated value of R^2_{trial} in the event of weak trial-level association, given that, although we proposed a better approach to evaluate surrogate endpoints, bias problems persisted in the estimation of R^2_{trial}. Overall we observed a moderate 95% coverage rate of R^2_{trial}. However, this could be improved with an increased sample size. This new joint surrogate model reduces the numerical problems encountered with the standard copula model-based approach and will provide satisfactory conclusions concerning surrogate endpoint evaluation. To solve the question of surrogacy validity both at the individual and at the trial-level, an index combining Kendall's τ and R^2_{trial} with an appropriate threshold should be considered.

7 | SOFTWARE

All the models were estimated using an extension of the R package `frailtypack` [36]. The corresponding R package implementing the new joint surrogate model, together with a sample input data set and complete documentation is available on CRAN. The two-steps copula models and the one-step Poisson model were estimated using the R package `surrosurv` [37].

DATA ACCESSIBILITY

The authors are making the data associated with this paper available

SUPPLEMENTARY INFORMATION LEGEND

The reader is referred to the Supplementary information for technical appendices and additional simulations referenced in Sections 2,3 and 4.. Elements included are:

- **A.** Log-likelihood construction associated with the proposed joint surrogate model;
- **B.** Kendall's τ derivation;
- **C.** Non-convergence case management procedure;
- **D.** Additional simulation results.

ACKNOWLEDGMENTS

This work was supported by the Association pour la Recherche sur le Cancer, Grant/Award Number: PJA20161205147; Institut National du Cancer, Grant/Award Number: 2017-125; Institut national de la santé et de la recherche médicale; Région Aquitaine

The authors thank the GASTRIC Group for permission to use their data. The investigators who contributed to GASTRIC are listed in [30,29]. The GASTRIC Group data used in the present paper can be downloaded from [14] for research purposes, under the condition that (a) the research is scientifically appropriate, (b) the confidentiality of individual patient data is protected, (c) the results of the analyses are shared with the GASTRIC Group prior to public communication, (d) the source of data is fully acknowledged as above, and (e) resulting data and results are further shared with the research community. Computer times were provided by the computing facilities of MCIA (Mésocentre de Calcul Intensif Aquitain) at the Université de Bordeaux and the Université de Pau et des Pays de l'Adour.

References

1. Fleming TR, Prentice RL, Pepe MS, Glidden D. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and aids research. *Stat Med.* 1994; 13(9): 955–968.
2. Ellenberg SS, Hamilton JM. Surrogate endpoints in clinical trials: Cancer. *Stat Med.* 1989; 8(4): 405–413.
3. Baker SG, Kramer BS. A perfect correlate does not a surrogate make. *BMC Med Res Methodol.* 2003; 3(1): 16.
4. Fleming TR, DeMets DL. Surrogate end points in clinical trials: Are we being misled?. *Ann Intern Med.* 1996; 125(7): 605-613.
5. Prentice RL. Surrogate endpoints in clinical trials: Definition and operational criteria. *Stat Med.* 1989; 8(4): 431–440.
6. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Stat Med.* 1992; 11(2): 167–178.
7. Buyse M, Molenberghs G. The validation of surrogate endpoints in randomized experiments. *Biometrics.* 1998; 54: 1014–1029.
8. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics.* 2000; 1(1): 49–67.
9. Alonso A, Molenberghs G, Geys H, Buyse M, Vangeneugden T. A unifying approach for surrogate marker validation based on Prentice's criteria. *Stat Med.* 2006; 25(2): 205–221.
10. Renfro LA, Shi Q, Sargent DJ, Carlin BP. Bayesian adjusted R2 for the meta-analytic evaluation of surrogate time-to-event endpoints in clinical trials. *Stat Med.* 2012; 31(8): 743–761.
11. Liu Y, Taylor JMG, Elliott MR, Sargent DJ. Causal assessment of surrogacy in a meta-analysis of colorectal cancer trials. *Biostatistics.* 2011; 12(3): 478–492.

12. Frangakis CE, Rubin DB. Principal Stratification in Causal Inference. *Biometrics.* 2002; 58(1): 21–29.

13. Rotolo F, Paoletti X, Burzykowski T, Buyse M, Michiels S. A Poisson approach to the validation of failure time surrogate endpoints in individual patient data meta-analyses. *Stat Methods Med Res.* 2017; 0(0): 1-14.

14. Buyse M, Molenberghs G, Paoletti X, et al. Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. *Biom J.* 2016; 58(1): 104–132.

15. Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *J R Stat Soc Ser C Appl Stat.* 2001; 50(4): 405–422.

16. Burzykowski T, Molenberghs G, Buyse M, Geys H. *The evaluation of Surrogate Endpoints*. Springer, New-york, NK . 2005.

17. Emura T, Nakatochi M, Murotani K, Rondeau V. A joint frailty-copula model between tumour progression and death for meta-analysis. *Stat Methods Med Res.* 2015; 0(0): 1-21.

18. Rondeau V, Pignon JP, Michiels S. A joint model for the dependence between clustered times to tumour progression and deaths: A meta-analysis of chemotherapy in head and neck cancer. *Stat Methods Med Res.* 2015; 24(6): 711-729.

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

20. Yu Z, Liu L, Bravata DM, Williams LS. Joint model of recurrent events and a terminal event with time-varying coefficients. *Biom J.* 2014; 56(2): 183–197.

21. Huang X, Liu L. A Joint Frailty Model for Survival and Gap Times Between Recurrent Events. *Biometrics.* 2007; 63(2): 389–397.

22. O'Sullivan F. Fast Computation of Fully Automated Log-Density and Log-Hazard Estimators. *SIAM J Sci and Stat Comput.* 1988; 9(2): 363-379.

23. Joly P, Commenges D, Letenneur L. A Penalized Likelihood Approach for Arbitrarily Censored and Truncated Data: Application to Age-Specific Incidence of Dementia. *Biometrics.* 1998; 54(1): 185-194.

24. Ramsay JO. Monotone Regression Splines in Action. *Statist Sci.* 1988; 3(4): 425-441.

25. Rondeau V, Commenges D, Joly P. Maximum penalized likelihood estimation in frailty models. *Lifetime Data Anal.* 2003; 9(2): 139-153.

26. Marquardt DW. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *SIAM J Appl Math.* 1963; 11(2): 431-441.

27. Kendall MG. *Rank correlation methods*. Oxford, England: Hafner Publishing Co. 2nd ed. 1995.

28. Duchateau L, Janssen P. *The Frailty Model*. Springer . 2008.

29. The GASTRIC Group . Benefit of adjuvant chemotherapy for resectable gastric cancer: A meta-analysis. *JAMA*. 2010; 303(17): 1729-1737.

30. Oba K, Paoletti X, Alberts S, et al. Disease-Free Survival as a Surrogate for Overall Survival in Adjuvant Trials of Gastric Cancer: A Meta-Analysis. *J Natl Cancer Inst.* 2013; 105(21): 1600-1607.

31. Crowther MJ, Riley RD, Staessen JA, Wang J, Gueyffier F, Lambert PC. Individual patient data meta-analysis of survival data using Poisson regression models. *BMC Med Res Methodol.* 2012; 12(1): 34.

32. Mazroui Y, Mathoulin-Pelissier S, Soubeyran P, Rondeau V. General joint frailty model for recurrent event data with a dependent terminal event: Application to follicular lymphoma data. *Stat Med.* 2012; 31(11-12): 1162–1176.

33. Rizopoulos D. Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive Gaussian quadrature rule. *Comput Statist Data Ana.* 2012; 56(3): 491 - 501.

34. Ferrer L, Rondeau V, Dignam J, Pickles T, Jacqmin-Gadda H, Proust-Lima C. Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer. *Stat Med.* 2017; 35(22): 3933-3948.

35. Shi Q, Renfro LA, Bot BM, Burzykowski T, Buyse M, Sargent DJ. Comparative assessment of trial-level surrogacy measures for candidate time-to-event surrogate endpoints in clinical trials. *Comput Statist Data Ana.* 2011; 55(9): 2748 - 2757.

36. Kròl A, Mauguen A, Mazroui Y, Laurent A, Michiels S, Rondeau V. Tutorial in Joint Modeling and Prediction: A Statistical Software for Correlated Longitudinal Outcomes, Recurrent Events and a Terminal Event. *J Stat Softw.* 2017; 81(3): 1–52.

37. Rotolo F, Paoletti X, Michiels S. surrosurv: An R package for the evaluation of failure time surrogate endpoints in individual patient data meta-analyses of randomized clinical trials. *Comput Methods Programs Biomed.* 2018; 155: 189 - 198.

TABLE 1 Estimates (Mean), mean of the standard errors (SE), empirical standard errors (SD) and percentage of coverage (CP). Estimation based on a Monte-Carlo (MC), a combination of MC with non-adaptive Gaussian Hermite (MC-GH) and with pseudo-adaptive Gaussian Hermite (MC-PGH) integration[‡], for M=500 samples.

		N=600 subjects, G=30 trials				N=600 subjects, G=10 trials			
Parameters	True	Mean	SD	SE	CP	Mean	SD	SE	CP
				MC-PGH[†]					
θ	1.000	1.211	0.545	0.526	95	1.165	0.578	0.508	92
ζ	1.000	1.016	0.584	0.384	91	1.285	1.613	0.404	88
γ	0.800	0.935	0.397	0.336	93	0.852	0.540	0.319	77
α	1.000	0.984	0.187	0.203	95	1.003	0.325	0.220	90
$\sigma^2_{v_S}$	0.700	0.768	0.436	0.385	89	0.772	0.599	0.380	81
$\sigma^2_{v_T}$	0.700	0.737	0.477	0.366	89	0.810	0.852	0.384	77
$\sigma_{v_{ST}}$	0.630	0.652	0.351	0.296	90	0.664	0.491	0.297	75
β_S	-1.250	-1.249	0.276	0.247	92	-1.290	0.393	0.293	86
β_T	-1.250	-1.209	0.234	0.236	92	-1.259	0.371	0.286	86
R^2_{trial}	0.810	0.815	0.194	0.185	75	0.801	0.233	0.174	68
τ	0.378	0.383	0.039	-	91	0.367	0.063	-	73
				MC-GH[†]					
θ	1.000	1.192	0.486	0.502	95	1.188	0.475	0.502	95
ζ	1.000	0.996	0.368	0.385	91	1.010	0.443	0.388	91
γ	0.800	0.929	0.385	0.329	93	0.867	0.534	0.321	77
α	1.000	0.987	0.181	0.203	95	0.995	0.235	0.216	93
$\sigma^2_{v_S}$	0.700	0.761	0.423	0.378	89	0.785	0.608	0.385	82
$\sigma^2_{v_T}$	0.700	0.730	0.448	0.364	89	0.749	0.585	0.381	79
$\sigma_{v_{ST}}$	0.630	0.648	0.350	0.294	90	0.666	0.492	0.302	76
β_S	-1.250	-1.247	0.270	0.244	92	-1.296	0.376	0.295	88
β_T	-1.250	-1.211	0.233	0.236	92	-1.261	0.371	0.288	86
R^2_{trial}	0.810	0.815	0.193	0.189	75	0.809	0.226	0.172	68
τ	0.378	0.383	0.037	-	92	0.374	0.05	-	85
				MC[†]					
θ	1.000	1.109	0.288	0.243	96	1.100	0.266	0.240	96
ζ	1.000	0.996	0.371	0.385	92	1.014	0.428	0.397	93
γ	0.800	0.948	0.467	0.342	92	0.923	0.735	0.341	73
α	1.000	0.983	0.184	0.203	95	0.993	0.233	0.217	94
σ^2_S	0.700	0.784	0.521	0.395	89	0.778	0.681	0.378	80
σ^2_T	0.700	0.725	0.441	0.361	89	0.733	0.614	0.372	78
$\sigma_{v_{ST}}$	0.630	0.651	0.353	0.295	89	0.642	0.461	0.293	75
β_S	-1.250	-1.258	0.300	0.250	92	-1.301	0.400	0.295	87
β_T	-1.250	-1.209	0.230	0.234	93	-1.251	0.368	0.283	86
R^2_{trial}	0.810	0.814	0.193	0.187	76	0.803	0.230	0.172	68
τ	0.378	0.381	0.038	-	91	0.374	0.058	-	79

[†]No convergence issue, except with MC where 1 dataset did not converged, [‡] 300 samples for MC and 20 or 32 quadrature points for GH or PGH.

TABLE 2 Estimates (Mean), mean of the standard errors (SE), empirical standard errors (SD) and percentage of coverage (CP). Estimation based on a combination of Monte-Carlo (MC) with pseudo-adaptive Gaussian Hermite (PGH) integration[‡‡], for M=500 samples.

		High individual and trial-level association							
		G=30 trials				G=10 trials			
Parameters	True	Mean	SD	SE	CP	Mean	SD	SE	CP
Moderate censoring rate(∼40%)[†], N=600 subjects									
θ	3.500	3.648	0.771	0.670	92	3.674	0.699	0.703	94
ζ	1.500	1.615	1.173	0.313	81	1.527	0.660	0.331	84
γ	2.500	2.703	0.972	0.774	89	2.741	2.215	0.880	71
α	1.000	0.997	0.233	0.181	84	0.991	0.201	0.197	89
$\sigma^2_{v_S}$	0.700	0.744	0.523	0.498	88	0.762	0.686	0.477	82
$\sigma^2_{v_T}$	0.700	0.808	0.759	0.661	89	0.789	0.803	0.606	82
$\sigma_{v_{ST}}$	0.630	0.669	0.538	0.505	88	0.680	0.665	0.471	82
β_S	-1.250	-1.201	0.292	0.261	92	-1.228	0.363	0.322	92
β_T	-1.250	-1.194	0.393	0.350	92	-1.189	0.444	0.404	91
R^2_{trial}	0.810	0.817	0.239	0.404	72	0.816	0.247	0.286	67
τ	0.614	0.613	0.045	-	91	0.614	0.038	-	85
High censoring rate (70%)[‡], N=600 subjects									
θ	3.500	3.777	0.850	1.124	97	3.727	0.816	1.044	96
ζ	1.500	1.577	1.101	0.651	91	1.537	0.838	0.559	92
γ	2.500	2.832	1.117	0.995	92	2.945	2.093	1.194	75
α	1.000	0.989	0.211	0.353	93	0.990	0.208	0.312	95
$\sigma^2_{v_S}$	0.700	0.817	0.696	0.628	88	0.738	0.679	0.547	81
$\sigma^2_{v_T}$	0.700	0.952	1.110	1.023	86	0.794	0.766	0.831	83
$\sigma_{v_{ST}}$	0.630	0.738	0.679	0.651	88	0.652	0.612	0.554	83
β_S	-1.250	-1.230	0.322	0.323	94	-1.214	0.383	0.364	93
β_T	-1.250	-1.237	0.417	0.502	95	-1.207	0.521	0.519	93
R^2_{trial}	0.810	0.806	0.264	0.858	71	0.810	0.268	0.386	64
τ	0.614	0.617	0.047	-	93	0.616	0.042	-	89
Censoring rate (∼40%)[†], increasing sample size, N=1000 subjects									
θ	3.500	3.547	0.629	0.525	93	3.603	0.551	0.557	94
ζ	1.500	1.601	1.026	0.271	85	1.500	0.303	0.281	86
γ	2.500	2.591	0.819	0.629	89	2.789	2.020	0.722	56
α	1.000	1.016	0.210	0.154	86	0.999	0.179	0.166	88
$\sigma^2_{v_S}$	0.700	0.708	0.433	0.367	87	0.740	0.563	0.373	77
$\sigma^2_{v_T}$	0.700	0.733	0.572	0.478	87	0.721	0.631	0.460	77
$\sigma_{v_{ST}}$	0.630	0.627	0.401	0.369	88	0.644	0.531	0.364	77
β_S	-1.250	-1.198	0.244	0.218	91	-1.235	0.325	0.275	88
β_T	-1.250	-1.199	0.294	0.292	93	-1.217	0.372	0.339	90
R^2_{trial}	0.810	0.812	0.206	0.259	78	0.809	0.232	0.216	69
τ	0.614	0.613	0.041	-	91	0.616	0.03	-	78

[†]Convergence issues less than 3%, [‡]Convergence issues ∼ 10%, [‡‡] 300 samples for MC and 15 or 20 quadrature points for PGH.

TABLE 3 Estimates (Mean), Bias, and Mean square errors (MSE): comparison between two-step (Clayton and Plackett copulas) and one-step (Poisson and proposed joint surrogate) approaches, for M = 500 samples.

		600 subjects, 30 trials				600 subjects, 10 trials			
Parameters	True	Mean	Bias	MSE	n(%)‡‡	Mean	Bias	MSE	n(%)‡‡
High individual and trial-level association									
Clayton†					266(53)				10(2)
τ	0.614	0.565	0.049	0.003		0.543	0.071	0.006	
$R^2_{trial,adj}$	0.810	0.742	0.068	0.082		0.833	-0.023	0.061	
Plackett†					274(55)				14(3)
τ	0.614	0.559	0.055	0.004		0.538	0.076	0.006	
$R^2_{trial,adj}$	0.810	0.811	-0.001	0.069		0.844	-0.034	0.062	
Poisson‡					1(0)				8(2)
τ	0.614	0.467	0.147	0.022		0.468	0.146	0.022	
R^2_{trial}	0.810	0.853	-0.043	0.048		0.838	-0.028	0.060	
Joint Surrogate††					1(0)				4(1)
τ	0.614	0.613	0.001	0.002		0.614	$< 10^{-3}$	0.001	
R^2_{trial}	0.810	0.817	-0.007	0.057		0.816	-0.006	0.061	
High individual weak trial-level association									
Clayton†					276(55)				12(2)
τ	0.614	0.568	0.046	0.003		0.543	0.071	0.006	
$R^2_{trial,adj}$	0.360	0.583	-0.223	0.153		0.625	-0.265	0.199	
Plackett†					303(61)				12(2)
τ	0.614	0.561	0.053	0.003		0.538	0.076	0.006	
$R^2_{trial,adj}$	0.360	0.617	-0.257	0.181		0.631	-0.271	0.202	
Poisson‡					13(3)				12(2)
τ	0.614	0.460	0.154	0.024		0.462	0.152	0.024	
R^2_{trial}	0.360	0.679	-0.319	0.191		0.644	-0.284	0.194	
Joint Surrogate††					1(0)				0(0)
τ	0.614	0.609	0.005	0.004		0.606	0.008	0.005	
R^2_{trial}	0.360	0.542	-0.182	0.131		0.520	-0.160	0.134	
Weak individual and trial-level association									
Clayton†					236(47)				8(2)
τ	0.378	0.277	0.101	0.011		0.257	0.121	0.016	
$R^2_{trial,adj}$	0.360	0.446	-0.086	0.105		0.525	-0.165	0.139	
Plackett†					238(48)				10(2)
τ	0.378	0.278	0.100	0.011		0.271	0.107	0.012	
$R^2_{trial,adj}$	0.360	0.511	-0.151	0.124		0.512	-0.152	0.139	
Poisson‡					1(0)				0(0)
τ	0.378	0.223	0.155	0.025		0.227	0.151	0.023	
R^2_{trial}	0.360	0.586	-0.226	0.127		0.593	-0.233	0.151	
Joint Surrogate††					0(0)				0(0)
τ	0.378	0.382	-0.004	0.001		0.375	0.003	0.004	
R^2_{trial}	0.360	0.442	-0.082	0.080		0.478	-0.118	0.106	

Estimation using: †BFGS algorithm, ‡ Bobyqa algorithm, †† Marquardt algorithm; ‡‡convergence issues.

TABLE 4 Estimates (Mean), Bias, and Mean square errors (MSE): comparison between two-step (Clayton and Plackett copulas) and one-step (Poisson and proposed joint surrogate) approaches, by generating the data from a Clayton copula model, for M = 500 samples.

		600 subjects, 30 trials				600 subjects, 10 trials			
Parameters	True	Mean	Bias	MSE	n(%)[‡‡]	Mean	Bias	MSE	n(%)[‡‡]
Clayton[†]					404 (81)				53 (11)
τ	0.614	0.636	-0.022	0.001		0.619	-0.005	$< 10^{-}3$	
$R^2_{trial,adj}$	0.360	0.369	-0.009	0.046		0.457	-0.097	0.080	
Plackett[†]					409 (82)				52 (10)
τ	0.614	0.573	0.041	0.002		0.554	0.060	0.004	
$R^2_{trial,adj}$	0.360	0.348	0.012	0.051		0.458	-0.098	0.081	
Poisson[‡]					0 (0)				1 (0)
τ	0.614	0.316	0.298	0.090		0.336	0.278	0.081	
R^2_{trial}	0.360	0.350	0.010	0.055		0.370	-0.010	0.083	
Joint Surrogate[††]					2 (0)				23 (5)
τ	0.614	0.520	0.094	0.014		0.523	0.091	0.026	
R^2_{trial}	0.360	0.377	-0.017	0.063		0.407	-0.047	0.083	

Estimation using: [†]BFGS algorithm, [‡]Bobyqa algorithm, [††] Marquardt algorithm; [‡‡]convergence issues.

TABLE 5 Evaluation of individual and trial-level surrogacy of DFS for OS, results comparing the two-step and the Poisson one-step approaches[13] to the proposed one-step joint surrogate model.

Parameters	R^2_{trial}	Kendall's τ
One-step approach		
Proposed joint surrogate	0.99 (0.84-1.14)	0.68 (0.65-0.70)
Poisson	1.00 (0.08-1.00)	0.74 (0.73-0.76)
Two-step approach using R		
Clayton adjusted	0.97 (0.46-1.00)	0.81 (0.80-0.91)
Plackett adjusted	1.00 (0.69-1.00)	0.82 (0.81-0.83)
Hougaard adjusted	0.94 (0.08-1.00)	0.18 (0.17-0.19)

SUPPLEMENTARY INFORMATION

APPENDIX

A LOG-LIKELIHOOD CONSTRUCTION ASSOCIATED WITH THE PROPOSED JOINT SURROGATE MODEL

Let S_{ij} and T_{ij} denote the follow-up times associated respectively with surrogate and true endpoints for subject j belonging in trial i. Similarly, δ_{ij} and δ_{ij}^{*} denote the progression and death indicators.

Let $\Phi = (\sigma^2_{\upsilon_S}, \sigma^2_{\upsilon_T}, \sigma_{\upsilon_{ST}}, \theta, \gamma, \lambda_{0T}(.), \lambda_{0S}(.), \beta_{Sk}, \beta_{Tk})$ denote the vector containing all the unknown parameters of the proposed joint surrogate model. The conditional contribution to the likelihood $L_i(\Phi|\omega_{ij}, u_i, \upsilon_{S_i}, \upsilon_{T_i})$ for all the n_i subjects in cluster i is obtained based on the product of the conditional likelihood for each subject by:

$$L_i(\Phi|.) = \prod_{j=1}^{ni} \left\{ \lambda_{Sij}(T_{ij}|\omega_{ij}, u_i, \upsilon_{S_i})^{\delta_{ij}} \exp(-\int_0^{T_{ij}} \lambda_{Sij}(t|\omega_{ij}, u_i, \upsilon_{S_i}, Z_{ij})dt) \times \right.$$
$$\left. \lambda_{Tij}(D_{ij}|\omega_{ij}, u_i, \upsilon_{T_i})^{\delta^{*}_{ij}} \exp(-\int_0^{D_{ij}} \lambda_{Tij}(t|\omega_{ij}, u_i, \upsilon_{T_i}, Z_{ij})dt) \right\}.$$

The marginal contribution to the likelihood $L_i(\Phi)$ for each trial i is obtained by integrating the conditional likelihood upon the random effects:

$$L_i(\Phi) = \int_{u_i} \int_{\upsilon_{S_i}} \int_{\upsilon_{T_i}} \left\{ \prod_{j=1}^{ni} \int_{\omega_{ij}} L_i(\Phi|.) f(\omega_{ij}) f(\upsilon_{S_i}, \upsilon_{T_i}) f(u_i) d\omega_{ij} \right\} d\upsilon_{S_i} d\upsilon_{T_i} du_i,$$

where the density probability functions in cluster i are given by:

$$f(\upsilon_{S_i}, \upsilon_{T_i}) = \frac{1}{(2\pi)\sqrt{|\Sigma_{\upsilon_i}|}} \exp\left[-\frac{1}{2}(\upsilon_{S_i}, \upsilon_{T_i})\Sigma_{\upsilon_i}{}^{-1}(\upsilon_{S_i}, \upsilon_{T_i})'\right],$$

$$f(\omega_{ij}, \theta) = \frac{1}{\sqrt{2\pi\theta}} \exp(-\frac{1}{2}\frac{\omega^2_{ij}}{\theta}),$$

and

$$f(u_i, \gamma) = \frac{1}{\sqrt{2\pi\gamma}} \exp(-\frac{1}{2}\frac{u_i^2}{\gamma})$$

respectively for correlated normal distributed random effects $(\upsilon_{S_i}, \upsilon_{T_i})$, for individual normal distributed random effect ω_{ij} and trial random effects u_i. Therefore, using the previous expressions, we obtained the full marginal likelihood associated with the proposed joint surrogate model by using $l(\Phi) = \log \prod_{i=1}^{G} L_i(\Phi)$, where G is the total number of considered trials.

B KENDALL'S τ DERIVATION

We assume the joint conditional distributions of S_{ij} and T_{ij} after adjusting for both the trial and the treatment effects defined by:

$$\begin{cases} \lambda_{S,ij}(t|\omega_{ij}, u_i, Z_{ijk}) = \lambda_{0S}(t) \exp(\omega_{S_{ij}} + u_i + \sum_{k=1}^{p} \beta_{Sk} Z_{ijk}) \\ \lambda_{T,ij}(t|\omega_{ij}, u_i, Z_{ijk}) = \lambda_{0T}(t) \exp(\zeta \omega_{T_{ij}} + \alpha u_i + \sum_{k=1}^{p} \beta_{Tk} Z_{ijk}) \end{cases} \quad \text{(B1)}$$

where

$$\omega_{ij} \sim N(0, \theta), \quad u_i \sim N(0, \gamma), \quad \text{(B2)}$$

in which θ, γ, α and ζ are estimated using the proposed joint surrogate model. Kendal's τ could be rewrite by:

$$\tau = E\{sign[(S_{ij} - S_{i'j'})(T_{ij} - T_{i'j'})]\},$$

where $sign(x) = -1, 0$ and 1 for $x < 0$, $x = 0$ and $x > 0$ respectively. Assume two randomly chosen individual j and j' from two randomly chosen cluster i and i' and assume that for bivariate survival data, the covariate information is the same in each cluster, i.e. $X_i = (x_{ij}, x_{i'j'}) = (x_i, x_{i'}) = X$. An alternative formulation for continuous distributions is given by the Probability of matching pairs minus the probability of discordant pairs as:

$$\begin{aligned}\tau &= P[(S_{ij} - S_{i'j'})(T_{ij} - T_{i'j'}) > 0] - P[(S_{ij} - S_{i'j'})(T_{ij} - T_{i'j'}) < 0] \\ &= 2P[(S_{ij} - S_{i'j'})(T_{ij} - T_{i'j'}) > 0] - 1 \\ &= 2\{P[(S_{ij} > S_{i'j'}) \cap (T_{ij} > T_{i'j'})] + P[(S_{i'j'} > S_{ij}) \cap (T_{i'j'} > T_{ij})]\} - 1. \end{aligned} \tag{B3}$$

One can show that (B3) could be rewrite as :

$$\tau = 2\left\{ \int_0^\infty \int_0^\infty S_{S_{i'j'},T_{i'j'}}(s,t) f_{S_{ij},T_{ij}}(s,t) ds dt + \int_0^\infty \int_0^\infty S_{S_{ij},T_{ij}}(s,t) f_{S_{i'j'},T_{i'j'}}(s,t) ds dt \right\} - 1,$$

where $f(s,t) = \int_u f(s,t|u) f_u du$, and $S(s,t) = \int_u S(s,t|u) f_u du$ are the marginal density and survival functions, and f_u the density probability function of the random effect u with $u = (\omega_{S_{ij}}, \omega_{T_{ij}}, u_{S_i}, u_{T_i})$. Otherwise, $f_{S_{ij},T_{ij}}(s,t) = \lambda_{S_{ij},T_{ij}}(s,t) S_{S_{ij},T_{ij}}(s,t)$ and conditional to random effects, T_{ij} and S_{ij} are independents $\forall i, j$.

Thereby, for the reduced joint surrogate model (B1), from the previous formulation ok Kendall's τ, one can show that

$$\begin{aligned}\tau = 2 \int_{u_i} \int_{\omega_{ij}} \int_{u_{i'}} \int_{\omega_{i'j'}} & \frac{\exp(\omega_{ij} + u_i + \zeta\omega_{ij} + \alpha u_i) + \exp(\omega_{i'j'} + u_{i'} + \zeta\omega_{i'j'} + \alpha u_{i'})}{(\exp(\omega_{i'j'} + u_{i'}) + \exp(\omega_{ij} + u_i))(\exp(\zeta\omega_{i'j'} + \alpha u_{i'}) + \exp(\zeta\omega_{ij} + \alpha u_i))} \\ & \frac{1}{\sqrt{2\pi\theta}} \exp\left[-\frac{1}{2}\frac{\omega_{i'j'}^2}{\theta}\right] \frac{1}{\sqrt{2\pi\gamma}} \exp\left[-\frac{1}{2}\frac{u_{i'}^2}{\gamma}\right] d\omega_{i'j'} du_{i'} \\ & \frac{1}{\sqrt{2\pi\theta}} \exp\left[-\frac{1}{2}\frac{\omega_{ij}^2}{\theta}\right] \frac{1}{\sqrt{2\pi\gamma}} \exp\left[-\frac{1}{2}\frac{u_i^2}{\gamma}\right] d\omega_{ij} du_i - 1.\end{aligned}$$

C NON-CONVERGENCE CASE MANAGEMENT PROCEDURE

Special attention must be given to initializing model parameters, the choice of the number of spline knots, the smoothing parameters and the number of quadrature points to solve convergence issues. We firstly initialized parameters with the true values used during data generation; the number of quadrature point was set to 20 (12 for Full PGH); the number of knots to 6 and the smoothing parameters were based on cross-validation. When numerical or convergence problems were encountered, the model was fitted again using a combination of the following strategies: varying the number of quadrature point (20 to 32 or 12 to 15), dividing or multiplying k_1 or k_2 by 10 or 100 according to their preceding values, or using parameter vectors obtained during the last iteration (with a modification of the number of quadrature points and smoothing parameters). Using this strategy, we usually obtained the rejection rate described above. A sensitivity analysis was conducted without this strategy, and similar results were obtained on the converged samples with about a 23% rejection rate.

D ADDITIONAL SIMULATION RESULTS

TABLE D1 Estimates (Mean), mean of the standard errors (SE), empirical standard errors (SD) and percentage of coverage (CP). Estimation based on a pseudo-adaptive Gaussian-Hermite (PGH) integration with 12 quadrature points, for M=400 samples.

		N=600 subjects, G=30 trials			
Parameters	True	Mean	SD	SE	CP
		PGH			
θ	1.000	1.011	0.422	0.365	89
ζ	1.000	1.433	1.611	0.374	86
γ	0.800	0.760	0.260	0.272	90
α	1.000	1.096	0.364	0.192	91
$\sigma^2_{v_S}$	0.700	0.646	0.344	0.322	83
$\sigma^2_{v_T}$	0.700	0.738	0.503	0.364	84
$\sigma_{v_{ST}}$	0.630	0.585	0.301	0.290	86
β_S	-1.250	-1.256	0.247	0.222	93
β_T	-1.250	-1.313	0.359	0.236	91
R^2_{trial}	0.810	0.823	0.204	0.137	54
τ	0.378	0.368	0.042	-	88

TABLE D2 Estimates (Mean), mean of the standard errors (SE), empirical standard errors (SD) and percentage of coverage (CP). Estimation based on a combination of Monte-Carlo (MC) with non-adaptive Gaussian Hermite (GH) and with pseudo-adaptive Gaussian Hermite (PGH) integration‡, for M=500 samples.

		N=600 subjects, G=30 trials				N=600 subjects, G=10 trials			
Parameters	True	Mean	SD	SE	CP	Mean	SD	SE	CP
		MC-PGH†							
θ	3.500	3.648	0.771	0.67	92	3.674	0.699	0.703	94
ζ	1.500	1.615	1.173	0.313	81	1.527	0.660	0.331	84
γ	2.500	2.703	0.972	0.774	89	2.741	2.215	0.880	71
α	1.000	0.997	0.233	0.181	84	0.991	0.201	0.197	89
$\sigma^2_{v_S}$	0.700	0.744	0.523	0.498	88	0.762	0.686	0.477	82
$\sigma^2_{v_T}$	0.700	0.808	0.759	0.661	89	0.789	0.803	0.606	82
$\sigma_{v_{ST}}$	0.630	0.669	0.538	0.505	88	0.680	0.665	0.471	82
β_S	-1.250	-1.201	0.292	0.261	92	-1.228	0.363	0.322	92
β_T	-1.250	-1.194	0.393	0.350	92	-1.189	0.444	0.404	91
R^2_{trial}	0.810	0.817	0.239	0.404	72	0.816	0.247	0.286	67
τ	0.614	0.613	0.045	-	91	0.614	0.038	-	85
		MC-GH†							
θ	3.500	3.397	0.528	0.526	94	3.372	0.456	0.534	95
ζ	1.500	1.583	0.324	0.319	85	1.594	0.318	0.337	88
γ	2.500	2.666	0.971	0.753	88	2.617	1.644	0.834	68
α	1.000	1.041	0.192	0.187	87	1.048	0.197	0.203	90
$\sigma^2_{v_S}$	0.700	0.731	0.518	0.481	88	0.714	0.601	0.442	81
$\sigma^2_{v_T}$	0.700	0.850	0.753	0.704	91	0.820	0.791	0.648	84
$\sigma_{v_{ST}}$	0.630	0.690	0.550	0.515	90	0.672	0.613	0.470	84
β_S	-1.250	-1.183	0.274	0.255	92	-1.197	0.351	0.314	91
β_T	-1.250	-1.229	0.386	0.360	94	-1.227	0.450	0.419	92
R^2_{trial}	0.810	0.826	0.232	0.356	71	0.814	0.254	0.307	64
τ	0.614	0.614	0.025	-	91	0.612	0.032	-	82

†Convergence issues less than 1%, ‡ 300 samples for MC and 20 or 32 quadrature points for GH or PGH.

TABLE D3 Estimates (Mean), mean of the standard errors (SE), empirical standard errors (SD) and percentage of coverage (CP). Estimation based on a combination of Monte-Carlo (MC) with non-adaptive Gaussian Hermite (GH) integration[‡], for M=500 samples.

		G=30 trials				G=10 trials			
Parameters	True	Mean	SD	SE	CP	Mean	SD	SE	CP
Weak individual and trial level association[†], N=600 subjects									
θ	1.000	1.198	0.497	0.510	96	1.184	0.471	0.506	95
ζ	1.000	0.987	0.359	0.392	90	0.998	0.384	0.387	90
γ	0.800	0.939	0.417	0.333	92	0.965	0.684	0.343	74
α	1.000	0.988	0.194	0.212	95	0.988	0.227	0.220	93
$\sigma^2_{v_S}$	0.700	0.754	0.440	0.381	88	0.798	0.610	0.378	78
$\sigma^2_{v_T}$	0.700	0.709	0.447	0.365	86	0.847	0.800	0.418	77
$\sigma_{v_{ST}}$	0.420	0.438	0.330	0.264	88	0.495	0.552	0.256	73
β_S	-1.250	-1.247	0.274	0.243	91	-1.300	0.380	0.282	86
β_T	-1.250	-1.246	0.238	0.241	93	-1.293	0.390	0.281	85
R^2_{trial}	0.360	0.442	0.270	0.249	77	0.482	0.303	0.204	60
τ	0.378	0.382	0.038	-	92	0.380	0.055	-	79
High individual, weak trial level association[†], N=600 subjects									
θ	3.500	3.381	0.525	0.525	93	3.351	0.452	0.528	96
ζ	1.500	1.585	0.322	0.309	87	1.605	0.320	0.325	86
γ	2.500	2.651	0.986	0.745	88	2.695	1.804	0.825	64
α	1.000	1.040	0.196	0.184	87	1.052	0.198	0.201	91
$\sigma^2_{v_S}$	0.700	0.734	0.527	0.479	86	0.727	0.622	0.439	81
$\sigma^2_{v_T}$	0.700	0.827	0.746	0.664	87	0.831	0.850	0.616	81
$\sigma_{v_{ST}}$	0.420	0.486	0.509	0.465	86	0.463	0.542	0.409	79
β_S	-1.250	-1.188	0.278	0.255	92	-1.201	0.361	0.306	89
β_T	-1.250	-1.263	0.376	0.358	94	-1.259	0.463	0.413	92
R^2_{trial}	0.360	0.537	0.314	0.485	77	0.525	0.325	0.408	68
τ	0.614	0.613	0.025	-	90	0.612	0.033	-	78
Weak individual, high trial level association[†], N=1000 subjects									
θ	1.000	1.132	0.369	0.392	96	1.118	0.388	0.389	94
ζ	1.000	0.993	0.306	0.320	92	1.005	0.315	0.324	93
γ	0.800	0.899	0.310	0.270	92	0.994	1.028	0.275	65
α	1.000	0.995	0.159	0.166	93	1.004	0.176	0.176	95
$\sigma^2_{v_S}$	0.700	0.759	0.365	0.302	88	0.808	0.591	0.300	72
$\sigma^2_{v_T}$	0.700	0.701	0.319	0.287	88	0.763	0.594	0.304	69
$\sigma_{v_{ST}}$	0.630	0.649	0.292	0.238	87	0.669	0.491	0.230	66
β_S	-1.250	-1.241	0.233	0.210	91	-1.313	0.406	0.231	80
β_T	-1.250	-1.210	0.214	0.205	92	-1.287	0.371	0.227	80
R^2_{trial}	0.810	0.822	0.148	0.136	80	0.775	0.223	0.140	67
τ	0.378	0.383	0.030	-	91	0.384	0.055	-	71

[†]Convergence issues less than 0.5%, [‡] 300 samples for MC and 15 or 20 quadrature points for GH.

TABLE D4 Estimates (Mean), Bias, and Mean square errors (MSE): comparison between two-step (Clayton and Plackett copulas) and one-step (Poisson and proposed joint surrogate) approaches, for M=500 samples.

		600 subjects, 30 trials				600 subjects, 10 trials			
Parameters	True	Mean	Bias	MSE	n(%)‡	Mean	Bias	MSE	:n(%)‡
High individual and trial level association									
Clayton†					360(72)				49(10)
τ	0.614	0.457	0.157	0.034		0.499	0.115	0.023	
$R^2_{trial,adj}$	0.810	0.726	0.084	0.105		0.808	0.002	0.074	
Plackett†					357(71)				31(6)
τ	0.614	0.476	0.138	0.020		0.486	0.128	0.019	
$R^2_{trial,adj}$	0.810	0.799	0.011	0.094		0.804	0.006	0.077	
Poisson†					1(0)				8(2)
τ	0.614	0.467	0.147	0.022		0.468	0.146	0.022	
R^2_{trial}	0.810	0.853	-0.043	0.048		0.838	-0.028	0.060	
Joint Surrogate††					1(0)				4(1)
τ	0.614	0.613	0.001	0.002		0.614	0.000	0.001	
R^2_{trial}	0.810	0.817	-0.007	0.057		0.816	-0.006	0.061	
High individual weak trial level association									
Clayton†					361(72)				39(8)
τ	0.614	0.453	0.161	0.036		0.505	0.109	0.022	
$R^2_{trial,adj}$	0.360	0.655	-0.295	0.203		0.583	-0.223	0.195	
Plackett†					365(73)				57(11)
τ	0.614	0.473	0.141	0.022		0.498	0.116	0.017	
$R^2_{trial,adj}$	0.360	0.666	-0.306	0.224		0.589	-0.229	0.194	
Poisson†					13(3)				12(2)
τ	0.614	0.460	0.154	0.024		0.462	0.152	0.024	
R^2_{trial}	0.360	0.679	-0.319	0.191		0.644	-0.284	0.194	
Joint Surrogate††					1(0)				0(0)
τ	0.614	0.609	0.005	0.004		0.606	0.008	0.005	
R^2_{trial}	0.360	0.542	-0.182	0.131		0.520	-0.160	0.134	
Weak individual and trial level association									
Clayton†					340(68)				15(3)
τ	0.378	0.201	0.177	0.032		0.212	0.166	0.029	
$R^2_{trial,adj}$	0.360	0.501	-0.141	0.12		0.504	-0.144	0.137	
Plackett†					260(52)				13(3)
τ	0.378	0.213	0.165	0.030		0.231	0.147	0.026	
$R^2_{trial,adj}$	0.360	0.519	-0.159	0.135		0.497	-0.137	0.135	
Poisson†					1(0)				0(0)
τ	0.378	0.223	0.155	0.025		0.227	0.151	0.023	
R^2_{trial}	0.360	0.586	-0.226	0.127		0.593	-0.233	0.151	
Joint Surrogate††					0(0)				0(0)
τ	0.378	0.382	-0.004	0.001		0.375	0.003	0.004	
R^2_{trial}	0.360	0.442	-0.082	0.080		0.478	-0.118	0.106	

Estimation using: †Bobyqa algorithm, †† Marquardt algorithm; ‡convergence issues.

3.2 Annexes article

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Supplementary Information to: One-step validation method for surrogate endpoints using data from multiple randomized cancer clinical trials with failure-time endpoints

Casimir Ledoux SOFEU | Takeshi Emura | Virginie Rondeau

APPENDIX

A LOG-LIKELIHOOD CONSTRUCTION ASSOCIATED WITH THE PROPOSED JOINT SURROGATE MODEL

Let S_{ij} and T_{ij} denote the observed follow-up times associated respectively with surrogate and true endpoints for subject j belonging in trial i. Similarly, δ_{ij} and δ^*_{ij} denote the progression and death indicators.

Let $\Phi = (\sigma^2_{v_S}, \sigma^2_{v_T}, \sigma_{v_{ST}}, \theta, \gamma, \alpha, \zeta, \lambda_{0T}(.), \lambda_{0S}(.), \beta_{Sk}, \beta_{Tk})$ denote the vector containing all the unknown parameters of the proposed joint surrogate model. The conditional contribution to the likelihood $L_i(\Phi|\omega_{ij}, u_i, v_{S_i}, v_{T_i})$ for all the n_i subjects in cluster i is obtained based on the product of the conditional likelihood for each subject by:

$$L_i(\Phi|.) = \prod_{j=1}^{ni} \left\{ \lambda_{Sij}(S_{ij}|\omega_{ij}, u_i, v_{S_i})^{\delta_{ij}} \exp(-\int_0^{S_{ij}} \lambda_{Sij}(t|\omega_{ij}, u_i, v_{S_i}, Z_{ij})dt) \times \right.$$
$$\left. \lambda_{Tij}(T_{ij}|\omega_{ij}, u_i, v_{T_i})^{\delta^*_{ij}} \exp(-\int_0^{T_{ij}} \lambda_{Tij}(t|\omega_{ij}, u_i, v_{T_i}, Z_{ij})dt) \right\}.$$

The marginal contribution to the likelihood $L_i(\Phi)$ for each trial i is obtained by integrating the conditional likelihood upon the random effects:

$$L_i(\Phi) = \int_{u_i} \int_{v_{S_i}} \int_{v_{T_i}} \left\{ \prod_{j=1}^{ni} \int_{\omega_{ij}} L_i(\Phi|.) f(\omega_{ij}) f(v_{S_i}, v_{T_i}) f(u_i) d\omega_{ij} \right\} dv_{S_i} dv_{T_i} du_i,$$

where the density probability functions in cluster i are given by:

$$f(v_{S_i}, v_{T_i}) = \frac{1}{(2\pi)\sqrt{|\Sigma_{v_i}|}} \exp\left[-\frac{1}{2}(v_{S_i}, v_{T_i})\Sigma_{v_i}{}^{-1}(v_{S_i}, v_{T_i})'\right],$$

$$f(\omega_{ij},\theta)=\frac{1}{\sqrt{2\pi\theta}}\exp(-\frac{1}{2}\frac{\omega_{ij}^2}{\theta}),$$

and

$$f(u_i,\gamma)=\frac{1}{\sqrt{2\pi\gamma}}\exp(-\frac{1}{2}\frac{u_i^2}{\gamma})$$

respectively for correlated normal distributed random effects $(\upsilon_{S_i},\upsilon_{T_i})$, for individual normal distributed random effect ω_{ij} and trial random effects u_i. Therefore, using the previous expressions, we obtained the full marginal likelihood associated with the proposed joint surrogate model by using $l(\Phi)=\log\prod_{i=1}^{G}L_i(\Phi)$, where G is the total number of considered trials.

B DERIVATION OF KENDALL'S τ

We assume the joint conditional distributions of S_{ij} and T_{ij} under $\upsilon_{S_i}=\upsilon_{T_i}=0$ (i.e. $\sigma^2_{\upsilon_S}=\sigma^2_{\upsilon_T}=0$), after adjusting for both the trial and the treatment effects defined by:

$$\begin{cases}\lambda_{S,ij}(t|\omega_{ij},u_i,Z_{ijk})=\lambda_{0S}(t)\exp(\omega_{S_{ij}}+u_i+\sum_{k=1}^{p}\beta_{Sk}Z_{ijk})\\ \lambda_{T,ij}(t|\omega_{ij},u_i,Z_{ijk})=\lambda_{0T}(t)\exp(\zeta\omega_{T_{ij}}+\alpha u_i+\sum_{k=1}^{p}\beta_{Tk}Z_{ijk})\end{cases}\tag{B1}$$

where

$$\omega_{ij}\sim N(0,\theta),\quad u_i\sim N(0,\gamma),\tag{B2}$$

in which θ, γ, α and ζ are estimated using the proposed joint surrogate model. Kendal's τ could be rewrite by:

$$\tau=E\{sign[(S_{ij}-S_{i'j'})(T_{ij}-T_{i'j'})]\},$$

where $sign(x)=-1$, 0 and 1 for $x<0$, $x=0$ and $x>0$ respectively. Assume two randomly chosen individual j and j' from two randomly chosen cluster i and i' and assume that for bivariate survival data, the covariate information is the same in each cluster, i.e. $X_i=(x_{ij},x_{i'j'})=(x_i,x_{i'})=X$. An alternative formulation for continuous distributions is given by the Probability of matching pairs minus the probability of discordant pairs as:

$$\begin{aligned}\tau&=P[(S_{ij}-S_{i'j'})(T_{ij}-T_{i'j'})>0]-P[(S_{ij}-S_{i'j'})(T_{ij}-T_{i'j'})<0]\\&=2P[(S_{ij}-S_{i'j'})(T_{ij}-T_{i'j'})>0]-1\\&=2\{P[(S_{ij}>S_{i'j'})\cap(T_{ij}>T_{i'j'})]+P[(S_{i'j'}>S_{ij})\cap(T_{i'j'}>T_{ij})]\}-1.\end{aligned}\tag{B3}$$

One can show that (B3) could be rewrite as :

$$\tau=2\left\{\int_0^\infty\int_0^\infty S_{S_{i'j'},T_{i'j'}}(s,t)f_{S_{ij},T_{ij}}(s,t)dsdt+\int_0^\infty\int_0^\infty S_{S_{ij},T_{ij}}(s,t)f_{S_{i'j'},T_{i'j'}}(s,t)dsdt\right\}-1,$$

where $f(s,t)=\int_u f(s,t|u)f_u du$, and $S(s,t)=\int_u S(s,t|u)f_u du$ are the marginal density and survival functions, and f_u the density probability function of the random effect u with $u=(\omega_{S_{ij}},\omega_{T_{ij}},u_{S_i},u_{T_i})$. Otherwise, $f_{S_{ij},T_{ij}}(s,t)=\lambda_{S_{ij},T_{ij}}(s,t)S_{S_{ij},T_{ij}}(s,t)$ and conditional to random effects, T_{ij} and S_{ij} are independents $\forall i,j$.

Thereby, for the reduced joint surrogate model (B1), from the previous formulation ok Kendall's τ, one can show that

$$\begin{aligned}\tau=2\int_{u_i}\int_{\omega_{ij}}\int_{u_{i'}}\int_{\omega_{i'j'}}&\frac{\exp(\omega_{ij}+u_i+\zeta\omega_{ij}+\alpha u_i)+\exp(\omega_{i'j'}+u_{i'}+\zeta\omega_{i'j'}+\alpha u_{i'})}{(\exp(\omega_{i'j'}+u_{i'})+\exp(\omega_{ij}+u_i))(\exp(\zeta\omega_{i'j'}+\alpha u_{i'})+\exp(\zeta\omega_{ij}+\alpha u_i))}\\&\frac{1}{\sqrt{2\pi\theta}}\exp\left[-\frac{1}{2}\frac{\omega_{i'j'}^2}{\theta}\right]\frac{1}{\sqrt{2\pi\gamma}}\exp\left[-\frac{1}{2}\frac{u_{i'}^2}{\gamma}\right]d\omega_{i'j'}du_{i'}\\&\frac{1}{\sqrt{2\pi\theta}}\exp\left[-\frac{1}{2}\frac{\omega_{ij}^2}{\theta}\right]\frac{1}{\sqrt{2\pi\gamma}}\exp\left[-\frac{1}{2}\frac{u_i^2}{\gamma}\right]d\omega_{ij}du_i-1.\end{aligned}$$

C PROCEDURE OF NON-CONVERGENCE CASES

Special attention must be given to initializing model parameters, the choice of the number of spline knots, the smoothing parameters and the number of quadrature points to solve convergence issues. We firstly initialized parameters with the true values used during data generation; the number of quadrature point was set to 20 (12 for Full PGH); the number of knots to 6 and the smoothing parameters were based on cross-validation. When numerical or convergence problems were encountered, the model was fitted again using a combination of the following strategies: varying the number of quadrature point (20 to 32 or 12 to 15), dividing or multiplying k_1 or k_2 by 10 or 100 according to their preceding values, or using parameter vectors obtained during the last iteration (with a modification of the number of quadrature points and smoothing parameters). Using this strategy, we usually obtained the rejection rate described above. A sensitivity analysis was conducted without this strategy, and similar results were obtained on the converged samples with about a 23% rejection rate.

D ADDITIONAL SIMULATION RESULTS

TABLE D1 Estimates (Mean), mean of the standard errors (SE), empirical standard errors (SD) and percentage of coverage (CP). Estimation based on a pseudo-adaptive Gaussian-Hermite (PGH) integration with 12 quadrature points, for M=400 samples.

		N=600 subjects, G=30 trials			
Parameters	True	Mean	SD	SE	CP
		PGH			
θ	1.000	1.011	0.422	0.365	89
ζ	1.000	1.433	1.611	0.374	86
γ	0.800	0.760	0.260	0.272	90
α	1.000	1.096	0.364	0.192	91
$\sigma^2_{v_S}$	0.700	0.646	0.344	0.322	83
$\sigma^2_{v_T}$	0.700	0.738	0.503	0.364	84
$\sigma_{v_{ST}}$	0.630	0.585	0.301	0.290	86
β_S	-1.250	-1.256	0.247	0.222	93
β_T	-1.250	-1.313	0.359	0.236	91
R^2_{trial}	0.810	0.823	0.204	0.137	54
τ	0.378	0.368	0.042	-	88

TABLE D2 Estimates (Mean), mean of the standard errors (SE), empirical standard errors (SD) and percentage of coverage (CP). Estimation based on a combination of Monte-Carlo (MC) with non-adaptive Gaussian Hermite (GH) and with pseudo-adaptive Gaussian Hermite (PGH) integration‡, for M=500 samples.

		N=600 subjects, G=30 trials				N=600 subjects, G=10 trials			
Parameters	True	Mean	SD	SE	CP	Mean	SD	SE	CP
				MC-PGH†					
θ	3.500	3.648	0.771	0.67	92	3.674	0.699	0.703	94
ζ	1.500	1.615	1.173	0.313	81	1.527	0.660	0.331	84
γ	2.500	2.703	0.972	0.774	89	2.741	2.215	0.880	71
α	1.000	0.997	0.233	0.181	84	0.991	0.201	0.197	89
$\sigma^2_{v_S}$	0.700	0.744	0.523	0.498	88	0.762	0.686	0.477	82
$\sigma^2_{v_T}$	0.700	0.808	0.759	0.661	89	0.789	0.803	0.606	82
$\sigma_{v_{ST}}$	0.630	0.669	0.538	0.505	88	0.680	0.665	0.471	82
β_S	-1.250	-1.201	0.292	0.261	92	-1.228	0.363	0.322	92
β_T	-1.250	-1.194	0.393	0.350	92	-1.189	0.444	0.404	91
R^2_{trial}	0.810	0.817	0.239	0.404	72	0.816	0.247	0.286	67
τ	0.614	0.613	0.045	-	91	0.614	0.038	-	85
				MC-GH†					
θ	3.500	3.397	0.528	0.526	94	3.372	0.456	0.534	95
ζ	1.500	1.583	0.324	0.319	85	1.594	0.318	0.337	88
γ	2.500	2.666	0.971	0.753	88	2.617	1.644	0.834	68
α	1.000	1.041	0.192	0.187	87	1.048	0.197	0.203	90
$\sigma^2_{v_S}$	0.700	0.731	0.518	0.481	88	0.714	0.601	0.442	81
$\sigma^2_{v_T}$	0.700	0.850	0.753	0.704	91	0.820	0.791	0.648	84
$\sigma_{v_{ST}}$	0.630	0.690	0.550	0.515	90	0.672	0.613	0.470	84
β_S	-1.250	-1.183	0.274	0.255	92	-1.197	0.351	0.314	91
β_T	-1.250	-1.229	0.386	0.360	94	-1.227	0.450	0.419	92
R^2_{trial}	0.810	0.826	0.232	0.356	71	0.814	0.254	0.307	64
τ	0.614	0.614	0.025	-	91	0.612	0.032	-	82

†Convergence issues less than 1%, ‡ 300 samples for MC and 20 or 32 quadrature points for GH or PGH.

TABLE D3 Estimates (Mean), mean of the standard errors (SE), empirical standard errors (SD) and percentage of coverage (CP). Estimation based on a combination of Monte-Carlo (MC) with non-adaptive Gaussian Hermite (GH) integration[‡], for M=500 samples.

		G=30 trials				G=10 trials			
Parameters	True	Mean	SD	SE	CP	Mean	SD	SE	CP
Weak individual and trial level association[†], N=600 subjects									
θ	1.000	1.198	0.497	0.510	96	1.184	0.471	0.506	95
ζ	1.000	0.987	0.359	0.392	90	0.998	0.384	0.387	90
γ	0.800	0.939	0.417	0.333	92	0.965	0.684	0.343	74
α	1.000	0.988	0.194	0.212	95	0.988	0.227	0.220	93
$\sigma^2_{v_S}$	0.700	0.754	0.440	0.381	88	0.798	0.610	0.378	78
$\sigma^2_{v_T}$	0.700	0.709	0.447	0.365	86	0.847	0.800	0.418	77
$\sigma_{v_{ST}}$	0.420	0.438	0.330	0.264	88	0.495	0.552	0.256	73
β_S	-1.250	-1.247	0.274	0.243	91	-1.300	0.380	0.282	86
β_T	-1.250	-1.246	0.238	0.241	93	-1.293	0.390	0.281	85
R^2_{trial}	0.360	0.442	0.270	0.249	77	0.482	0.303	0.204	60
τ	0.378	0.382	0.038	-	92	0.380	0.055	-	79
High individual, weak trial level association[†], N=600 subjects									
θ	3.500	3.381	0.525	0.525	93	3.351	0.452	0.528	96
ζ	1.500	1.585	0.322	0.309	87	1.605	0.320	0.325	86
γ	2.500	2.651	0.986	0.745	88	2.695	1.804	0.825	64
α	1.000	1.040	0.196	0.184	87	1.052	0.198	0.201	91
$\sigma^2_{v_S}$	0.700	0.734	0.527	0.479	86	0.727	0.622	0.439	81
$\sigma^2_{v_T}$	0.700	0.827	0.746	0.664	87	0.831	0.850	0.616	81
$\sigma_{v_{ST}}$	0.420	0.486	0.509	0.465	86	0.463	0.542	0.409	79
β_S	-1.250	-1.188	0.278	0.255	92	-1.201	0.361	0.306	89
β_T	-1.250	-1.263	0.376	0.358	94	-1.259	0.463	0.413	92
R^2_{trial}	0.360	0.537	0.314	0.485	77	0.525	0.325	0.408	68
τ	0.614	0.613	0.025	-	90	0.612	0.033	-	78
Weak individual, high trial level association[†], N=1000 subjects									
θ	1.000	1.132	0.369	0.392	96	1.118	0.388	0.389	94
ζ	1.000	0.993	0.306	0.320	92	1.005	0.315	0.324	93
γ	0.800	0.899	0.310	0.270	92	0.994	1.028	0.275	65
α	1.000	0.995	0.159	0.166	93	1.004	0.176	0.176	95
$\sigma^2_{v_S}$	0.700	0.759	0.365	0.302	88	0.808	0.591	0.300	72
$\sigma^2_{v_T}$	0.700	0.701	0.319	0.287	88	0.763	0.594	0.304	69
$\sigma_{v_{ST}}$	0.630	0.649	0.292	0.238	87	0.669	0.491	0.230	66
β_S	-1.250	-1.241	0.233	0.210	91	-1.313	0.406	0.231	80
β_T	-1.250	-1.210	0.214	0.205	92	-1.287	0.371	0.227	80
R^2_{trial}	0.810	0.822	0.148	0.136	80	0.775	0.223	0.140	67
τ	0.378	0.383	0.030	-	91	0.384	0.055	-	71

[†]Convergence issues less than 0.5%, [‡] 300 samples for MC and 15 or 20 quadrature points for GH.

TABLE D4 Estimates (Mean), Bias, and Mean square errors (MSE): comparison between two-step (Clayton and Plackett copulas) and one-step (Poisson and proposed joint surrogate) approaches, for M=500 samples.

		600 subjects, 30 trials				600 subjects, 10 trials			
Parameters	True	Mean	Bias	MSE	n(%)‡	Mean	Bias	MSE	:n(%)‡
High individual and trial level association									
Clayton†					360(72)				49(10)
τ	0.614	0.457	0.157	0.034		0.499	0.115	0.023	
$R^2_{trial,adj}$	0.810	0.726	0.084	0.105		0.808	0.002	0.074	
Plackett†					357(71)				31(6)
τ	0.614	0.476	0.138	0.020		0.486	0.128	0.019	
$R^2_{trial,adj}$	0.810	0.799	0.011	0.094		0.804	0.006	0.077	
Poisson†					1(0)				8(2)
τ	0.614	0.467	0.147	0.022		0.468	0.146	0.022	
R^2_{trial}	0.810	0.853	-0.043	0.048		0.838	-0.028	0.060	
Joint Surrogate††					1(0)				4(1)
τ	0.614	0.613	0.001	0.002		0.614	0.000	0.001	
R^2_{trial}	0.810	0.817	-0.007	0.057		0.816	-0.006	0.061	
High individual weak trial level association									
Clayton†					361(72)				39(8)
τ	0.614	0.453	0.161	0.036		0.505	0.109	0.022	
$R^2_{trial,adj}$	0.360	0.655	-0.295	0.203		0.583	-0.223	0.195	
Plackett†					365(73)				57(11)
τ	0.614	0.473	0.141	0.022		0.498	0.116	0.017	
$R^2_{trial,adj}$	0.360	0.666	-0.306	0.224		0.589	-0.229	0.194	
Poisson†					13(3)				12(2)
τ	0.614	0.460	0.154	0.024		0.462	0.152	0.024	
R^2_{trial}	0.360	0.679	-0.319	0.191		0.644	-0.284	0.194	
Joint Surrogate††					1(0)				0(0)
τ	0.614	0.609	0.005	0.004		0.606	0.008	0.005	
R^2_{trial}	0.360	0.542	-0.182	0.131		0.520	-0.160	0.134	
Weak individual and trial level association									
Clayton†					340(68)				15(3)
τ	0.378	0.201	0.177	0.032		0.212	0.166	0.029	
$R^2_{trial,adj}$	0.360	0.501	-0.141	0.12		0.504	-0.144	0.137	
Plackett†					260(52)				13(3)
τ	0.378	0.213	0.165	0.030		0.231	0.147	0.026	
$R^2_{trial,adj}$	0.360	0.519	-0.159	0.135		0.497	-0.137	0.135	
Poisson†					1(0)				0(0)
τ	0.378	0.223	0.155	0.025		0.227	0.151	0.023	
R^2_{trial}	0.360	0.586	-0.226	0.127		0.593	-0.233	0.151	
Joint Surrogate††					0(0)				0(0)
τ	0.378	0.382	-0.004	0.001		0.375	0.003	0.004	
R^2_{trial}	0.360	0.442	-0.082	0.080		0.478	-0.118	0.106	

Estimation using: †Bobyqa algorithm, †† Marquardt algorithm; ‡convergence issues.

3.3 Discussion

Il est important de noter que dans le modèle développé dans ce chapitre, nous avons émis l'hypothèse d'homogénéité des effets du traitement entre les essais cliniques afin d'obtenir la formulation du τ de Kendall qui a été proposée. Cette hypothèse se ramène également à la définition du τ de Kendall chez des sujets non traités, lorsqu'on considère le modèle complet. En effet, sans cette hypothèse, l'expression du τ à partir du modèle complet est fonction des variables indicatrices de traitement. Ainsi, ne pas émettre l'hypothèse revient à considérer 3 formulations différentes du τ, suivant les bras de traitement des sujets ij et $i'j'$ considérée (Z_{ij1} et $Z_{i'j'1}$) pour sa définition. Cette hypothèse bien que peu réaliste en pratique permet de bien apprécier la relation au niveau individuelle entre les deux critères de jugement. En revanche, nous notons que dans ce modèle nous prenons en compte l'hétérogénéité au niveau essai associée aux fonctions de risque de base et l'hétérogénéité au niveau individuel. Par ailleurs, les études de simulations ont montré une robustesse dans l'estimation du taux de kendall, sauf lorsque les méta-analyses considérées avaient très peu d'essais (10), pour lesquelles on observait des taux de couverture $< 80\%$).

3.4 Annexes supplémentaires

On peut écrire la log-vraisemblance marginale suivant la méthode d'intégration comme suit:

3.4.1 Formulation de la log-vraisemblance marginale avec intégration par Monte-Carlo (MC)

$$
\begin{aligned}
l(\Phi) \;=\; & \sum_{i=1}^{G}\Bigg\{\log\Bigg\{\int_{u_i}\int_{v_{S_i}}\int_{v_{T_i}}\Bigg\{\prod_{j=1}^{ni}\int_{\omega_{ij}}\exp\Bigg[\delta_{ij}\Bigg(\log\lambda_{0S}(T_{ij})+\sum_{k=1}^{p}\beta_k^S Z_{ijk}\Bigg)\\
& +\delta_{ij}^*\Bigg(\log\lambda_{0T}(D_{ij})+\sum_{k=1}^{p}\beta_k^T Z_{ijk}\Bigg)+u_i(\delta_{ij}+\delta_{ij}^*\alpha)+(v_{S_i}\delta_{ij}+v_{T_i}\delta_{ij}^*)Z_{ij1}\\
& +\omega_{ij}(\delta_{ij}+\delta_{ij}^*\zeta)-\Lambda_{0S}(T_{ij})\exp(\sum_{k=1}^{p}\beta_k^S Z_{ijk})\exp(\omega_{ij}+u_i+v_{S_i}Z_{ij1})\\
& -\Lambda_{0T}(D_{ij})\exp(\sum_{k=1}^{p}\beta_k^T Z_{ijk})\exp(\zeta\omega_{ij}+\alpha u_i+v_{T_i}Z_{ij1})\Bigg]\times\frac{1}{\sqrt{2\pi\theta}}\exp(-\frac{1}{2}\frac{\omega_{ij}^2}{\theta})d\omega_{i_1}\Bigg\}\\
& \times\frac{1}{(2\pi)\sqrt{|\Sigma_{v_i}|}}\exp\left[-\frac{1}{2}(v_{S_i},v_{T_i}){\Sigma_{v_i}}^{-1}(v_{S_i},v_{T_i})'\right]\times\frac{1}{\sqrt{2\pi\gamma}}\exp(-\frac{1}{2}\frac{u_i^2}{\gamma})dv_{T_i}dv_{S_i}du_i\Bigg\}\Bigg\}
\end{aligned}
$$

3.4.2 Formulation de la log-vraisemblance marginale avec intégration par MC au niveau essai et quadrature de GH au niveau individuel

$$
\begin{aligned}
l(\Phi) \;=\; & \sum_{i=1}^{G}\Bigg\{\log\Bigg\{\int_{u_i}\int_{v_{S_i}}\int_{v_{T_i}}\Bigg\{\prod_{j=1}^{ni}\int_{\omega_{ij}}\exp\Bigg[\delta_{ij}\Bigg(\log\lambda_{0S}(T_{ij})+\sum_{k=1}^{p}\beta_k^S Z_{ijk}\Bigg)\\
& +\delta_{ij}^{*}\Bigg(\log\lambda_{0T}(D_{ij})+\sum_{k=1}^{p}\beta_k^T Z_{ijk}\Bigg)+u_i(\delta_{ij}+\delta_{ij}^{*}\alpha)+(v_{S_i}\delta_{ij}+v_{T_i}\delta_{ij}^{*})Z_{ij1}\\
& +\omega_{ij}(\delta_{ij}+\delta_{ij}^{*}\zeta)-\Lambda_{0S}(T_{ij})\exp(\sum_{k=1}^{p}\beta_k^S Z_{ijk})\exp(\omega_{ij}+u_i+v_{S_i}Z_{ij1})\\
& -\Lambda_{0T}(D_{ij})\exp(\sum_{k=1}^{p}\beta_k^T Z_{ijk})\exp(\zeta\omega_{ij}+\alpha u_i+v_{T_i}Z_{ij1})-\frac{1}{2}\log(2\pi\theta)-\frac{\omega_{ij}^2}{2\theta}\Bigg]d\omega_{i_1}\Bigg\}\\
& \times\frac{1}{(2\pi)\sqrt{|\Sigma_{v_i}|}}\exp\left[-\frac{1}{2}(v_{S_i},v_{T_i})\Sigma_{v_i}{}^{-1}(v_{S_i},v_{T_i})'\right]\times\frac{1}{\sqrt{2\pi\gamma}}\exp(-\frac{1}{2}\frac{u_i^2}{\gamma})dv_{T_i}dv_{S_i}du_i\Bigg\}\Bigg\}
\end{aligned}
$$

Chapter 4

Développement d'un package R pour la validation en une étape des critères de substitution à l'aide d'un modèle conjoint à fragilités

---

4.1 Article

Dans le Chapitre 3, nous avons proposé une nouvelle approche méthodologique pour la validation en une étape des critères de substitution. Cette méthode qui s'appuyait sur un nouveau modèle conjoint à fragilités était assez robuste et a permis de réduire les problèmes de convergence et d'estimation souvent rencontrés dans l'approche standard. Dans ce travail, nous avons fixé comme objectif, de vulgariser la méthode en la rendant accessible aux cliniciens et à la communauté scientifique à travers un package `R` et de proposer un tutoriel approprié pour son utilisation.

Plus spécifiquement, nous présentons dans ce chapitre la fonction `R jointSurroPenal()` qui permet d'estimer les paramètres du modèle développé au Chapitre 3. Les arguments de cette fonction ainsi que les objets disponibles en sortie ont été documentés en profondeur et peuvent être consultés à partir de l'aide sur cette fonction. En complément à l'estimation, nous avons proposé de nouveaux outils applicables à l'objet `R` issu de la fonction `jointSurroPenal()`. Ces outils permettent de présenter un résumé des sorties; de faire de la prédiction des effets du traitement sur le critère de jugement principal à partir des effets du traitement observés sur le critère de substitution dans de nouveaux essais; d'évaluer la précision des prédictions à partir d'une variante de la validation croisée, le `loocv` (leave-one-out cross-validation). Nous avons

également implémenté l'effet minimum d'un critère de substitution (ou surrogate threshold effect, STE) qui est une quantité proposée par Burzykowski et al. (2005) pour capter l'effet minimum du traitement observable sur le critère de substitution, pour prédire un effet significatif du traitement sur le critère de jugement principal. Nous nous sommes appuyé sur le STE et le R^2_{trial} pour orienter les utilisateurs sur la validité du critère de substitution, en suivant la classification proposée par l'agence Allemande d'évaluation des technologies de la santé, Institute for Quality and Efficiency in Health Care (2011). Nous avons par la suite proposé d'autres fonctions pour générer les données et conduire les études de simulation. Chaque nouvelle fonction était accompagnée d'une documentation conséquente.

Tous nos programmes ont été inclus dans `frailtypack`, qui est un package `R` destiné à l'estimation des paramètres d'une variété de modèles à fragilités, contenant un ou plusieurs effets aléatoires corrélés, ou des fragilités partagées (Król et al. 2017). Par exemple, tous les modèles à fragilités présentés dans les sections (2.5) et (2.6) y sont implémentés. Ces développements majeurs ont fait passer `frailtypack` de la version `2.13.2` à la version `3.0.1`, qui a été publiée sur le `CRAN` (Comprehensive `R` Archive Network) en `Novembre 2018`. Par ailleurs, afin d'accélérer les calculs, tous nos programmes sont développés en `Fortran 90`, et parallélisés suivant l'interface de programmation OpenPM (Open Multi-Processing). Par conséquent, `R` nous sert seulement d'interface pour l'appel des fonctions `Fortran` et la présentation des résultats.

Afin de faciliter l'utilisation du package, nous avons écrit un article scientifique qui sert de tutoriel pour la mise en œuvre du modèle et des fonctions développées. Nous discutons dans cet article le choix des couples arguments/valeurs, la gestion des problèmes de convergence et l'interprétation des sorties des fonctions.

Ce travail est en révision dans *Plos One* (Casimir L. Sofeu et Virginie Rondeau, 2019). Dans le chapitre 5, nous proposons un nouveau type de modèle conjoint à fragilités et à copules, qui a également été inclus dans `frailtypack`.

How to use `frailtypack` for validating failure-time surrogate endpoints using individual patient data from meta-analyses of randomized controlled trials

Casimir Ledoux SOFEU*, Virginie Rondeau

INSERM U1219 - Biostatistics, Bordeaux, France
Université de Bordeaux, ISPED, Bordeaux, France

* casimir.sofeu@u-bordeaux.fr, scl.ledoux@gmail.com

Abstract

Background and Objective: The use of valid surrogate endpoints can accelerate the development of phase III trials. Numerous validation methods have been proposed with the most popular used in a context of meta-analyses, based on a two-step analysis strategy. For two failure time endpoints, two association measures are usually considered, Kendall's τ at individual level and adjusted R2 (adjR^2_{trial}) at trial level. However, adjR^2_{trial} is not always available mainly due to model estimation constraints. More recently, we proposed a one-step validation method based on a joint frailty model, with the aim of reducing estimation issues and estimation bias on the surrogacy evaluation criteria. The model was quite robust with satisfactory results obtained in simulation studies. This study seeks to popularize this new surrogate endpoints validation approach by making the method available in a user-friendly R package. *Methods:* We provide numerous tools in the `frailtypack` `R` package, including more flexible functions, for the validation of candidate surrogate endpoints using data from multiple randomized clinical trials. *Results:* We implemented the surrogate threshold effect which is used in combination with R^2_{trial} to make decisions concerning the validity of the surrogate endpoints. It is also possible thanks to `frailtypack` to predict the treatment effect on the true endpoint in a new trial using the treatment effect observed on the surrogate endpoint. The leave-one-out cross-validation is available for assessing the accuracy of the prediction using the joint surrogate model. Other tools include data generation, simulation study and graphic representations. We illustrate the use of the new functions with both real data and simulated data. *Conclusion:* This article proposes new attractive and well developed tools for validating failure time surrogate endpoints.

Introduction

The choice of endpoint for assessing the efficacy of a new treatment is a key step in setting up clinical trials. The use of the true endpoint increases the cost and duration of trials, and usually induces an alteration of the treatment effects over time [1, 2]. For example, in oncology, overall survival is a common clinical endpoint used during phase 3 trials to evaluate the clinical benefit of new treatments. However, its use requires a sufficiently long follow-up time and a sufficiently high sample size to show a significant difference in the treatment effect. To overcome this problem, there has been a lot of interest over the last three decades in the use of alternative criteria or surrogate

endpoints to reduce the cost and shorten the duration of phase 3 trials [1–4]. A good surrogate endpoint should predict the effect of treatment on the primary endpoint [3].

Prentice (1989) [5] enumerated four criteria to be fulfilled by a putative surrogate endpoint. The fourth criterion, often called Prentice's criterion, stipulates that a surrogate endpoint must capture the full treatment effect upon the true endpoint. The validation of Prentice's criterion based on a clinical trial was quite difficult, mainly due to a lack of power and the difficulty to verify an assumption related to the relation between the treatment effects upon the true and the surrogate endpoints. Therefore, to verify this assumption and obtain a consistent sample size, Buyse *et al.* (2000) [6] like other authors [7] suggested basing validation on the meta-analytic (or multicenter) data. An important point when dealing with meta-analytic data is to take heterogeneity between trials into account, for the purpose of prediction outside the scope of the trial. Thus, a validated surrogate endpoint from meta-analytic data can be used to predict the treatment effect upon the true endpoint in any trial.

In the meta-analysis framework, when both the surrogate and the true endpoints are failure times, the current consensus is to base validation on the two-stage analysis strategy proposed by Burzykowski *et al.* [8]. In the first stage, the association between the surrogate and true endpoints is evaluated using a bivariate copula model after taken the trial specific treatment effects into account. In the second stage, the prediction of the treatment effect on the true endpoint based on the observed treatment effect on the surrogate endpoint is assessed using the adjusted coefficient of determination (adjR^2_{trial}). adjR^2_{trial} is obtained from the regression model on the estimates of the trial-specific treatment effects on both the surrogate and the true endpoints, after adjusting on the estimation errors obtained in the first-stage model. The programs that implement this method are available in the `R` package `surrosurv` [9] and the `SAS` macro `%COPULA` [10]. However, the practical use of the two-stage copula model is often difficult, mainly due to convergence issues or model estimation with the adjustment on the estimation errors [11–13]. This drawback led to the development since Burzykowski *et al.* [8] of alternative approaches [11, 13–17].

Most of the novel methods, except that of Sofeu *et al.* [17] and Rotolo *et al.* [13], are based on a two-stage validation strategy. Alonso and Molenberghs [14] proposed an information theory approach, with a new definition and quantification of surrogacy at the individual level and the trial level. The drawback of this method was the difficulty to provide a hard cut-off value in the information-theoretic measure, to discriminate between good and bad surrogates. Buyse et *et al.* [15] suggested a two-stage validation approach in which individual-level surrogacy was evaluated through the association between the trial-specific Kaplan-Meier estimates of the true endpoint versus Kaplan-Meier estimates of the surrogate endpoint at a fixed time point. It is also possible to base validation at the individual level on a bivariate copula model. In the trial-level evaluation, a weighted linear regression on the treatment effects on the surrogate and true endpoints was fitted and the coefficient of determination (R^2) was used to quantify the proportion of variance explained by the regressions. The available programs also make it possible to account for variability between trials using a robust sandwich estimator of Lin and Wei [18].

For the approaches described in the previous paragraph, the `R` package `surrogate` [19] , the `SAS` macros `%TWOSTAGECOX and %TWOSTAGEKM`, and the `SAS` programs available in Alonso *et al.* [10] were provided to carry out the evaluation exercise. Rotolo *et al.* [13] proposed a one-step validation approach based on auxiliary mixed Poisson models, which employs a bivariate survival model with an individual random effect shared between the two endpoints and correlated treatment-by-trial interactions. Simulation results described by the authors showed estimation biases on the surrogacy assessment measures , especially in the event of a high association and

when heterogeneity of baseline risk is taken into account. The associated program was implemented in the R package surrosurv [9]. Renfro *et al.* [11] suggested estimating the second-stage model in a Bayesian framework and the estimate of the adjusted R^2_{trial} was then based on the posterior distribution of the parameters of the adjusted model. The corresponding trial-level surrogacy can be evaluated by adapting the WinBUGS and R programs described in Bujkiewicz *et al.* [20]. This approach emphasizes a decrease in estimation performance of the adjusted R^2_{trial} when the data characteristics are close to reality (for example, low trial size or number of trial).

More recently, we proposed a one-step validation approach based on a joint frailty model [17]to reduce convergence issues and estimation biases on the surrogacy evaluation criteria. In this novel method, we used a flexible form of the baseline hazard functions using splines to obtain smooth risk functions, which represent incidence in epidemiology. Several integration strategies were considered to compute integrals over the random effect, present in the marginal log-likelihood. The proposed joint surrogate model showed satisfactory results compared to the existing two-step copula and one-step Poisson approaches.

We aim in this paper to popularize this new surrogate endpoints validation approach by making the method available in a user-friendly R package (frailtypack). We have developed a prediction tool for the treatment effect on true endpoints based on the observed treatment effect on surrogate endpoints. Interpretation of R^2_{trial} and decision-making about the validity of the candidate surrogate endpoint are possible thanks to the classification suggested by the Institute for Quality and Efficiency in Health Care [21], and surrogate threshold effect (STE) introduced by Burzykowski and Buyse [22]. Other tools are for displaying the basic risks and survival functions, for model assessment, and for data generation based on the joint surrogate model. Another attractive goal of this article is to provide a tool to perform simulation studies.

frailtypack is an R package that fits a variety of frailty models containing one or more random effects, or shared frailty. It includes a shared frailty model, a joint frailty model for recurrent events and terminal event, others forms of advanced joint frailty models [23], and now a joint frailty model for evaluating surrogate endpoints in meta-analyses of randomized controlled trials with failure-time endpoints. In this paper we focus on a particular subset of features applicable for evaluating surrogate endpoints.

The rest of this paper is organized as follows. In the next section, we summarize the joint surrogate model with the estimation methods and the surrogacy evaluation criteria. We end it with the definition of STE. In the third section, we introduce the functions developed in the R-package frailtypack to estimate the parameters of the joint surrogate model, as well as the new functions related to the surrogacy evaluation. In the fourth section, we illustrate the new functions using generated data and individual patient data from the Ovarian Cancer Meta-Analysis Project [24]. Finally, we present a concluding discussion.

Methodology

In this section, we present the one-step joint surrogate model for evaluating a candidate surrogate endpoint [17]. The model estimation and the surrogacy evaluation criteria are also discussed here.

Model and estimation

Joint surrogate model definition

Let us consider data from a meta-analysis (or a multi-center study); let S_{ij} and T_{ij} be two time-to-event endpoints associated respectively with the surrogate endpoint and the

true endpoint such that $S_{ij} < T_{ij}$ or $S_{ij} = T_{ij}$ in the event of right censoring. We denote Z_{ij1} the treatment indicator. S_{ij} can be the progression-free survival time (defined as the time from randomization to clinical progression of the disease or death) in patients treated for cancer and T_{ij} the overall survival (defined as the time from randomization to death from any cause). For the j^{th} subject ($j = 1, \ldots, n_i$) of the i^{th} trial ($i = 1, \ldots, G$), the joint surrogate model is defined as follows [17]:

$$\begin{cases} \lambda_{S,ij}(t|\omega_{ij}, u_i, v_{S_i}, Z_{ij1}) & = \lambda_{0S}(t)\exp(\omega_{ij} + u_i + v_{S_i}Z_{ij1} + \beta_S Z_{ij1}) \\ \lambda_{T,ij}(t|\omega_{ij}, u_i, v_{T_i}, Z_{ij1}) & = \lambda_{0T}(t)\exp(\zeta\omega_{ij} + \alpha u_i + v_{T_i}Z_{ij1} + \beta_T Z_{ij1}) \end{cases} \quad (1)$$

where,

$$\omega_{ij} \sim N(0, \theta), u_i \sim N(0, \gamma), \omega_{ij} \perp u_i, u_i \perp v_{S_i}, u_i \perp v_{T_i}$$

and

$$\begin{pmatrix} v_{S_i} \\ v_{T_i} \end{pmatrix} \sim MVN\left(\mathbf{0}, \Sigma_v\right), with\ \Sigma_v = \begin{pmatrix} \sigma^2_{v_S} & \sigma_{v_{ST}} \\ \sigma_{v_{ST}} & \sigma^2_{v_T} \end{pmatrix}$$

In this model, $\lambda_{0S}(t)$ is the baseline hazard function associated with the surrogate endpoint and β_S the fixed treatment effect (or log-hazard ratio); $\lambda_{0T}(t)$ is the baseline hazard function associated with the true endpoint and β_T the fixed treatment effect. ω_{ij} is a shared individual-level frailty that serve to take into account the heterogeneity in the data at the individual level due to unobserved covariates; u_i is a shared frailty effect associated with the baseline hazard function that serve to take into account the heterogeneity between trials of the baseline hazard function, associated with the fact that we have several trials in this meta-analytical design. Coefficients ζ and α distinguish both individual and trial-level heterogeneities between the surrogate and the true endpoint. v_{S_i} and v_{T_i} are two correlated random effects treatment-by-trial interactions.

Estimation

Marginal log-likelihood Let δ_{ij} and δ^*_{ij} be the progression and the death indicators. Sofeu *et al.* [17] showed that the marginal log-likelihood from model (1) includes two integration levels and is defined as follows:

$$l(\Phi) = \log\left\{\prod_{i=1}^{G}\int_U\left[\prod_{j=1}^{ni}\int_{\omega_{ij}} \lambda_{Sij}^{\delta_{ij}} \cdot S(S_{ij}) \cdot \lambda_{Tij}^{\delta^*_{ij}} \cdot S(T_{ij}) f(\omega_{ij}) d\omega_{ij}\right] f(v_{S_i}, v_{T_i}) f(u_i) dU\right\} \quad (2)$$

where $\Phi = (\hat{\sigma}^2_{v_S}, \hat{\sigma}^2_{v_T}, \hat{\sigma}_{v_{ST}}, \hat{\theta}, \hat{\gamma}, \hat{\lambda}_{0T}(.), \hat{\lambda}_{0S}(.), \hat{\beta}_S, \hat{\beta}_T)$ is the vector of the model parameters and $U = (u_i, v_{S_i}, v_{T_i})$ is the vector of trial random effects. $\hat{\lambda}_{0S}(.)$ and $\hat{\lambda}_{0T}(.)$ are estimates for the baseline hazard functions associated with the surrogate endpoint and the true endpoint.

Parameters estimation The model parameters Φ were estimated by a semi-parametric approach using the maximization of the penalized likelihood. We used the robust Marquardt algorithm [25], which is a mixture between the newton-Raphson and the steepest descent algorithm. For more details on the penalized likelihood, see the S1A Appendix in S1 Appendix or [26]. In order to estimate the integrals present in (2), different numerical integration strategies were considered, including a mixture of the Monte-Carlo integration with the Pseudo-adaptive or the classical Gauss-Hermite quadrature.

Surrogacy evaluation criteria and interpretation

We have already proposed new definitions of Kendall's τ and coefficient of determination as individual-level and trial-level association measures to evaluate a candidate surrogate endpoint [17]. We recall in the S1B and S1C Appendix in S1 Appendix the formulation of these association measures.

Prediction and surrogate threshold effect (STE)

Gail *et al.* [27] underlined some issues in using R^2_{trial} for assessing a candidate surrogate endpoint. The first problem is the difficulty in interpreting R^2_{trial}. For perfect prediction of the treatment effect on the true endpoints, R^2_{trial} must be equal to 1. However, such a situation is impossible in practice. Therefore, for $R^2_{trial} \neq 1$, it is not clear what threshold would be sufficient for a valid surrogate endpoint. Another problem raised by Gail *et al.* [27] is that, unless $R^2_{trial} = 1$, the variance of the prediction of the treatment effect on the true endpoint in a new trial cannot be reduced to 0, even in the absence of any estimation error in the trial. Furthermore, if this effect is estimated directly from data on the true endpoint, this estimation error can theoretically be made arbitrarily close to 0 by increasing the trial's sample size. To address these issues, Burzykowski and Buyse [22] proposed a new concept, the surrogate threshold effect. One of the most interesting features of STE is its natural interpretation from a clinical point of view. STE represents the minimum treatment effect on the surrogate necessary to predict a non-zero (significant) effect on the true endpoint. We show in S1D Appendix in S1 Appendix that STE, based on model (1), can be obtained by solving one of the following quadratic equations:

$$E(\beta_T + v_{T0}|\beta_{S_0}, \vartheta) - z_{1-(\gamma/2)}\sqrt{Var(\beta_T + v_{T0}|\beta_{S_0}, \vartheta)} = 0 \quad (3)$$

for the lower prediction limit function of the treatment effect on the true endpoint based on the observed treatment effect on the surrogate endpoint, or

$$E(\beta_T + v_{T0}|\beta_{S_0}, \vartheta) + z_{1-(\gamma/2)}\sqrt{Var(\beta_T + v_{T0}|\beta_{S_0}, \vartheta)} = 0, \quad (4)$$

for the upper prediction limit function. Elements in equations (3)-(4) are defined in S1D Appendix in S1 Appendix.

Readers can refer to S1E Appendix in S1 Appendix for the interpretation of STE, in combination with R^2_{trial} and decision-making as suggested by the German Institute for Quality and Efficiency in Health Care (IQWiG) [21]

Available functions in the `frailtypack` `R` package for surrogacy evaluation

In this section, we introduce the new `R` functions, used to estimate model (1). Functions for data generation and simulation studies are also described.

Estimation of joint surrogate model and surrogacy evaluation

The `jointSurroPenal()` function

Model (1) can be fitted using the `jointSurroPenal()` function defined as follows:

```
jointSurroPenal(data, maxit = 40, indicator.zeta = 1, indicator.alpha = 1,
  frail.base = 1, n.knots = 6, LIMlogl = 0.001, LIMparam = 0.001,
  LIMderiv = 0.001, nb.mc = 300, nb.gh = 32, nb.gh2 = 20, adaptatif = 0,
```

```
  int.method = 2, nb.iterPGH = 5, nb.MC.kendall = 10000,
  nboot.kendall = 1000, true.init.val = 0, theta.init = 1,
 sigma.ss.init = 0.5, scale = 1, sigma.tt.init = 0.5, sigma.st.init = 0.48,
  gamma.init = 0.5, alpha.init = 1, zeta.init = 1, betas.init = 0.5,
  betat.init = 0.5, random.generator = 1, kappa.use = 4, random = 0,
  seed = 0, random.nb.sim = 0, init.kappa = NULL, nb.decimal = 4,
  print.times = TRUE, print.iter = FALSE)
```

The mandatory argument of this function is `data`, the dataset to use for the estimations. Argument `data` refers to a dataframe including at least 7 variables: `patienID`, `trialID`, `timeS`, `statusS`, `timeT`, `status` and `trt`. The description of these variables, like other arguments of the function, can be found in S2A Appendix in S2 Appendix, or via the R command help(jointSurroPenal). The rest of the arguments can be set to their default values. In addition, details on the required arguments/values are given in the illustration section.

The `jointSurroPenal` object

The function `jointSurroPenal()` returns an object of class `'jointSurroPenal'`, if the joint surrogate model has been estimated. We describe in S2A Appendix in S2 Appendix some of the relevant returned values, as well as the functions which can be applied to this object. A full description can be found by displaying the help on the function `jointSurroPenal()`.

Data Generation using the `R` function `jointSurrSimul()`

For data generation purposes, we implemented the algorithm described in Sofeu *et al.* [17] in the `R` function `jointSurrSimul()`. The generation procedure is based on model (1). A variant of this algorithm is to base generation on a model that includes just a shared frailty term at the individual level as described by Rondeau *et al.* [28]. This function is defined as follows:

```
jointSurrSimul(n.obs = 600, n.trial = 30, cens.adm = 549.24, alpha = 1.5,
  theta = 3.5, gamma = 2.5, zeta = 1, sigma.s = 0.7, sigma.t = 0.7,
  rsqrt = 0.8, betas = -1.25, betat = -1.25, frailt.base = 1,
  lambda.S = 1.8, nu.S = 0.0045,lambda.T = 3, nu.T = 0.0025, ver = 1,
  typeOf = 1, equi.subj.trial = 1, equi.subj.trt = 1,
  prop.subj.trial = NULL, full.data = 0, prop.subj.trt = NULL,
  random.generator = 1, random = 0, random.nb.sim = 0, seed = 0,
  nb.reject.data = 0)
```

Arguments of the `jointSurrSimul()` function are accessible using the `R` command `help(jointSurrSimul)`. An exhaustive description is presented in S2B Appendix in S2 Appendix.

Simulation studies based on the joint surrogate model

It is possible to perform simulation studies based on model (1), using the function `jointSurroPenalSimul()` defines as follows:

```
jointSurroPenalSimul(nb.dataset = 1, nbSubSimul = 1000, ntrialSimul = 30,
  equi.subj.trial = 1, prop.subj.trial = NULL, equi.subj.trt = 1,
  prop.subj.trt = NULL, theta2 = 3.5, zeta = 1, gamma.ui = 2.5,
  alpha.ui = 1, sigma.s = 0.7, sigma.t = 0.7, R2 = 0.81, betas = -1.25,
```

```
  betat = -1.25, lambdas = 1.8, nus = 0.0045, lambdat = 3, nut = 0.0025,
  time.cens = 549, indicator.zeta = 1,indicator.alpha = 1, frail.base = 1,
  init.kappa = NULL, n.knots = 6, maxit=40, LIMparam = 0.001,
  LIMlogl = 0.001, LIMderiv = 0.001, int.method = 2, adaptatif = 0,
  nb.iterPGH = 5, nb.mc = 300, nb.gh = 32, nb.gh2 = 20,
  nb.MC.kendall = 10000, nboot.kendall = 1000, true.init.val = 0,
  theta.init = 1, zeta.init = 1, gamma.init = 0.5, alpha.init = 1,
  sigma.ss.init = 0.5, sigma.tt.init = 0.5, sigma.st.init = 0.48,
  betas.init = 0.5, betat.init = 0.5, kappa.use = 4,
  random.generator = 1, random = 0, random.nb.sim = 0, seed = 0,
  nb.decimal = 4, print.times = TRUE, print.iter = FALSE)
```

Most of the arguments in this function are mandatory for the user, taking into account the simulation design. S2B Appendix in S2 Appendix describes all the arguments, as well as the elements of the `'jointSurroPenalSimul'` object.

Kendall's τ estimation using the function `jointSurroTKendall`

The function `jointSurroTKendall()` is used to estimate Kendall's τ described in S1B Appendix in S1 Appendix, based on the estimates from the model (1). It is possible to perform the numerical integration with the Monte-Carlo or the Gauss-Hermite quadrature method. The `jointSurroTKendall()` function is defined as shown below, with arguments described in S2D Appendix in S2 Appendix. This function returns the estimated value of Kendall's τ

```
jointSurroTKendall(object = NULL, theta, gamma, alpha = 1, zeta = 1,
  int.method = 0, sigma.v = matrix(rep(0,4),2,2), nb.gh = 32,
  nb.MC.kendall = 10000, random.generator = 1, random.nb.sim = 0,
  random = 0, seed = 0, ui = 1)
```

Illustrations

Computational details and package installation

Estimations in the proposed functions are based on `Fortran` programs, with parallel computing using `OpenMP`, to speed up calculations. Thus, we used `R` as an interface between the user and the `Fortran` compiler. The stable version of `frailtypack` is available on the Comprehensive `R` Archive Network (CRAN) [29]. Furthermore, the ongoing version can be found on GitHub at `https://github.com/socale/frailtypack`. A list of other models implemented in `frailtypack` [23] can be found in Fig in S1 Fig. The results in this paper were obtained using `R` version `3.5.2` and the `frailtypack` package version 3.0.3 , using a processor `Intel(R) Xeon(R) CPU E5-2690 v2 @ 3.00GHz` including 40 cores and a Read Only Memory (RAM) of 378 Gb. A standard laptop and a desktop PC under recent versions of `R` can be used to fit the model. The results will be the same, but with longer computing time. For example, using a standard desktop PC in the application, the fit took around 1 hours compared to 9 min with a server including 40 cores and a RAM of 378 Go.

The `frailtypack` package can be installed in any R session using the `install.packages` command as follows:

```
install.packages("frailtypack", dependencies = T, type ="source",
repos = "https://cloud.r-project.org")
```

Installation via `GitHub` is possible thanks to the `devtools` package. All dependencies required by `frailtypack` must be installed first. The installation commands are:

```
install.packages(c("survC1","doBy","statmod"), repos = "https://cloud.r
-project.org")
devtools::install_github("socale/frailtypack", ref = "surrogacy_submetted_
3-0-3")
```

Finally, `frailtypack` must be loaded using the command:

```
library(frailtypack)
```

Data source

We illustrate the use of the developed functions with the individual patient data of the Ovarian Cancer Meta-Analysis Project [24] and a generated dataset based on model (1). We also describe the simulation studies at the end of this section.

Description of `dataOvarian` dataset

The `dataOvarian` dataset combines data that were collected in four double-blind randomized clinical trials in advanced ovarian cancer. In the first two trials of this study, data were available on the centers in which patients were treated, and each of the the two trials were considered as a homogeneous group according to the investigators. Finally, the statistical unit in the first two trials was center and it was trial for the last two trials. Therefore, a total of 50 units were available for surrogacy evaluation. The objective in these studies was to examine the efficacy of cyclophosphamide plus cisplatin (CP) versus cyclophosphamide plus adriamycin plus cisplatin (CAP) to treat advanced ovarian cancer. The candidate surrogate endpoint **S** was progression-free survival time (PFS), defined as the time (in years) from randomization to clinical progression of the disease or death. The true endpoint **T** was survival time, defined as the time (in years) from randomization to death from any cause. The dataset includes 1192 subjects with 82% of PFS-related events at a median survival time of 78.7 days [Interquartile range (IQR): 36.6 - 202.5], and 79.8% of deaths at a median survival time of 111.4 days [IQR: 56.0 - 275.9]. Data can be loaded as follows:

```
data("dataOvarian", package = "frailtypack")
```

By displaying the structure of this dataset, we can find the same structure as in the function `jointSurroPenal()`, with 7 variables. The column `trialID` here refers to the analysis unit

```
str(dataOvarian)

 'data.frame': 1192 obs. of  7 variables:
  $ patienID: int  1 2 3 4 5 6 7 8 9 10 ...
  $ trialID : num  2 2 2 2 2 2 2 2 2 2 ...
  $ trt     : int  0 0 0 1 0 1 0 0 1 1 ...
  $ timeS   : num  0.1052 0.8952 0.079 1.7393 0.0913 ...
  $ statusS : int  1 1 1 0 1 1 1 1 1 1 ...
  $ timeT   : num  0.186 1.409 0.126 1.739 0.127 ...
  $ statusT : int  1 1 1 0 1 1 1 1 1 1 ...
```

Generated dataset

In the example below, we generate a meta-analysis including 600 subjects in 30 trials. Arguments α, θ, ζ and γ are fixed to obtain a Kendall's τ of 0.61, which is obtained using the `jointSurroTKendall()` function as follows:

```
jointSurroTKendall(theta = 3.5, gamma = 2.5, alpha = 1.5, zeta = 1)
```

```
 [1] 0.6062975
```

Otherwise, the trial level surrogacy, R^2_{trial} is fixed to 0.8. This could correspond to simulation design including high trial level and high individual level surrogacy. The treatment effects β_S and β_T are set to -1.25 to consider protective effects both on the surrogate endpoint and the true endpoint. The code below is used to generate the dataset using the `jointSurrSimul()` function introduced previously, and display the head.

```
data.sim <- jointSurrSimul(n.obs = 600, n.trial = 30, alpha = 1.5,
    theta = 3.5, gamma = 2.5, zeta = 1, sigma.s = 0.7, sigma.t = 0.7,
    rsqrt = 0.8, betas = -1.25, betat = -1.25, random.generator = 1,
    seed = 0, nb.reject.data = 0)

head(data.sim)
```

```
  patienID trialID trt      timeS statusS     timeT statusT
1        1       1   0   8.243721       1  38.41068       1
2        2       1   1 446.169009       0 446.16901       1
3        3       1   1 110.418853       0 110.41885       1
4        4       1   1  70.262075       0  70.26207       1
5        5       1   1 382.973632       1 549.24000       0
6        6       1   0  61.148254       1 230.24486       1
```

Surrogacy evaluation

In this section, we use the dataset previously described to illustrate the evaluation of the surrogate endpoints based on the one-step joint surrogate model (1). Different arguments of the associated functions will be explored as the returned values.

Model estimation based on the Advanced Ovarian Cancer meta-analysis dataset

From a practical point of view, the most important arguments for using the `jointSurroPenal()` function beyond the standard argument (`data`) concern the following: the parametrization of the model (with arguments `indicator.zeta` and `indicator.alpha`), the method of integration and associated arguments (`int.method`, `n.knots`, `nb.mc`, `nb.gh`, `nb.gh2`, `adaptatif`), the smoothing parameters (`init.kappa` and `kappa.use`) and the scale of survival times (`scale`). Although optional, all these arguments can be used to manage the convergence issues. The choice of the values to assign to these arguments can be based on the convergence of model. When the convergence issues are fixed, users can implement the likelihood cross-validation criteria to evaluate the goodness of fit of different models, as shown later in this section. In the first step, users can try the model with the default values.

In the event of convergence issues, we recommend the following strategy: Changing the number of samples for Monte-Carlo integration (`nb.mc`) by choosing a numerical

value between 100 and 300; varying the number of nodes for the Gaussian-Hermite quadrature integration (**nb.gh** and **nb.gh2**) by choosing the values between 15, 20 and 32; varying the number of nodes for spline (**n.knots**) by a numerical value between 6 and 10; providing new values for the smoothing parameters (**init.kappa**). Users can also set the arguments α or ζ to 1 (**indicator.zeta = 1** or **indicator.alpha = 1**) to avoid estimating these parameters. We also recommend changing the integration method with the arguments **int.method** and **adaptatif**. For example, by using **adaptatif = 1** for integration over the random effects at the individual level, one could use the pseudo-adaptive quadrature Gaussian-Hermite integration instead of the classical quadrature Gaussian-Hermite method. By changing the scale of the survival times (argument **scale**) and considering years instead of days, it is possible to solve some of the numerical issues.

Using the default values based on the advanced ovarian cancer dataset, the model did not converge. By changing the value of some arguments, we obtained the following set of arguments/values which allowed it to convergence of model:

```
joint.surro.ovar <- jointSurroPenal(data = dataOvarian, n.knots = 8,
    indicator.alpha = 0, nb.mc = 200, scale = 1/365)
```

In this model, we fix the coefficient α to 1, and thereby do not estimate it. We consider 8 spline nodes for the baseline hazards. By default, we use the fixed initial values and obtain smoothing parameters by cross-validation on reduced models. We approximate integrals over the random effects using a combination of Monte-Carlo with 200 samples and classical Gauss-Hermite quadrature with 32 nodes. To solve numerical problems during estimation, we re-scale the survival times by converting days to years. This parametrization of the model provided the results described in the next section.

Summary of results

By applying the function `summary()` on the object `joint.surro.ovar`, the following results are displayed in the event of convergence:

```
summary(joint.surro.ovar)

 Estimates for variance parameters of random effects
          Estimate Std Error      z       P
 theta       6.848    0.3786 18.086  < e-10 ***
 zeta        1.792    0.0714 25.095  < e-10 ***
 gamma       0.045    0.0774  0.576  0.5645
 sigma2_S    0.610    0.3733  1.633  0.1025
 sigma2_T    1.830    1.0202  1.794 0.07287   .
 sigma_ST    1.056    0.6067  1.741  0.0817   .

 Estimates for fixed treatment effects
        Estimate Std Error      z        P
 beta_S   -0.596    0.2298 -2.595 0.009463 **
 beta_T   -0.841    0.3936 -2.136  0.03264  *
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 hazard ratios (HR) and confidence intervals for fixed treatment effects
        exp(coef) Inf.95.CI Sup.95.CI
 beta_S     0.551     0.351     0.864
 beta_T     0.431     0.199     0.933

 Surrogacy evaluation criterion
              Level Estimate Std Error Inf.95.CI Sup.95.CI Strength
 Ktau    Individual    0.683        --     0.664     0.696
 R2trial      Trial    1.000     0.001     0.998     1.002     High
 R2.boot      Trial    0.982        --     0.896     1.000     High
 ---
 Association strength: <= 0.49 'Low'; ]0.49-0.72[ 'Medium'; >= 0.72 'High'
 ---
 Surrogate threshold effect (STE): -0.273 (HR = 0.761 )

 Convergence parameters
 Penalized marginal log-likelihood =  -10892.611
 Number of iterations =  29
 LCV = the approximate likelihood cross-validation criterion
       in semi-parametrical case     =  9.162
 Convergence criteria:
   parameters =  9.573e-06 likelihood =  8.426e-08 gradient =  4.507e-08
```

The results are organized in five parts. We first present estimates for the variance parameters of the random effects and the coefficients ζ and α (if applicable). This includes standard errors, z-statistics and p value of the Wald test. Results suggest a strong heterogeneity at the individual level, observed on the endpoints ($\theta = 6.848$ compare to 0), and more pronounced on the true endpoint ($\zeta = 1.792$ compare to 1). Observation on γ suggests homogeneous baseline hazards across trials ($\gamma = 0.045$, $p > 0.5$), both on the surrogate endpoint and on the true endpoint. This could explain

the identification problem encountered by considering the coefficient α in the model. The parameters σ_S^2, σ_T^2, σ_{ST} suggest the presence of heterogeneity at trial level interacting with the treatment ($p <= 0.10$). The next two parts of the results show estimates for the fixed treatment effects β_S given the random effects (u_i, v_{S_i}) and β_T given (u_i, v_{T_i}), with the associated hazard ratios and confidence intervals. These parameters can be interpreted as usual, but adjustment on the random effects taking into account. We observed significant protective effects of the treatment on the surrogate endpoint and on the true endpoint ($p < 0.05$).

The fourth part of the results describes the surrogacy evaluation criterion. Kendall's τ, R^2_{trial} and $R^2_{trial,boot}$ (obtained using parametric bootstrap) are available with the associated confidence intervals as is the standard error of R^2_{trial} obtained by the Delta method [30]. Arguments `int.method.kt` and `nb.gh` of the `summary()` function can be used to choose between the Monte-Carlo and the Gauss-Hermite quadrature which integration method is to be used to estimate Kendall's τ, and set the number of quadrature nodes when appropriate. Using at least `500` samples for the Monte-carlo integration and at least `15` quadrature nodes the two integration methods generally yield the same results for Kendall's τ.

These results suggest high association measurement at the individual level (Kendall's $\tau = 0.68$ $[0.66 - 0.70]$), and high correlation strength at the trial level ($R^2_{trial,boot} = 0.98$ $[0.90 - 1.00]$) between the surrogate endpoint and the true endpoints, according to the classification of the surrogacy criteria proposed by the Institute of Quality and Efficiency in Health Care [31, 32]. Given that Kendall's τ is adjusted on random effects at the individual level [17], it is quite difficult to observe a value `> 0.7` compared to unadjusted ones from the two-step copula approach of Burzykowski `et al.` [8]. A very high value suggests extreme values for the parameters α, ζ, θ or γ, although such values are difficult to observe in practice. Therefore, a value `around 0.65` can be considered as sufficient for validating surrogacy at the individual level.

We also compute and display the surrogate threshold effect with the associated hazard risk. We obtain an acceptable value of STE (- 0.273, HR = 0.761), which illustrates the high validity of the surrogate. As mentioned by [22], unrealistically large/small values of STE (e.g., corresponding to a HR of less than 0.5) would indicate too wide prediction limits and, consequently, poor validity of the surrogate. Therefore, as observed previously [8], PFS can be considered as a valid surrogate endpoint for OS when evaluating new treatments for advanced ovarian cancer.

The last part of the results describes the convergence parameters.

Model estimation based on generated dataset

Here, we estimate two joint surrogate models for the purpose of model comparison, based on the generated dataset `data.sim`. Integrals are approximated using a combination of Monte Carlo and classical Gauss-Hermite in the first model and a combination of Monte Carlo and Pseudo-adaptive Gauss-Hermite integration in the second one. The codes for both models are described as follows:

```
joint.surro.sim.MCGH <- jointSurroPenal(data = data.sim, int.method = 2,
    nb.mc = 300, nb.gh = 20)
joint.surro.sim.MCPGH <- jointSurroPenal(data = data.sim, int.method = 2,
    nb.mc = 300, nb.gh = 20, adaptatif = 1)
```

A relevant question in this case might be how to compare different models, or how to choose the optimal value of the number of knots for spline, the number of quadrature points, the number of samples for Monte-Carlo, or the optimal integration method. We propose in this package to base comparison on the approximated likelihood

cross-validation criterion. The lower the value obtained for this parameter, the better the associated model will be.

Choice of model based on `LCV`

The LCV for models `joint.surro.sim.MCGH` and `joint.surro.sim.MCPGH` are respectively

```
joint.surro.sim.MCGH$LCV

 [1] 8.29982

joint.surro.sim.MCPGH$LCV

 [1] 8.31713
```

As expected [17], the two models are quite similar based on the observed values of LCV. The `summary()` function applied to these objects also gives similar results.

```
summary(joint.surro.sim.MCGH)
```

```
Estimates for variance parameters of random effects
         Estimate Std Error     z         P
theta       3.450    0.4928 7.001    < e-10 ***
zeta        1.506    0.2364 6.369 1.899e-10 ***
gamma       1.881    0.5602 3.358 0.0007853 ***
alpha       0.916    0.1443 6.348 2.183e-10 ***
sigma2_S    0.703    0.4289 1.640    0.1011
sigma2_T    1.096    0.6147 1.783   0.07451   .
sigma_ST    0.442    0.3974 1.113    0.2657

Estimates for fixed treatment effects
       Estimate Std Error      z        P
beta_S   -2.046    0.2667 -7.673   < e-10 ***
beta_T   -1.844    0.3562 -5.177 2.25e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

hazard ratios (HR) and confidence intervals for fixed treatment effects
       exp(coef) Inf.95.CI Sup.95.CI
beta_S     0.129     0.077     0.218
beta_T     0.158     0.079     0.318

Surrogacy evaluation criterion
             Level Estimate Std Error Inf.95.CI Sup.95.CI Strength
Ktau    Individual    0.596        --     0.542     0.625
R2trial      Trial    0.254     0.276    -0.288     0.796      Low
R2.boot      Trial    0.290        --     0.002     0.767      Low
---
Association strength: <= 0.49 'Low'; ]0.49-0.72[ 'Medium'; >= 0.72 'High'
---
Surrogate threshold effect (STE): -8.523 (HR = 0 )

Convergence parameters
Penalized marginal log-likelihood =  -4957.842
Number of iterations =  14
LCV = approximate likelihood cross-validation criterion
      in the semi-parametrical case     =  8.3
Convergence criteria:
  parameters =  3.833e-05 likelihood =  0.0002426 gradient =  1.137e-06
```

```
summary(joint.surro.sim.MCPGH)
```

```
Estimates for variance parameters of random effects
         Estimate Std Error     z         P
theta       2.640    0.4295 6.148 7.854e-10 ***
zeta        2.277    0.4010 5.679 1.356e-08 ***
gamma       1.355    0.4174 3.246   0.00117  **
alpha       1.135    0.2285 4.965 6.855e-07 ***
sigma2_S    0.593    0.3471 1.709    0.0875   .
sigma2_T    0.664    0.5771 1.151    0.2498
sigma_ST    0.380    0.3219 1.181    0.2376

Estimates for fixed treatment effects
       Estimate Std Error      z         P
beta_S   -1.643    0.2277 -7.216    < e-10 ***
beta_T   -1.640    0.3573 -4.589 4.463e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

hazard ratios (HR) and confidence intervals for fixed treatment effects
       exp(coef) Inf.95.CI Sup.95.CI
beta_S     0.193     0.124     0.302
beta_T     0.194     0.096     0.391

Surrogacy evaluation criterion
             Level Estimate Std Error Inf.95.CI Sup.95.CI Strength
Ktau    Individual    0.577        --     0.522     0.607
R2trial      Trial    0.367     0.358    -0.334     1.068      Low
R2.boot      Trial    0.407        --     0.007     0.964      Low
---
Association strength: <= 0.49 'Low'; ]0.49-0.72[ 'Medium'; >= 0.72 'High'
---
Surrogate threshold effect (STE): -4.922 (HR = 0.007 )

Convergence parameters
Penalized marginal log-likelihood =  -4968.465
Number of iterations =  20
LCV = the approximate likelihood cross-validation criterion
      in the semi-parametrical case     =  8.317
Convergence criteria:
  parameters =  5.962e-05 likelihood =  0.0004484 gradient =  2.465e-06
```

Graphical representation of baseline hazard and survival functions

By using the generic function `plot()`, it is possible to plot the baseline hazard and survival functions for both surrogate and true endpoints. The definition of this function is shown below, and the associated arguments are described in S2E Appendix in S2 Appendix.

```
plot(x, endpoint = 2, scale = 1, type.plot = "Hazard", xmin = 0,
  conf.bands = TRUE, xmax = NULL, ylim = c(0,1), Xlab = "Time",
  pos.legend = "topright", main, cex.legend = 0.7,
  Ylab = "Baseline hazard function")
```

Fig 1 represents the baseline survival and hazard functions for model, for both the surrogate and the true endpoints using the advanced ovarian cancer meta-analysis dataset. We limit survival times to 8 months since after this threshold, the estimated survival probabilities are almost equal to `0`. The code below produces the plots given in Fig 1.

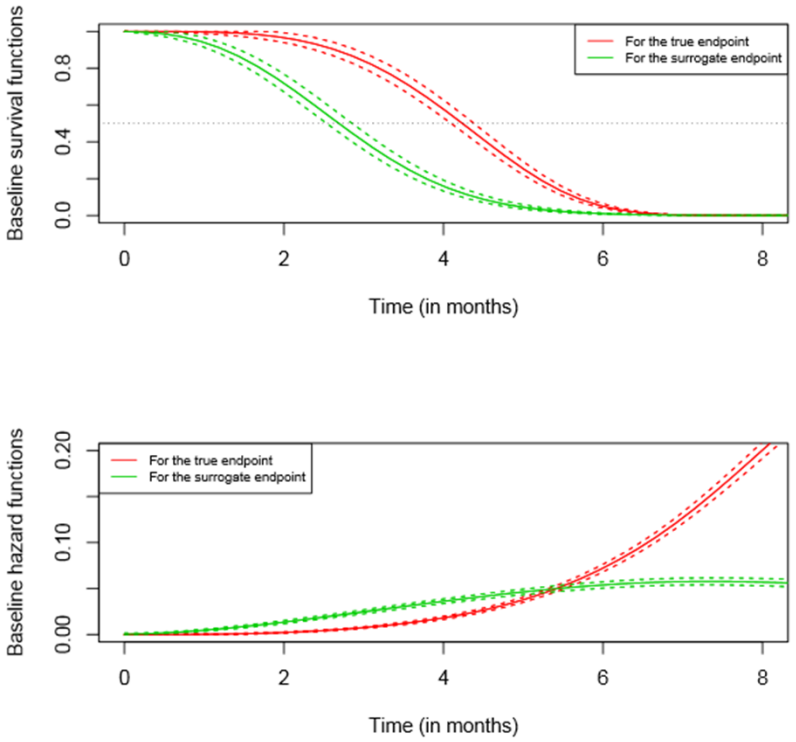


Fig 1. Baseline hazard and survival functions for surrogate endpoint and true endpoint truncated at 8 months using the advanced ovarian cancer meta-analysis dataset.

```
par(mfrow=c(2,1))
plot(joint.surro.ovar,type.plot = "Su", xmax = 8, Xlab = "Time (in months)",
    scale = 12)
plot(joint.surro.ovar, xmax = 8, ylim = c(0,0.2), Xlab = "Time (in months)",
    scale = 12, pos.legend = "topleft")
```

Fig 2 shows another representation of the baseline survival and hazard functions for the surrogate and the true endpoints. We use the object `joint.surro.sim.MCPGH` for this purpose, which is based on the generated data.

The following code is used to produces Fig 2:

```
par(mfrow=c(2,2))
```

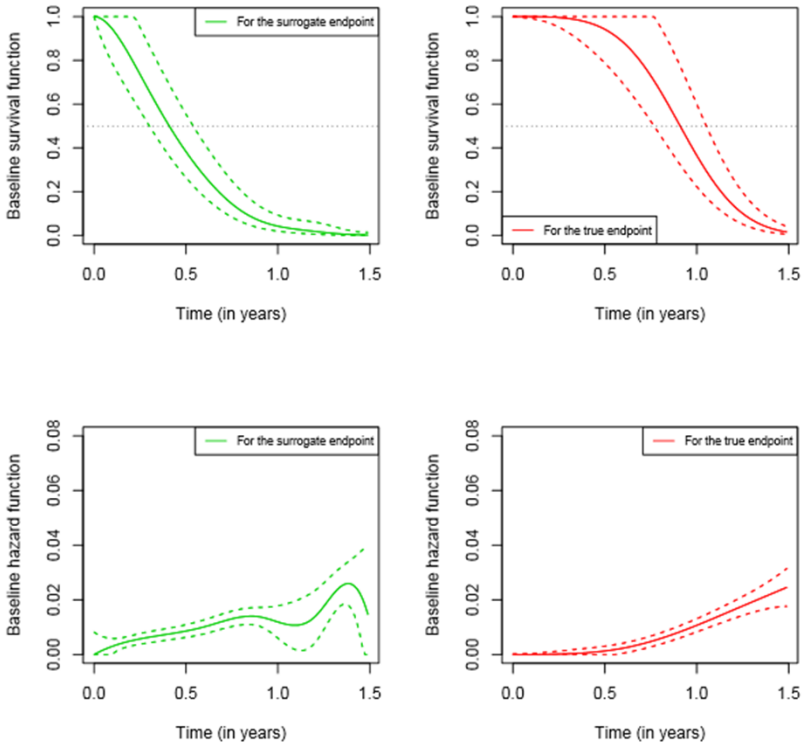


Fig 2. Baseline hazard and survival functions for surrogate endpoint and true endpoint, using simulated meta-analysis of 600 subjects and 30 trials.

```
plot(joint.surro.sim.MCPGH, type.plot = "Su", endpoint = 0, scale = 1/365,
    Xlab = "Time (in years)")
plot(joint.surro.sim.MCPGH, type.plot = "Su", endpoint = 1, scale = 1/365
    ,pos.legend = "bottomleft", Xlab = "Time (in years)")
plot(joint.surro.sim.MCPGH, type.plot = "Ha", endpoint = 0, scale = 1/365
    ,ylim =  c(0,0.08),
    Xlab = "Time (in years)")
plot(joint.surro.sim.MCPGH, type.plot = "Ha", endpoint = 1, scale = 1/365
    ,ylim = c(0,0.08),
    Xlab = "Time (in years)")
```

Model evaluation and prediction

To assess the accuracy of the prediction using estimates from model (1), the leave-one-out cross validation criteria (loocv) described in S2F Appendix in S2 Appendix can be performed as follows:

```
dloocv <- loocv(object = joint.surro.sim.MCGH, unusedtrial = 26,
var.used = "error.estim")
```

We found the following result:

```
dloocv$result

   trialID ntrial  beta.S  beta.T beta.T.i Inf.95.CI Sup.95.CI
1        1     20  -2.145  -0.582   -2.038    -2.663    -1.412
2        2     20  -1.480  -0.799   -1.464    -2.135    -0.793 *
3        3     20  -0.285  -0.422   -0.195    -1.801     1.411 *
4        4     20   0.307   0.487   -0.248    -2.347     1.852 *
5        5     20  -1.087  -1.230   -0.983    -2.007     0.041 *
6        6     20 -21.305  -1.496  -13.951   -32.636     4.733 *
7        7     20  -0.796  -1.943   -0.687    -1.889     0.515
8        8     20  -1.578  -1.302   -1.545    -2.167    -0.923 *
9        9     20  -1.909  -1.402   -1.736    -2.241    -1.230 *
10      10     20  -1.752  -0.053   -1.505    -2.174    -0.836
11      11     20 -21.304  -0.342  -16.325   -35.269     2.619 *
12      12     20  -2.766 -20.920   -2.236    -3.201    -1.271
13      13     20  -0.474  -1.025   -0.835    -2.289     0.618 *
14      14     20   0.056  -0.148   -0.561    -2.603     1.481 *
15      15     20  -1.337  -1.154   -1.250    -2.218    -0.282 *
16      16     20  -0.191  -0.291   -0.125    -1.833     1.582 *
17      17     20   0.264   0.161   -0.540    -3.006     1.926 *
18      18     20  -2.589  -0.657   -2.197    -2.968    -1.426
19      19     20  -1.795  -1.654   -1.562    -2.263    -0.861 *
20      20     20  -0.630   1.599   -1.128    -2.451     0.195
21      21     20  -0.593  -0.510   -0.602    -1.988     0.785 *
22      22     20  -0.682  -1.645   -0.555    -1.827     0.716 *
23      23     20  -0.787  -0.179   -0.850    -2.061     0.362 *
24      24     20  -3.019  -2.735   -2.227    -3.504    -0.949 *
25      25     20  -2.393  -1.577   -2.099    -2.879    -1.319 *
26      27     20  -1.640  -1.063   -1.630    -2.248    -1.012 *
27      28     20  -1.386  -1.672   -1.220    -2.057    -0.383 *
28      29     20  -0.207  -0.722   -0.535    -2.220     1.150 *
29      30     20   0.299   0.185   -0.377    -3.215     2.461 *
```

The returned object, of class `jointSurroPenalloocv` includes for each trial the number of included subjects (`ntrial`), the observed treatment effect on the surrogate endpoint (`beta.S`), the observed treatment effect on the true endpoint (`beta.T`) and the predicted treatment effect on the true endpoint (`beta.T.i`) with the associated prediction interval (`Inf.95.CI, Sup.95.CI`). If the observed treatment effect on the true endpoint is included into the prediction interval, the last column contains `"*"`, indicating a good prediction.

Simulation studies

In this section, we show an example of simulation studies in the `frailtypack` package, based on model (1).

Estimations

Using the function `jointSurroPenalSimul()` simulation studies can be performed as follows:

```
joint.simul10 <- jointSurroPenalSimul(nb.dataset = 10, nbSubSimul =600 ,
```

```
 ntrialSimul = 30, LIMparam = 0.001, LIMlogl = 0.001, LIMderiv = 0.001,
 nb.mc = 200, nb.gh = 20, nb.gh2 = 32, true.init.val = 1, print.iter = F)
```

This function serves to perform simulation studies with `10` meta-analyses, each study including `600` subjects and `30` trials. By default, each generated meta-analysis includes the same proportion of subjects per trial and the same proportion of treated subjects per trial. In the event of an identification problem, the model is re-estimated using `32` quadrature nodes. All unused simulation parameters are set to the initial value, as presented in the function `jointSurroPenalSimul()`. Using default values, we expect `0.81` for R^2_{trial} and `0.595` for Kendall's τ.

Simulation results

Simulation results can be displayed using the S3 method `summary()`. This function allows argument `R2boot` to specify whether the confidence interval of R^2_{trial} will be computed using parametric bootstrapping `(1)` or the Delta method `(0)`.

```
summary(joint.simul10, R2boot = 0)

 Simulation and estimation pamareters
 nb.subject =  600
 nb.trials =  30
 nb.simul =  10
 int.method =  2
 nb.gh =  20
 nb.gh2 =  32
 nb.mc =  200
 kappa.use =  4
 n.knots =  6
 n.iter =  14

 Simulation results
    Parameters True value   Mean Empirical SE Mean SE CP(%)
 2       theta        3.5  3.451        0.711   0.545    80
 3        zeta          1  1.049         0.22   0.177    70
 4       gamma        2.5  2.642        0.957   0.711    80
 5       alpha          1  1.009        0.135   0.138    90
 6     sigma.S        0.7  0.608        0.361   0.426    90
 7     sigma.T        0.7  0.627        0.347   0.459    80
 8    sigma.ST       0.63  0.555        0.314   0.389    90
 9      beta.S      -1.25 -1.368        0.233   0.251    90
 10     beta.T      -1.25 -1.397        0.238   0.269   100
 11    R2trial       0.81   0.82        0.181   0.521    80
 12      K.tau      0.595  0.592        0.032       -    80
 Rejected datasets : n(%) =  0(0)
```

In the first part of the results, we present a brief summary of simulation and estimation parameters, and the average number of iterations to reach convergence `(n.iter = 14)`.

The next part presents a table of simulation results. Each row of the table corresponds to a model parameter. The first column is the name of the parameter, followed by the value assigned to the parameter during simulation. The next three

columns correspond to the average of the estimates observed for all the generated datasets , the empirical standard errors and the mean of the estimated standard errorThe last column is the coverage probability (CP), which is the proportion (%) of the 95% confidence intervals of the estimate that includes the true value of the parameter. We considered 10 meta-analyses here, although simulation studies more often require around `500` datasets of meta-analysis.

The last row of the results indicates the number of rejected datasets due to convergence issues.

Discussion

This paper presents new tools for validating candidate surrogate endpoints using data from multiple randomized clinical trials, with failure time endpoints. Since version `3.0.1`, The `R` `frailtypack` package implements the joint-surrogate model, which is a more attractive approach than two-step approaches for evaluating surrogate endpoints based on a one-step analysis strategy. The joint-surrogate model demonstrated better performances than the two-step copula model or the one-step Poisson approach [17]. Furthermore, the new model showed stable results even with a moderate trial size or number of trial as commonly encountered in practice, whereas the adjusted model estimated with the Bayesian framework showed unstable results [11]

By varying the values of the arguments in the `jointSurroPenal` function , convergence of the model is not always guaranteed. Therefore, it is important in the event of convergence issues to know how to play with the couples arguments/values couple as shown in the previous section. Thus, users can choose the method of integration, initial values, the number of nodes for splines and the smoothing parameters, the number of nodes to use for the Gauss-Hermite quadrature and the number of samples for the Monte-Carlo integration when applicable, the random number generator, and other necessary arguments. It is also possible to set some parameters of the model in the event of identifiability issues. This underlines the flexibility of the `frailtypack` package in managing convergence issues. This flexibility is quite different from that obtained with the `surrosurv` package [9] or macros `SAS` [10] for evaluating surrogate endpoints using the two-step Copula model or one-step Poisson model. Other advantages of our model compared to existing approaches [8, 13] are in the reduction of convergences and numerical issues, the robustness to model misspecification, the surrogacy evaluation based on a one-step approach and therefore the estimation of R^2_{trial} without need for adjustment on estimation errors. In addition, as underlined in the illustration section, the interpretation of Kendall's τ is different from that in the two-step copula approach.

Our previous paper [17] demonstrated the robustness of the joint surrogate model to model misspecification, numerical integration and variations in data characteristics regarding the surrogacy evaluation criteria (R^2_{trial} and Kendall's τ). It is robust to variations in the values of the arguments regarding the surrogacy evaluation criteria. Thus, in the event of convergence, change in arguments/values mostly produced similar results. For example, when we reduced the number of samples for Monte-Carlo integration to `100 (nb.mc = 100)` in the application based on the advanced ovarian cancer meta-analysis dataset, we observed `R2trial = 1.000 [95%CI: 0.998 - 1.002), R2boot = 0.981 [95%CI: 0.891 - 1.000]`, `Kendall's` τ `= 0.683 [95%CI: 0.664 - 0.695]`, `STE = -0.291 (HR = 0.747)` and `LCV = 9.161`. These results are quite similar to those using `nb.mc = 200` (see illustration section in manuscript). In addition, if we integrate over the random effect at the individual level using the pseudo-adaptive Gaussian-Hermite quadrature (argument `adaptatif = 1`) instead of the classical Gaussian-Hermite quadrature, the results are similar with

R2trial = 1.000 [95%CI: 0.998 - 1.002], R2boot = 0.982 [95%CI: 0.897 - 1.000], Kendall's τ = 0.683 [95%CI: 0.664 - 0.696], STE = -0.272 (HR = 0.762) and LCV = 9.162. These examples confirm the robustness of the model previously discussed by Sofeu et al. (2019) using simulation studies.

Moreover, thanks to the `jointSurroPenalSimul()` function, it is possible to perform simulation studies in order to plan a new trial and define the optimal number of clusters when evaluating surrogate endpoints given the joint surrogate model. For example, if a given meta-analysis includes few trials, simulation studies may help in establishing the minimum number of centers to obtain the best estimate of the surrogacy evaluation criteria. Jurgen et al. [33] suggested using clinical trial simulations to optimize adaptive trial designs. As they explained, the typical goal of a clinical trial simulation is to identify a design that has a high probability of success based on the most likely conditions but which can also perform well, or at least acceptably, under more extreme conditions if neccesary. Simulation studies can help if the recommended values for the arguments do not make it possible to reach convergence or involve longer computer time when fitting the joint surrogate model. Given the data characteristics, they can help in choosing optimal values for some arguments (the number of quadrature nodes, the number of samples for the Monte-Carlo integration and the number of nodes for splines) and in anticipating their impact on estimating the model parameters. The management of the convergence issues by the program itself is described in S2G Appendix in S2 Appendix.

Numerous tools have been presented in this paper for evaluating surrogacy. We have the following: the surrogate threshold effect which is used in combination with R^2_{trial} to assess the validity of the potential surrogate endpoint; the `predict()` function used in a new trial to predict the treatment effect of the true endpoint based on the observed treatment effect on the surrogate endpoint; and the leave-one-out cross-validation which can be used to assess the accuracy of the prediction using model (1). Furthermore, a graphical representation of the baseline hazard and survival functions is possible using the `plot()` function.

The `jointSurroPenal()` function can also be used in interim analyses to estimate the fixed treatment effect on the surrogate endpoint, taking into account competing risk of death and heterogeneity in the data at the individual level and at the trial level in interaction with treatment. This is an alternative to the joint frailty-copula model between tumor progression and death for meta-analysis proposed in [34].

We now plan to extend the model (1) and the `jointSurroPenal()` function to take into account interval censoring for endpoints where the exact event times are unknown. This extension will also make it possible to model the baseline hazard functions parametrically, using a Weibull distribution. To improve the use of `frailtypack`, intuition can be gained by developing an associated interactive web app using the `R` package `Shiny` available at `https://CRAN.R-project.org/package=Shiny`.

Supporting information

S1 Fig. Package characteristics (version 3.0.3.1). Blue cross is for the option available for a given type of model in the package on CRAN, orange cross is for the option included in the package but not yet on CRAN yet. Empty cells mean that an option is not available for a given type of model. RE = Recurrent Event. TE = Terminal Event. LO = Longitudinal Outcome. STE = Surrogate Threshold Effect. ODE = Ordinary Differential Equation.

S1A Appendix. Penalized likelihood approach.

S1B Appendix. Individual-level surrogacy.

S1C Appendix. Trial-level surrogacy.

S1D Appendix. Derivation of the surrogate threshold effect.

S1E Appendix. Interpretation of STE and decision-making.

S2A Appendix. The `jointSurroPenal()` function.

S2B Appendix. The `jointSurrSimul()` function.

S2C Appendix. The `jointSurroPenalSimul()` function.

S2D Appendix. The `jointSurroTKendall()` function.

S2E Appendix. The `plot()` function.

S2F Appendix. The `loocv()` function.

S2G Appendix. Management of convergence issues.

Acknowledgments

The authors thank the Ovarian Cancer Meta-Analysis Project for sharing the data used to illustrate the programs. This work was supported by the Association pour la Recherche sur le Cancer, Grant/Award Number: PJA20161205147; Institut National du Cancer, Grant/Award Number: 2017-125; Institut national de la santé et de la recherche médicale; Région Aquitaine. We also thank Antoine Barbieri, INSERM U1219, for his support for programming the Bayesian approach. We gratefully acknowledge very helpful and constructive comments and suggestions from the academic editor and the three anonymous referees, which lead to significant improvements of this manuscript

References

1. Fleming TR, DeMets DL. Surrogate End Points in Clinical Trials: Are We being Misled? Annals of Internal Medicine. 1996;125(7):605–613. doi:10.7326/0003-4819-125-7-199610010-00011.

2. Matulonis UA, Oza AM, Ho TW, Ledermann JA. Intermediate Clinical Endpoints: A Bridge Between Progression-Free Survival and Overall Survival in Ovarian Cancer Trials. Cancer. 2015;121(11):1737–1746. doi:10.1002/cncr.29082.

3. Ellenberg SS, Hamilton JM. Surrogate Endpoints in Clinical Trials: Cancer. Statistics in Medicine. 1989;8(4):405–413. doi:10.1002/sim.4780080404.

4. Booth CM, Eisenhauer EA. Progression-Free Survival: Meaningful or Simply Measurable? Journal of Clinical Oncology. 2012;30(10):1030–1033. doi:10.1200/JCO.2011.38.7571.

5. Prentice RL. Surrogate Endpoints in Clinical Trials: Definition and operational criteria. Statistics in Medicine. 1989;8(4):431–440. doi:10.1002/sim.4780080407.

6. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The Validation of Surrogate Endpoints in Meta-Analyses of Randomized Experiments. Biostatistics. 2000;1(1):49–67.

7. Burzykowski T, Molenberghs G, Buyse M, Geys H. The Evaluation of Surrogate Endpoints. Springer-Verlag, New-york, NK; 2005.

8. Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of Surrogate End Points in Multiple Randomized Clinical Trials with Failure Time End Points. Journal of the Royal Statistical Society C (Applied Statistics). 2001;50(4):405–422. doi:10.1111/1467-9876.00244.

9. Rotolo F. `surrosurv`: Evaluation of Failure Time Surrogate Endpoints in Individual Patient Data Meta-Analyses; 2017. Available from: `https://CRAN.R-project.org/package=surrosurv`.

10. Alonso A, Bigirumurame T, Burzykowski T, Buyse M, Molenberghs G, Muchene L, et al. Applied Surrogate Endpoint Evaluation Methods with `SAS` and `R`. Chapman and Hall/CRC; 2017.

11. Renfro LA, Shi Q, Sargent DJ, Carlin BP. Bayesian Adjusted R2 for the Meta-Analytic Evaluation of Surrogate Time-To-Event Endpoints in Clinical Trials. Statistics in Medicine. 2012;31(8):743–761. doi:10.1002/sim.4416.

12. Shi Q, Renfro LA, Bot BM, Burzykowski T, Buyse M, Sargent DJ. Comparative Assessment of Trial-Level Surrogacy Measures for Candidate Time-to-Event Surrogate Endpoints in Clinical Trials. Computational Statistics & Data Analysis. 2011;55(9):2748 – 2757. doi:https://doi.org/10.1016/j.csda.2011.03.014.

13. Rotolo F, Paoletti X, Burzykowski T, Buyse M, Michiels S. A Poisson Approach to the Validation of Failure Time Surrogate Endpoints in Individual Patient Data Meta-Analyses. Statistical Methods in Medical Research. 2019;28(1):170–183. doi:10.1177/0962280217718582.

14. Alonso A, Molenberghs G. Surrogate Marker Evaluation from an Information Theory Perspective. Biometrics. 2007;63(1):180–186. doi:10.1111/j.1541-0420.2006.00634.x.

15. Buyse M, Michiels S, Squifflet P, Lucchesi KJ, Hellstrand K, Brune ML, et al. Leukemia-free Survival as a Surrogate End Point for Overall Survival in the Evaluation of Maintenance Therapy for Patients with Acute Myeloid Leukemia in Complete Remission. Haematologica. 2011;96(8):1106–1112. doi:10.3324/haematol.2010.039131.

16. Buyse M, Molenberghs G, Paoletti X, Oba K, Alonso A, der Elst WV, et al. Statistical Evaluation of Surrogate Endpoints with Examples from Cancer Clinical Trials. Biometrical Journal. 2016;58(1):104–132. doi:10.1002/bimj.201400049.

17. Sofeu CL, Emura T, Rondeau V. One-step validation method for surrogate endpoints using data from multiple randomized cancer clinical trials with failure-time endpoints. Statistics in Medicine. 2019;38(16):2928–2942. doi:10.1002/sim.8162.

18. Lin DY, Wei LJ. The Robust Inference for the Cox Proportional Hazards Model. Journal of the American Statistical Association. 1989;84(408):1074–1078. doi:10.1080/01621459.1989.10478874.

19. Van der Elst W, Meyvisch P, Alonso A, Ensor HM, Molenberghs CJWG. `Surrogate`: Evaluation of Surrogate Endpoints in Clinical Trials; 2018. Available from: `https://CRAN.R-project.org/package=Surrogate`.

20. Bujkiewicz S, Thompson JR, Riley RD, Abrams KR. Bayesian Meta-Analytical Methods to Incorporate Multiple Surrogate Endpoints in Drug Development Process. In: Statistics in medicine; 2016.

21. Institute for Quality and Efficiency in Health Care. Validity of Surrogate Endpoints in Oncology: Executive Summary; 2011. Available from: `www.iqwig.de/download/A10-05_Executive_Summary_v1-1_Surrogate_endpoints_in_oncology.pdf`.

22. Burzykowski T, Buyse M. Surrogate Threshold Effect: An Alternative Measure for Meta-Analytic Surrogate Endpoint validation. Pharmaceutical Statistics. 2006;5(3):173–186. doi:10.1002/pst.207.

23. Król A, Mauguen A, Mazroui Y, Laurent A, Michiels S, Rondeau V. Tutorial in Joint Modeling and Prediction: A Statistical Software for Correlated Longitudinal Outcomes, Recurrent Events and a Terminal Event. Journal of Statistical Software, Articles. 2017;81(3):1–52. doi:10.18637/jss.v081.i03.

24. Ovarian cancer Meta-Analysis Project. Cyclophosphamide Plus Cisplatin Plus Adriamycin Persus Cyclophosphamide, Doxorubicin, and Cisplatin Chemotherapy of Ovarian Carcinoma: A Meta-Analysis. Classic Papers and Current Comments. 1991;3:237–234.

25. Marquardt DW. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. Journal of the Society for Industrial and Applied Mathematics. 1963;11(2):431–441. doi:10.1137/0111030.

26. Joly P, Commenges D, Letenneur L. A Penalized Likelihood Approach for Arbitrarily Censored and Truncated Data: Application to Age-Specific Incidence of Dementia. Biometrics. 1998;54(1):185–194.

27. Gail MH, Pfeiffer R, van Houwelingen HC, Carroll RJ. On Meta-Analytic Assessment of Surrogate Outcomes . Biostatistics. 2000;1(3):231–246. doi:10.1093/biostatistics/1.3.231.

28. Rondeau V, Mathoulin-Pelissier S, Jacqmin-Gadda H, Brouste V, Soubeyran P. Joint Frailty Models for Recurring Events and Death Using Maximum Penalized Likelihood Estimation: Application on Cancer Events. Biostatistics. 2007;8(4):708–721. doi:10.1093/biostatistics/kxl043.

29. Rondeau V, Gonzalez JR, Mazroui Y, Mauguen A, Diakite A, Laurent A, et al.. `frailtypack`: General Frailty Models: Shared, Joint and Nested Frailty Models with Prediction; Evaluation of Failure-Time Surrogate Endpoints; 2019. Available from: `https://CRAN.R-project.org/package=frailtypack`.

30. Dowd BE, Greene WH, Norton EC. Computation of Standard Errors. Health Services Research. 2014;49(2):731–750.

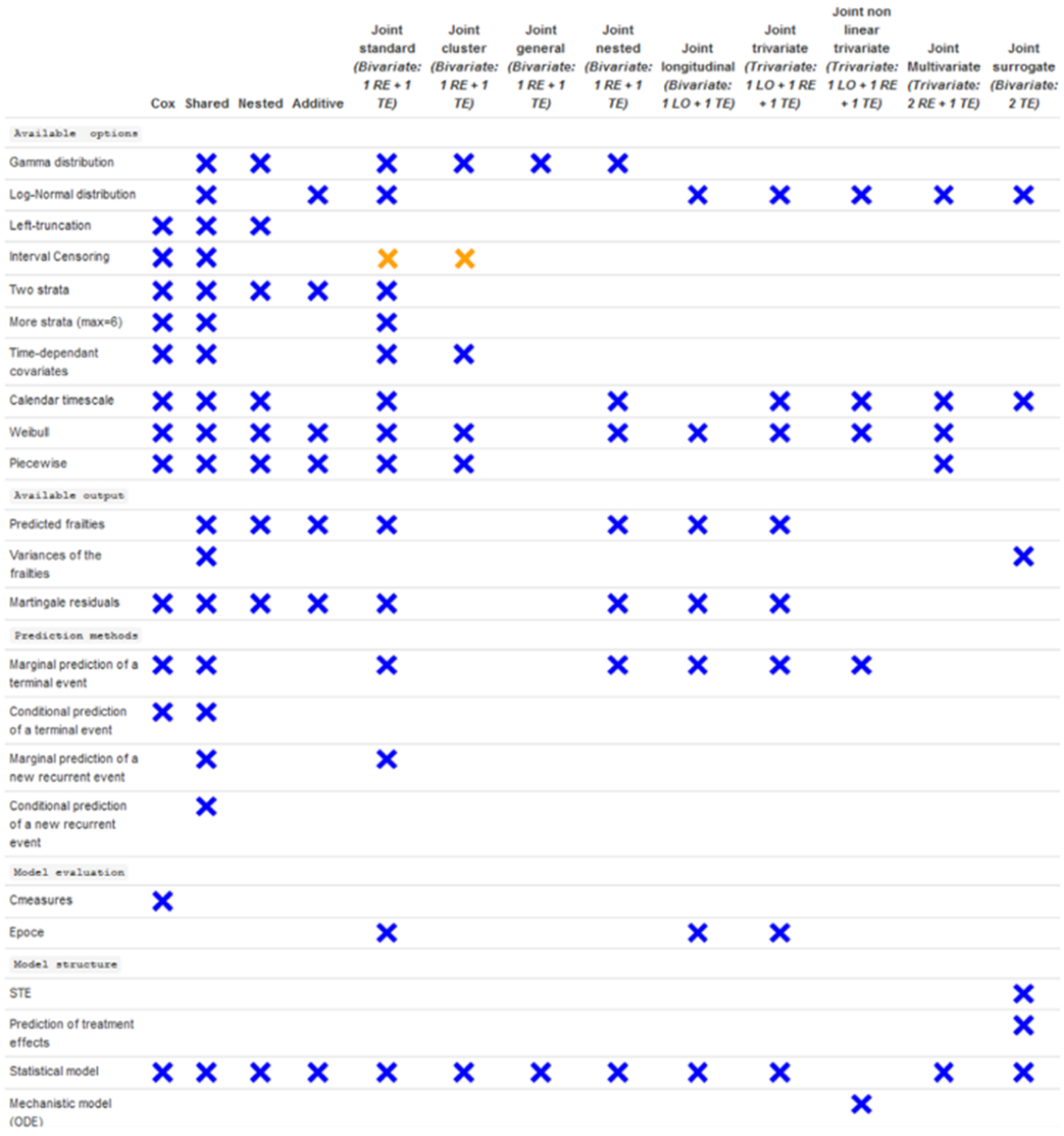
31. Prasad V, Kim C, Burotto M, Vandross A. The Strength of Association Between Surrogate End Points and Survival in Oncology: A Systematic Review of Trial-Level Meta-Analyses. JAMA Internal Medicine. 2015;175(8):1389–1398. doi:10.1001/jamainternmed.2015.2829.

32. Baker SG. Five Criteria for Using a Surrogate Endpoint to Predict Treatment Effect Based on Data from Multiple Previous Trials. Statistics in Medicine. 2018;37(4):507–518. doi:10.1002/sim.7561.

33. Jurgen H, Song W, John K. Using simulation to optimize adaptive trial designs: applications in learning and confirmatory phase trials. Clinical Investigation. 2015;5(4):401–413. doi:10.4155/CLI.15.14.

34. Emura T, Nakatochi M, Murotani K, Rondeau V. A Joint Frailty-Copula Model Between Tumour Progression and Death for Meta-Analysis. Statistical Methods in Medical Research. 2017;26(6):2649–2666. doi:10.1177/0962280215604510.

4.2 Annexes article



S1 Fig.

S1 Appendix

S1A Appendix. Penalized likelihood approach

The baseline hazard functions $\hat{\lambda}_{0S}(.)$ and $\hat{\lambda}_{0T}(.)$ are approximated using a linear combination of Cubic M-splines and polynomial functions of 3^{rd} order [1]. The cumulative hazard functions are approximated using I-splines (integrated M-splines). This allows smoothing functions, which is useful in epidemiology. In the semi-parametric approach, regardless of the expected smooth baseline hazard functions, the likelihood of the model is penalized by a term depending on the roughness of the functions [2]. Therefore, the penalized marginal log-likelihood is given by:

$$pl(\Phi) = l(\Phi) - \kappa_1 \int_0^\infty \lambda_{0S}^{''2}(t)dt - \kappa_2 \int_0^\infty \lambda_{0T}^{''2}(t)dt \tag{.1}$$

where $l(\Phi)$ is the full marginal log-likelihood. κ_1 and κ_2 are positive smoothing parameters which control the trade-off between the data fit and the smoothness of the functions. The smoothing parameters are chosen using a maximizing likelihood cross-validation criterion described in Joly *et al.* [2], based on two separate Cox proportional hazard models with no covariates. An alternative is to set κ_1 and κ_2 manually

S1B Appendix.Individual-level surrogacy

To measure the strength of the association between S_{ij} and T_{ij} after adjusting the marginal distributions for the trial and the treatment effects, we defined Kendall's τ as follows:

$$\begin{aligned}
\tau \quad = \quad & 2\int_{u_i}\int_{\omega_{ij}}\int_{u_{i'}}\int_{\omega_{i'j'}} \\
& \frac{\exp(\omega_{ij}+u_i+\zeta\omega_{ij}+\alpha u_i)+\exp(\omega_{i'j'}+u_{i'}+\zeta\omega_{i'j'}+\alpha u_{i'})}{(\exp(\omega_{i'j'}+u_{i'})+\exp(\omega_{ij}+u_i))(\exp(\zeta\omega_{i'j'}+\alpha u_{i'})+\exp(\zeta\omega_{ij}+\alpha u_i))} \\
& \frac{1}{\sqrt{2\pi\theta}}\exp\left[-\frac{1}{2}\frac{\omega_{i'j'}^2}{\theta}\right]\frac{1}{\sqrt{2\pi\gamma}}\exp\left[-\frac{1}{2}\frac{u_{i'}^2}{\gamma}\right]d\omega_{i'j'}du_{i'} \\
& \frac{1}{\sqrt{2\pi\theta}}\exp\left[-\frac{1}{2}\frac{\omega_{ij}^2}{\theta}\right]\frac{1}{\sqrt{2\pi\gamma}}\exp\left[-\frac{1}{2}\frac{u_i^2}{\gamma}\right]d\omega_{ij}du_i - 1
\end{aligned}$$

where θ, ζ, α and γ are estimated using the model (1). Kendall's τ is the difference between the probability of concordance and the probability of discordance of two realizations of S_{ij} and T_{ij}. It belongs to the interval [-1,1] and assumes a zero value when S_{ij} and T_{ij} are independent. We estimate it using Monte-Carlo or Gauss-Hermite quadrature integration methods, and parametric bootstrapping for confidence interval.

S1C Appendix. Trial-level surrogacy

The key reason for validating a surrogate endpoint is to be able to predict the effect of treatment on the true endpoint, based on the observed effect of treatment on the surrogate endpoint. As shown by Buyse *et al.*

[3], the coefficient of determination obtained from the covariance matrix Σ_v of the random effects treatment-by-trial interaction can be used to evaluate underlined prediction, and therefore as surrogacy evaluation measurement at trial level. It is defined by:

$$R^2_{trial} = \frac{\sigma^2_{v_{ST}}}{\sigma^2_{v_S}\sigma^2_{v_T}} \tag{.2}$$

The SEs of R^2_{trial} is calculated using the Delta method [4]. We also propose R^2_{trial} and 95% CI computed using parametric bootstrapping. Use of the Delta method can lead to confidence limits violating the [0,1], as noted by Burzykowski *et al.* [5]. However, using other methods would not significantly alter the findings of the surrogacy assessment.

S1D Appendix. Derivation of the surrogate threshold effect

Assume the distribution of the trial-specific treatment effects observed on the surrogate and true endpoints :

$$\begin{pmatrix}\beta_{S_i}\\ \beta_{T_i}\end{pmatrix} \sim MVN\left(\beta, \Sigma_v\right), \quad with \quad \beta = (\beta_S, \beta_T)^\top, \quad and \quad \Sigma_v = \begin{pmatrix}\sigma^2_{v_S} & \sigma_{v_{ST}}\\ \sigma_{v_{ST}} & \sigma^2_{v_T}\end{pmatrix},$$

where β and Σ_v represent the fixed treatment effects and the variance-covariance matrix described in model (1). Let $i = 0$ the new trial for which data are available on the surrogate endpoint but not on the true endpoint. If β_{S_0} represent the observed treatment effect in trial 0 from a Cox model and ϑ represents the fixed-treatment effects parameters and variance components related to model (1), as shown in [3], the conditional mean of the treatment effect in trial 0, $\beta_T + v_{T0}$ can be written as

$$E(\beta_T + v_{T_0}|\beta_{S_0}, \vartheta) = \beta_T + \frac{\sigma_{v_{ST}}}{\sigma_{v_{SS}}}(v_{S_0} - \beta_S), \tag{.3}$$

with the conditional variance:

$$Var(\beta_T + v_{T0}|\beta_{S_0}, \vartheta) = Var(v_{T0})(1 - R^2_{trial}), \tag{.4}$$

where R^2_{trial} is defined in (.2), and v_{T_0} is the random effect treatment-by-trial interaction associated with the true endpoint. In practice, ϑ and β_{S_0} are unknown and have to be estimated. Model (1) can be fitted to the data to estimate ϑ, and a Cox proportional hazard model can be fitted to the data from the new trial to estimate β_{S_0}. The corresponding estimates may be denoted by $\hat{\vartheta}$ and $\hat{v}_{S_0}$. Therefore, the formulation of the prediction variance is:

$$Var(\beta_T + v_{T0}|\beta_{S_0}, \vartheta) \approx f\{Var(\hat{\beta}_{S_0})\} + f\{Var(\hat{\vartheta}\} + Var(v_{T0})(1 - R^2_{trial}), \tag{.5}$$

where $f\{Var(\hat{\beta}_{S_0})\}$ and $f\{Var(\hat{\vartheta}\}$ are functions of the asymptotic variance-covariance matrices of $\hat{v}_{S_0}$ and $\hat{\vartheta}$. These functions describe the contribution to the variability due to the use of the estimates of these parameters. Assume the simple case where the estimation errors are present in the meta-analysis but not in

the new trial. This assumption requires an infinite sample size for the new trial. In this case, $f\{Var(\hat{\beta}_{S_0})\}$ can be reduced to 0, and (.5) therefore can be written as:

$$Var(\beta_T + v_{T0}|\beta_{S_0}, \vartheta) \approx f\{Var(\hat{\vartheta}\} + Var(v_{T0})(1 - R^2_{trial}). \tag{.6}$$

Let $x = (1, -\sigma_{v_{ST}}/\sigma_{SS})^\top$. Given that in linear mixed-effects models the maximum likelihood estimates of the covariance parameters are asymptotically independent of the fixed effects parameters, one can rewrite the prediction variance (.7) as a quadratic function of β_{S_0} [6] as:

$$Var(\beta_T + v_{T0}|\beta_{S_0}, \vartheta) \approx x^\top \left[V_\mu + \left(\frac{\beta_{S_0} - \beta_S}{\sigma_{v_{SS}}} \right)^2 V_D \right] x + \sigma_{v_{TT}}(1 - R^2_{trial}), \tag{.7}$$

where V_μ and V_D are the asymptotic variance-covariance matrices of $(\hat{\beta}_T, \hat{\beta}_S)^\top$ and $(\hat{\sigma}_{v_{ST}}, \hat{\sigma}_{v_{SS}})^\top$, respectively. The limits of the $(1-\gamma)100\%$ prediction interval for $\beta_T + v_{T0}$ are functions of v_{S_0}, and correspond to:

$$l(\beta_{S_0}) \equiv E(\beta_T + v_{T0}|\beta_{S_0}, \vartheta) - z_{1-(\gamma/2)} \sqrt{Var(\beta_T + v_{T0}|\beta_{S_0}, \vartheta)} \tag{.8}$$

for the lower prediction limit function, and

$$u(\beta_{S_0}) \equiv E(\beta_T + v_{T0}|\beta_{S_0}, \vartheta) + z_{1-(\gamma/2)} \sqrt{Var(\beta_T + v_{T0}|\beta_{S_0}, \vartheta)}, \tag{.9}$$

for the upper prediction limit function. In (.8)-(.9), $z_{1-(\gamma/2)}$ is the $(1-(\gamma/2))$ quantile of the standard normal distribution. Assume that $\sigma_{v_{ST}} > 0$ and that negative values of α_i indicate a beneficial treatment effect in trial i. One can compute a value of β_{S_0} such that the upper prediction limit

$$u(\beta_{S_0}) \equiv 0. \tag{.10}$$

This value represents the STE. The solution(s) of equation (.10) can be obtained by solving a quadratic equation. The number of solutions of the equation depends on the configuration of the parameters of $u(\beta_{S_0})$. In the event of two solutions, STE is the lower value. Therefore, an observed value of treatment effect on a surrogate endpoint higher than STE predicts a significant treatment effect on the true endpoint. Note that there may be two possible values for STE depending on the variance used to compute $u(\beta_{S_0})$. Thus, STE can be influenced by the prediction variance, and then by the characteristics of the meta-analysis, and the new trial.

S1E Appendix. Interpretation of STE and decision-making

A large value of STE would point to the need to observe a large treatment effect on the surrogate endpoint in order to conclude in a non-zero treatment effect on the true endpoint. If the STE is high, it would not be reasonable to use the surrogate [6], even if it were potentially valid (with $R^2_{trial} \approx 1$).

The IQWiG suggest basing the prediction of the treatment effect on the true endpoint both on the treatment effect observed on the surrogate endpoint, on R^2_{trial} and on STE [7]. The use of STE is mainly

required in the event of a moderate correlation with $0.7 < \mathrm{R} < 0.85$ ($0.49 < R^2_{trial} < 0.72$). In this case, if the upper bound of the 95% CI (or the upper bound of the 80% CI) of the treatment effect on the surrogate is lower than STE, there is at most an indication of an effect on the true endpoint. If the correlation is high ($\mathrm{R} \geq 0.85$ or $R^2_{trial} \geq 0.72$), and if there is a treatment effect on the surrogate, there is at most an indication of an effect on the true endpoint.

References

[1] J. O. Ramsay, Monotone regression splines in action, Statist. Sci. 3 (4) (1988) 425–441. doi:10.1214/ss/1177012761.
URL https://doi.org/10.1214/ss/1177012761

[2] P. Joly, D. Commenges, L. Letenneur, A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia, Biometrics 54 (1) (1998) 185–194.
URL http://www.jstor.org/stable/2534006

[3] M. Buyse, G. Molenberghs, T. Burzykowski, D. Renard, H. Geys, The validation of surrogate endpoints in meta-analyses of randomized experiments, Biostatistics 1 (1) (2000) 49–67.

[4] B. E. Dowd, W. H. Greene, E. C. Norton, Computation of standard errors, Health Services Research 49 (2) (2014) 731–750.
URL http://doi.org/10.1111/1475-6773.12122

[5] T. Burzykowski, G. Molenberghs, M. Buyse, H. Geys, D. Renard, Validation of surrogate end points in multiple randomized clinical trials with failure time end points, Journal of the Royal Statistical Society C (Applied Statistics) 50 (4) (2001) 405–422. doi:10.1111/1467-9876.00244.
URL http://dx.doi.org/10.1111/1467-9876.00244

[6] T. Burzykowski, M. Buyse, Surrogate threshold effect: An alternative measure for meta-analytic surrogate endpoint validation, Pharmaceutical Statistics 5 (3) (2006) 173–186. doi:10.1002/pst.207.
URL http://dx.doi.org/10.1002/pst.207

[7] Institute for Quality and Efficiency in Health Care, Validity of surrogate endpoints in oncology: Executive summary (2011).
URL www.iqwig.de/download/A10-05_Executive_Summary_v1-1_Surrogate_endpoints_in_oncology.pdf

S2 Appendix

S2A Appendix. The `jointSurroPenal()` function

Description:. The mandatory argument of this function is `data`, which must be a dataframe containing at least seven columns named:

- `patienID:` a numeric, that represents the patient's identifier and must be unique;
- `trialID:` a numeric, that represents the trial in which each patient was randomized;
- `timeS:` the follow up time associated with the surrogate endpoint;
- `statusS:` the event indicator associated with the surrogate endpoint, with 0 = no event, 1 = event;
- `timeT:` the follow up time associated with the true endpoint;
- `statusT:` the event indicator associated with the true endpoint, with 0 = no event, 1 = event;
- `trt:` the treatment indicator for each patient, with 1 = treated, 0 = untreated.

Argument `maxit` indicates the maximum number of iterations to be reached by the Marquardt algorithm, the default being `40`. To simplify the model, it is possible to set the coefficients α or ζ to `1` and not to estimate them, or to assume homogeneity on the baseline hazard functions. Arguments `indicator.zeta` and `indicator.alpha` are binary and indicates whether ζ and α should be estimated (1) or not (0). These parameters could be set to `0` in the event of convergence and identification issues. Argument `frail.base` indicates whether the heterogeneity between trials on the baseline risk using the shared cluster specific frailties (u_i) should be considered (`1`), or not (`0`).

Argument `n.knots` indicates the number of knots between 4 and 20, which corresponds to `n.knots + 2` splines functions for approximating the baseline hazard functions (the same number of hazard functions for both endpoints). This value is required in the penalized likelihood estimation. The convergence thresholds of the Marquardt algorithm are for the difference between the consecutive values of estimated coefficients (`LIMparam`), the difference between two consecutive penalized log-likelihoods (`LIMlogl`) and for the small gradient of the log-likelihood (`LIMderiv`). By default, these thresholds values are set to 10^{-3}.

The log-likelihood formulation associated with the joint model (1) includes two integration levels which cannot be solved analytically. Argument `int.method` is a numeric which indicates the integration method to be used for approximating integrals over the random effects. When this parameter is set to `0`, the full Monte Carlo integration method is used; if set to `1` the full Gauss-Hermite quadrature is used; if set to `2`, a combination of both Gauss-Hermite quadrature for integration over the individual-level random effects and Monte Carlo for integration over the trial-level random effects is used; if set to `4`, a combination of both Monte Carlo to integrate over the individual-level random effects and Gauss-Hermite quadrature to integrate over the trial-level random effects is used. In the event of Gauss-Hermite quadrature integration, argument

`adaptatif` is used to indicate whether the pseudo adaptive Gauss-Hermite quadrature `(1)` [1] or the classical Gauss-Hermite quadrature `(0)` should be used. Argument `nb.gh` indicates the number of nodes to be used in the Gauss-Hermite quadrature integration method and can be 5, 7, 9, 12, 15, 20 or 32. Argument `nb.gh2` is the number of nodes for the Gauss-Hermite quadrature in the event of convergence issues to re-estimate the model. Argument `nb.iterPGH` indicates the number of iterations before the re-estimation of the posterior random effects, in the event of the two-steps pseudo-adaptive Gauss-Hermite quadrature [2]. If this parameter is set to `0`, there is no re-estimation. With the Monte-Carlo integration method, argument `nb.mc` is used to indicate the number of samples considered, a value between 100 and 300 normally giving good results. However, beyond 300, the program takes a long computing time to obtain estimates.

By default, the integrals in Kendall's τ formulation are approximated using the Monte-Carlo integration. Argument `nb.MC.kendall` indicates the number of generated samples used in this case. It is advisable to use at least 4000 samples for stable results. Argument `nboot.kendall` specifies the number of samples considered in the parametric bootstrap to estimate the confidence interval of Kendall's τ, or R^2_{trial}. The default is `1000`.

Argument `true.init.val` is numeric with two values, `0` or `2`, that indicate how to consider initial values of model parameters. If set to `(0)`, initial values given to the parameters will be considered. If `true.init.val` is set to `2`, α and γ are initialized using two different shared frailty models [3]; $\sigma^2_{v_S}$, $\sigma^2_{v_T}$ and $\sigma_{v_{ST}}$ are set by the user; ζ, θ, β_S and β_T are initialized using a classical joint frailty model, considering individual level random effects [4]. If the joint frailty model is encounters to convergence issues, β_S and β_T are initialized using two shared frailty models. In all other scenarios, if the simplified model does not converge, default values are used. Initial values for the associated parameters of splines are set to `0.5`. Arguments `theta.init`, `sigma.ss.init`, `sigma.tt.init`, `sigma.st.init`, `gamma.init`, `betas.init`, `betat.init`, `alpha.init`, `zeta.init` are vectors of initial values for variances of random effects, regression coefficients, and α and ζ parameters. By default, θ, α and ζ are initialized to `1`, σ^2_S, σ^2_T, γ, $beta_S$, β_T to are initialized to `0.5`, and σ^2_{ST} are initialized to `0.48`.

As other arguments of this function, `scale` is a numeric that makes it possible to re-scale survival times and to avoid numerical problems for some convergence issues. If no change is needed the argument is set to 1, the default value. e.g.: 365 aims to convert days to years ".

Argument `init.kappa` is a vector of two smoothing parameters used to penalized the log likelihood. By default, `init.kappa = NULL` meaning that the values used should be obtained using the maximizing likelihood cross-validation criterion [5]. Argument `kappa.use` is a numeric with values `1`, `3` or `4` that indicate how to manage the smoothing parameters κ_1 and κ_2 in the event of convergence issues. If it is set to `1`, the given smoothing parameters or those obtained by cross-validation are used. If it is set to `3`, the associated smoothing parameters are successively divided by 10, in the event of convergence issues until 5 times. If it is set to `4`, the management of the smoothing parameter is as in case `1`, followed by the successive division as described in case `3` and preceded by the change of the number of nodes for the Gauss-Hermite quadrature. The default is `4`.

When the program requires a random number generator, argument `random. generator` indicates the random number generator to use by the `Fortran` compiler with a value `1` for the intrinsec subroutine `Random_number` and `2` for the subroutine `uniran()`; argument `random` is a binary that indicates wether the random number generator should be reset with a different environment at each call `(1)` or not `(0)`. if `random` is set to `1`, the computer clock is used as seed and it will not be possible to reproduce the generated data (or dataset); argument `random` is required if `random.generator` is set to 1. Argument `random.nb.sim` is a binary that indicates the number of generations that will be made, required if both `random.generator` and `random` are set to 1. The parameter `seed` is the seed to use for data (or samples) generation, required if `random.generator` is set to 1. `seed` must be a positive value. Otherwise, the program will not consider it.

The other arguments for this function are `nb.decimal`, which indicates the number of decimals required for presenting results, `print.times`, a logical that specifies wether the estimation time should be printed or not, and `print.iter`, which is a logical that specifies wether the iteration process should be printed or not. The default values for the last two arguments are `FALSE`.

The `jointSurroPenal` object:. value `EPS` is a vector containing the convergence thresholds obtained with the Marquardt algorithm for the parameters for the log-likelihood and for the gradient. Vector `b` refers to the estimates of the spline parameters, the coefficient ζ (if `indicator.zeta` is set to `1`), the standard error of the shared individual-level frailty ω_{ij} (θ), elements of the lower triangular matrix (L) from the Cholesky decomposition such that $\Sigma = LL^T$, with Σ the covariances of the random effects (v_{S_i}, v_{T_i}), the coefficient α (if `indicator.alpha` is set to `1`), the standard error of the random effect u_i and the regression coefficients β_S and β_T. Value `varH` is the covariance matrix of all parameters in `b` and `varHIH` is the corresponding robust estimation of the covariance matrix of estimates. Value `loglikPenal` represents the penalized log-likelihood. Value `LCV` represents the approximated likelihood cross-validation criterion (see equation .1), with H^{-1} the converged Hessian matrix and `l(.)` the full log-likelihood.

$$LCV = \frac{1}{n}(trace(H_{pl}^{-1}H) - l(.)) \tag{.1}$$

Value `xS` represents a vector of times where both survival and hazard functions are estimated; `survS`, `lamS`, `lamT` and `survT` are four arrays `(dim = 3)` for estimates and confidence bounds of: baseline survival and hazard function for surrogate endpoint, baseline survival and hazard function for true endpoint. `n.iter` is the number of iterations used to reach convergence. The `'jointSurroPenal'` object also contains estimates of distinct parameters of the model labeled `theta` for θ, `gamma` for γ, `alpha` for α `zeta` for ζ, `sigma.s` for σ_S, `sigma.t` for σ_T, `sigma.st` for σ_{ST}, `beta.s` for β_S and `beta.t` for β_T.

Other returned values include: `ktau` and `R2.boot` which represent Kendall's τ and R^2_{trial} with the corresponding 95 % CI computed using the parametric bootstrap; `Coefficients` which is a matrix of the estimates with the corresponding standard errors and the 95 % CI; and `kappa` which represents a vector of positive smoothing parameters used for convergence. These values may be different from the initial values if argument `kappa.use` is set to `3` or `4`; `data`, the dataset used in the model; and `varcov.Sigma`, which is the covariance

matrix of ($\hat{\sigma_S}$,$\hat{\sigma_T}$, $\sigma_{\hat{ST}}$) obtained from the Delta method. The latter is used to predict the treatment effect on the true endpoint as well as in the computation of the surrogate threshold effect (STE).

available functions associated with the `'jointSurroPenal'` *object.* For any object inherited from the class `'jointSurroPenal'`, the following `R` functions are available:

- `jointSurroTKendall()`: used to compute Kendall's τ;
- `summary()`: used to display the results of the surrogacy evaluation, based on model (1);
- `ste()`: used to compute the surrogate threshold effect;
- `loocv()`: used for the leave-one-out cross-validation, in order to assess the accuracy of the prediction using the model (1) ;
- `predict()`: this function allows for new trials to predict the treatment effect on the true endpoint based on the observed treatment effect on the surrogate;
- `plot()`: used to plot the baseline survival and the baseline hazard functions.

By calling `help()` on each of the above functions, one can obtain an exhaustive description of its use.

S2B Appendix. The `jointSurrSimul()` function

The first two arguments refer to the size of the meta-analysis (or the multicenter study) be considered, and include the total number of subjects represented by `n.obs` and the total number of trials (or centers), represented by the `n.trial` parameter. The fixed censorship time should be specified using the argument `cens.adm`. The desired model parameters should be specified in the arguments `theta` for the variance of the shared frailty term at the individual level (θ), `alpha` for the coefficient α, `gamma` for the variance of the shared frailty term at the trial level associated with the baseline hazard (γ), `zeta` for the coefficient ζ, `sigma.s` and `sigma.t` for the variances of the correlated random effects treatment-by-trial interaction (σ_S^2, σ_T^2). The desired level of correlation between v_{S_i} and v_{T_i} is given by the argument `rsqrt`, in such a way that the corresponding coefficient of determination $R^2_{trial} = rsqrt^2$. Arguments `betas` and `betat` represent the fixed treatment effects associated with the surrogate endpoint (β_S) and with the true endpoint (β_T).

Using the `jointSurrSimul()` function, it is possible to generate the dataset by considering homogeneous baseline hazard functions across trials. In this case, the argument `frailt.base` should be set `0`. To consider heterogeneity, this parameter should be set to `(1)`, i.e. the default. The desired parameters for the Weibull hazard function are given by the arguments `lambda.S` as scale parameter and `nu.S` as shape parameter, which are both associated with the Surrogate endpoint; `lambda.T` and `nu.T` as the scale and shape parameters associated with the True endpoint. The argument `ver` indicates the number of covariates to be considered in the joint model. To evaluate surrogacy, we just consider the treatment arm. Argument `typeOf` is used to

indicate the variant of the joint model used for data generation. It is coded `0` for the classical joint model with a shared individual frailty effect [4] and `1` for the joint surrogate model (1). If `typeOf` is set to `0`, arguments `gamma`, `zeta`, `sigma.s`, `sigma.t` and `rsqrt` are not required.

The parameter `equi.subj.trial` is a binary variable that indicates wether the same proportion of subjects should be included per trial (1) or not (0). If set to 0, the proportions of subject per trial are required in the argument `prop.subj.trial` (a vector of `n.trial` elements). Likewise, argument `equi.subj. trt` is a binary variable that indicates if the same proportion of subjects should be randomized per trial (1) or not (0). If set to 0, the proportions of subject per trial are required in the argument `prop.subj.trt`. The argument `full.data` with required values 0 or 1 indicates the structure of the returned dataset.

The other arguments of this function are assigned to the random data generator. Arguments `random.generator` , `random`, `random.nb.sim` and `seed` are detailed in the function `jointSurroPenal()` . Argument `nb.reject.data` specifies the number of generated datasets to be rejected before the considered one. This parameter is required in the event of data generation for simulation studies. With the same values and `random.generator` set to 1, all generated dataset will be the same. However, if the argument `nb.reject.data` varies, different datasets will be obtained during the data generation process.

If the argument `full.data` is set to 0, the function `jointSurrSimul()` returns a `dataframe` with the same columns as those detailed in the function `jointSurroPenal()`. However, if `full.data` is set to 1, additional columns corresponding to the random effects ω_{ij}, u_i, v_{S_i} and v_{T_i} are included in the `dataframe`. Note that in the latter case, the presence of u_i, v_{S_i} and v_{T_i} requires `typeOf` to be equal to `1`.

S2C Appendix. The `jointSurroPenalSimul()` function

Description:. argument `nb.dataset` indicates the number of datasets (or meta-analysis) to be generated. It can be set between 100 and 500. Arguments `nbSubSimul` and `ntrialSimul` specify respectively the number of subjects and the number of trials to be considered in each meta-analysis. Argument `equi.subj.trial` is a binary that indicates whether the same proportion of subjects per trial should be considered in each study design (1), the default or not (0). In the event of different trial sizes, put the proportions of subjects to be considered per trial in the vector `prop.subj.trial`. Argument `equi.subj.trt` indicates whether the same proportion of subjects randomized in the treatment arm should be considered per trial `(1)`, the default or not `(0)`. If `0`, put the proportions of subjects to be considered per trial in `prop.subj.trt`. The sizes of the vectors `prop.subj.trial` and `prop.subj.trt` are equal to the number of trials. Arguments `theta2`, `gamma.ui`, `sigma.s`, `sigma.t`, `betas`, `betat`, `zeta`, `alpha.ui` are simulation values for variances of random effects (θ, γ, σ_S^2, σ_T^2), regression coefficients (β_S, β_T) , and ζ and α parameters. The desired trial level association (R^2_{trial}) is specified using argument `R2`. Users can implement the `R` function `jointSurroKendall()` to set the desired Kendall's τ by varying the values assigned to the arguments α, γ, ζ and θ. Recall that Kendall's τ formulation depends on these variables.

Arguments `lambdas`, `lambdat`, `nus` and `nut` are desired scale and shape parameters for the `Weibull`

distribution associated with the Surrogate endpoint and with the True endpoint. Argument `time.cens` indicates the censorship time. The default for `time.cens` is `549` for about 40% of censored subjects, taking into account the Weibull parameters.

Argument `true.init.val` is a numeric used to manage initial values during estimations, with values `0`, `1` or `2`. If it is set to `(0)` (the default), models should be initialized using arguments `theta.init` for θ, `zeta.init` for ζ, `gamma.init` for γ, `alpha.init` for α, `sigma.ss.init` for $\sigma^2_{v_S}$, `sigma.tt.init` for $\sigma^2_{v_T}$, `sigma.st.init` for $\sigma_{v_{ST}}$, `betas.init` for β_S and `betat.init` for β_T. If `true.init.val` is set to `1`, the real simulation parameters are used as initial values. If set to `2`, α and γ are initialized using two distinct shared frailty models [3]; $\sigma^2_{v_S}$, $\sigma^2_{v_T}$ and $\sigma_{v_{ST}}$ are set by the user using the previous corresponding initial values; ζ, θ, β_S and β_T are initialized using the classical joint frailty model, considering individual level random effects [4]. If the joint frailty model is faced with convergence issues, β_S and β_T are initialized using two shared frailty models. In all other scenarios, default parameters values are used if the simplified model does not converge. Initial values for spline associated parameters are set to `0.5`.

Argument `kappa.use` is a numeric with values `0`, `1`, `2`, `3` or `4`, that indicates how to manage the smoothing parameters κ_1 and κ_2 in the event of convergence issues. If it is set to `0`, the first smoothing parameters that allowed convergence on the first dataset are used for all simulations. If it is set to `1`, smoothing parameters are estimated by cross-validation for each generated dataset. If it is set to `2`, the same process for choosing kappas as in case `1`. However, in the event of a convergence issue, the first smoothing parameters that allowed convergence among the three previous that worked are used. If it is set to `3`, the associated smoothing parameters are successively divided by 10, in the event of convergence issues until 5 times. If it is set to `4` (the default), the management of the smoothing parameters is as in case `2`, preceded by the successive division described in case `3` and by the changing of the number of nodes for the Gauss-Hermite quadrature.

The other arguments are detailed in the function `jointSurroPenal()`. Argument `random.nb.sim` can be set to `nb.dataset`.

The `jointSurroPenalSimul` object:. the function `jointSurroPenalSimul()` returns an object from the class `'jointSurroPenalSimul'` of which all elements can be obtained by displaying the help on this function. Some of the returned values include: `n.iter`, which is the mean number of iterations needed to reach converge; `dataTkendall` and `dataR2boot`, which are two matrices with `nb.dataset` line(s) and three columns of the estimates of Kendall's τ and R^2_{trial} with their confidence intervals; and `dataParamEstim`, which is a dataframe including all estimates with the associated standard errors for all simulations. All cases of non-convergence in previous matrices and data frames are represented by a line of 0.

S2D Appendix. The `jointSurroTKendall()` function

Argument `object` is an object inherited from the `'jointSurroPenal'` class. This argument can be set to `NULL` if users want to base computation on given parameters of Kendall's τ function. In that case, argument

`theta` indicates the variance of the individual-level random effect (ω_{ij}); `gamma`, the variance of the trial-level random effect associated with the baseline risk (u_i); `alpha`, the coefficient associated with u_i; `zeta`, the coefficient associated with ω_{ij}. A argument `sigma.v` is for the covariance matrix of the random effects treatment-by-trial interaction (v_{S_i}, v_{T_i}).

Argument `int.method` is a numeric that indicates the integration method. If `int.method` is set to `0`, the Monte-Carlo integration is used with the number of generated samples specified using argument `nb.MC.kendall`. However, if `int.method` is set to `1`, the Gauss-Hermite quadrature integration is used with the number of nodes specified using argument `nb.gh`. For better or stable results, one should use at least 4000 samples or `20` quadrature nodes. The actual value for nodes used is the maximum between `20` and `nb.gh`.

Argument `ui` is a binary that indicates wether heterogeneity at trial level associated with the baseline risk should be considered (`1`), the default or not (`0`). The other arguments of the function `jointSurroPenal()` are detailed in the function `jointSurroPenal()`.

S2E Appendix. The `plot()` function

In this function, arguments `x` is an object inherited from the `'jointSurroPenal'` class. Argument `endpoint` is a binary that indicates the endpoint to be plotted: `0` for the surrogate endpoint, `1` for the true endpoint, and `2` for both endpoints, the default. Argument `scale` is a numeric that allows survival times to be re-scaled. If no change is needed, the argument is set to 1, the default value. Otherwise, the survival times are multiplied by the scale value. Argument `type.plot` is a character string specifying the type of curve to be plotted. Possible values are "Hazard", or "Survival". It is possible to simply use the first letters "Haz" and "Su". Argument `conf.bands` is a logical that determines whether confidence bands will be plotted. Arguments `xmin` and `xmax` indicate the minimum and the maximum values for the x-axis, and `ylim` the range of the y-axis. The location of the legend in the graph can be specified by setting the argument `pos.legend` to a single keyword from the list `"bottomright"`, `"bottom"`, `"bottomleft"`, `"left"`, `"topleft"`, `"top"`, `"topright"`, `"right"` and `"center"`. Argument `cex.legend` allows the size of the legend to be changed. The last three arguments: `main`, `Xlab` and `Ylab` are labels for the title of the plot, for the x-axis and for the y-axis.

S2F Appendix. The `loocv()` function

Principle and description:. for each trial i, the loocv consists of the following: (a) estimating the parameters of model (1) using the dataset minus subjects in trial i; (b) predicting the treatment effect on the true endpoint based on both the observed treatment effect on the surrogate endpoint in trial i and the estimates from (a); (b) comparing the observed treatment effect on the true endpoint, in trial i with the predicted value. In (a) and (b), the observed treatment effects on the surrogate and the true endpoints are estimated

using two Cox proportional hazard models. The function `loocv()`, which can be applied to an object of class `'jointSurroPenal'`, has three main arguments.

Argument `object` is an object inherited from class `'jointSurroPenal'`; `unusedtrial` is an argument that specifies a list of trial(s) not to be considered in the cross-validation. This argument is useful when excluding some trials, when the model is facing a convergence issue. The last argument is `var.used` and has two values. The first one is `"error.estim"` and indicates whether the prediction variance takes the estimation errors from the estimates of the parameters into account. If estimates are supposed to be known, or if the dataset includes a sufficiently high number of trials with a sufficiently high number of subjects per trial, value `"No.error"` can be used. The default is `error.estim`.

S2G Appendix. Management of convergence issues

Special attention must be paid to initializing model parameters, the choice of the number of knots per spline, the smoothing parameters and the number of quadrature points for solving convergence issues. When numerical or convergence problems are encountered, the model is fitted again using a combination of the following strategies: varying the number of quadrature points, dividing or multiplying κ_1 or κ_2 by 10 or 100 according to their preceding values, or using parameter vectors obtained during the last iteration (with a modification of the number of quadrature points and smoothing parameters). Using this strategy, a lower rejection rate is usually obtained during simulation studies. A sensitivity analysis was conducted without this strategy and similar results were obtained on the converged samples with about 23% rejection rate.

References

[1] D. Rizopoulos, Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive gaussian quadrature rule, Computational Statistics & Data Analysis 56 (3) (2012) 491 – 501. doi:https://doi.org/10.1016/j.csda.2011.09.007.
URL http://www.sciencedirect.com/science/article/pii/S0167947311003264

[2] L. Ferrer, V. Rondeau, J. Dignam, T. Pickles, H. Jacqmin-Gadda, C. Proust-Lima, Joint modelling of longitudinal and multi-state processes: Application to clinical progressions in prostate cancer, Statistics in Medicine 35 (22) (2017) 3933–3948. doi:10.1002/sim.6972.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6972

[3] V. Rondeau, D. Commenges, P. Joly, Maximum penalized likelihood estimation in frailty models, Lifetime Data Analysis 9 (2) (2003) 139–153.

[4] V. Rondeau, S. Mathoulin-Pelissier, H. Jacqmin-Gadda, V. Brouste, P. Soubeyran, Joint frailty models for recurring events and death using maximum penalized likelihood estimation: Application on cancer events, Biostatistics 8 (4) (2007) 708–721. doi:10.1093/biostatistics/kxl043.
URL http://biostatistics.oxfordjournals.org/content/8/4/708.abstract

[5] P. Joly, D. Commenges, L. Letenneur, A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia, Biometrics 54 (1) (1998) 185–194.
URL http://www.jstor.org/stable/2534006

Chapter 5

Modèle conjoint à fragilités et à copules pour la validation en une étape des critères de substitution à temps d'évènements à partir des données issues de méta-analyses d'essais cliniques

---

5.1 Article

Le modèle conjoint à fragilités développé dans le chapitre 3 présente une limite particulière liée aux temps de calculs qui peuvent être très longs lorsque la taille des essais augmente. Ce problème vient de la présence dans la log-vraisemblance marginale des intégrales sur les effets aléatoires au niveau individuel. En effet, ces intégrales ne pouvant pas être exprimées de façon analytique demandent pour chaque essai de répéter la procédure d'intégration numérique autant de fois qu'on a de sujets dans l'étude. De plus, considérer les effets aléatoires au niveau individuel dans le modèle demande d'estimer deux paramètres à savoir la variance des effets aléatoires θ et le paramètre α. Suite à ce constat, nous nous sommes fixé comme objectif dans ce travail de considérer un autre type de modèle conjoint qui combine les effets aléatoires et les copules. Dans ce modèle, au lieu des effets aléatoires au niveau individuel, la dépendance résiduelle entre le critère de substitution (S) et le critère de jugement principal (T) est prise en compte par une fonction de copule. Nous avons considéré, compte tenu de leurs propriétés mathématiques, les fonctions de copules de Clayton et de Gumbel.

Avec le nouveau modèle, nous avons une simplification de la log-vraisemblance marginale

qui non seulement n'inclut plus les intégrales sur les effets aléatoires au niveau individuel, mais également inclut un seul paramètre lié à la dépendance au niveau individuel entre S et T : le paramètre de copule. Indépendamment de la fonction de copule considérée, nous avons défini le τ de Kendall comme dans l'approche standard en deux étapes de Burzykowski et al. (2001), en fonction du paramètre de copule. Par ailleurs, la nouvelle estimation du τ de Kendall peut être interprété de la même façon que celui obtenu dans l'approche standard en deux étapes, contrairement à celui qui a été estimé dans le modèle conjoint à fragilités dont l'interprétation est conditionnelle à ω_{ij}. Un autre objectif de ce travail était de comparer les performances de ces deux modèles. Les paramètres du modèle ont été estimés en maximisant comme précédemment la log-vraisemblance marginale pénalisée, à l'aide de l'algorithme de `Marquardt`.

Indépendamment du modèle utilisé pour la génération des données, les résultats ont montré de meilleures estimations du τ de Kendall, avec le modèle basé sur les copules de Clayton, comparées à celles issues du modèle conjoint à fragilités. De plus, en cas de très forte association au niveau individuel ($\tau > 0.8$), le modèle conjoint à fragilités faisait face à des problèmes de convergence, et avait du mal à estimer le τ de Kendall, en termes de taux de couverture. En revanche, le R^2_{trial} était mieux estimé avec le modèle conjoint à fragilités comparé au modèle de copule de Clayton. A l'issue des analyses de sensibilité, de la variation des méthodes d'intégration et de la variation des caractéristiques des données, nous avons noté une robustesse du modèle conjoint à fragilités et à copule. Pour conclure, nous recommandons le choix de ce modèle en première intention pour la validation des critères de substitution.

Ce travail a fait l'objet d'un article scientifique qui a été soumis pour considération dans *Biometrical journal* (Sofeu et al., 2019). Nous avons parallèlement implémenté le nouveau modèle dans `frailtypack`. La nouvelle version du package incluant ces extensions sera disponible très prochainement sur le CRAN. Toutefois, elle est accessible sur Git Hub.

Biometrical Journal **0** (2019) 0, zzz–zzz / DOI: 10.1002/bimj.200100000

A joint frailty-copula model for meta-analytic validation of failure time surrogate endpoints in clinical trials

Casimir L. Sofeu[*,1], **Takeshi Emura** [2], and **Virginie Rondeau** [1]

[1] INSERM U1219 (Biostatistics team), ISPED, Université Bordeaux Segalen, 146 rue Léo Saignat, 33076 Bordeaux Cedex, France

[2] Graduate Institute of Statistics, National Central University, Jhongda Road, Jhongli City, Taoyuan 32001, Taiwan

Received zzz, revised zzz, accepted zzz

Abstract

In a meta-analysis framework, the classical approach for the validation of time-to-event surrogate endpoint is based on a two-step analysis. This approach often raises estimation issues. Recently, we proposed a one-step validation approach based on a joint frailty model. This approach was quite time-consuming, despite parallel computing, due to individual level frailties used to take into account heterogeneity in the data at the individual level. We now propose an alternative one-step approach for evaluating surrogacy, using a joint frailty-copula model. The model includes two correlated random effects treatment-by-trial interaction and a shared random effect associated with the baseline risks. At the individual level, the joint survivor functions of time-to-event endpoints are linked using copula functions. We chose a non-parametric form of the baseline hazard functions using splines. We estimated parameters and hazard function using a semi-parametric penalized marginal likelihood method, considering various numerical integration methods. Both individual-level and trial-level surrogacy were evaluated using Kendall's Tau and coefficient of determination. The performance of the estimators was evaluated using simulation studies. The model was applied to individual patient data meta-analyses in advanced ovarian cancer to assess progression-free survival as a surrogate for overall survival, as part of the evaluation of new therapy. The model showed good performance and was quite robust regarding the integration methods and data variation, regardless of the surrogacy evaluation criteria. The Clayton Copula model well performed Kendall's Tau compared to the previous model. The novel model reduces the convergence and model estimation issues encountered in the two-step approach.

Key words: Joint frailty-copula model; Meta-analysis of clinical trials; Numerical integration; One-step validation method; Surrogate endpoint.

1 Introduction

The use of the classical two-step copula approach for validating the candidate time-to-event surrogate endpoint seems rather rare, as reported in a recent literature review (Savina et al., 2018; Branchoux et al., 2019). Studies frequently use alternative approaches where the trial level surrogacy (R^2_{trial}) is based on the square of the Spearman and the Pearson correlations of the treatment effects across trials. Other popular association measurements include the coefficient of determination from the weighted linear regression of treatment effects for the true endpoint on treatment effects for the surrogate endpoint across trials, and the unadjusted R^2_{trial} from the two-step copula model (Shi et al., 2011; Savina et al., 2018; Branchoux et al., 2019). Simulations by Shi et al. (2011) showed that all these methods are biased, mainly with moderate sample size as commonly encountered in meta-analyses. Restrictions in the use of the classical two-step copula approach and therefore the adjusted R^2_{trial} may be influenced by convergence issues or estimation errors, or by the unavailability of an individual patient data (IPD) meta-analysis.

A literature review by Matulonis *et al.* Matulonis et al. (2015) discussed inconsistent results in the use of progression-free survival (PFS) as surrogate endpoint for overall survival (OS). An improvement

*Corresponding author: e-mail: casimir.sofeu@u-bordeaux.fr, scl.ledoux@gmail.com, Phone: +33557579568

observed on PFS from clinical trials of targeted agents could not be observed on OS, possibly reflecting the long post-progression survival (PPS) period associated with the disease (Booth and Eisenhauer, 2012; Matulonis et al., 2015). This highlights the need for intermediate endpoints such as the time to second disease progression or death or the time to second subsequent therapy or death, to determine whether the improvement observed on PFS persists beyond the first disease progression and throughout subsequent lines of therapy. Therefore, the search for valid surrogate endpoints remains relevant for accelerating phase III clinical trials.

Recently, Sofeu et al. (2019) proposed a novel approach to validate a putative surrogate endpoint based on a joint frailty model. This approach was based on a one-step analysis strategy where trial level surrogacy was no longer needed to be adjusted on estimation errors, as in the classical Burzykowski et al. (2001) approach. The authors used a shared frailty term to take into account heterogeneity in the data at the individual level and thus, to assess the validity of surrogacy at the individual level. This led to estimating two parameters (α and θ) for heterogeneity at individual level and to using numerical integration to integrate over the individual level random effect in the marginal log-likelihood. The complexity of the model was quite time-consuming, despite parallel computing.

Moreover, from the application we observed that compared to the existing approach, the association measurement at the individual level was the smallest when estimates were based on the new joint surrogate model. Therefore, we propose an alternative approach for the validation of surrogate endpoints, based on an extension of the joint-frailty copula model of Emura et al. (2017). The new model includes two correlated random effects treatment-by-trial interaction and a shared random effect at trial level accounting for heterogeneity on baseline risks. At the individual level, the joint survivor function of failure time endpoints are linked using copula functions. Therefore, instead of assuming independence between endpoints conditional on a shared frailty term, we consider dependence using copula models (Prenen et al., 2017). Different copula functions are considered, and the individual level surrogacy is assessed using an intuitive formulation of Kendall's τ, based on the copula parameter, as in the classical two-step copula approach.

The article is organized as follows. In section 2 we recall the formulation of the joint surrogate model. In section 3, we define the joint frailty-copula model for surrogacy evaluation. The full penalized log-likelihood construction and the estimation methods are also described in this section. The surrogacy evaluation criteria are defined in Section 4. In Section 5 we present the simulation studies and in section 6 an application to individual patient data meta-analyses in advanced ovarian cancer is proposed. Finally, in section 7 we present the concluding discussion.

2 Background : the joint surrogate model

Assume S_{ij} and T_{ij} the failure times associated respectively with the surrogate and the true endpoints, for subject j ($j = 1, \cdots, n_i$) belonging to the trial i ($i = 1, \cdots, G$). Let $\mathbf{Z}_{S,ij} = (Z_{Sij1}, \cdots, Z_{Sijp})^{'}$ and $\mathbf{Z}_{T,ij} = (Z_{Tij1}, \cdots, Z_{Tijp})^{'}$ be covariates associated with S_{ij} and T_{ij}, respectively. The joint surrogate model can be defined according to Sofeu et al. (2019) as:

$$\begin{cases} \lambda_{S,ij}(t|\omega_{ij}, u_i, v_{Si}, \mathbf{Z}_{S,ij}) = \lambda_{0S}(t)\exp(\omega_{ij} + u_i + v_{Si}Z_{ij1} + \boldsymbol{\beta}_S\mathbf{Z}_{S,ij}) \\ \lambda_{T,ij}(t|\omega_{ij}, u_i, v_{Ti}, \mathbf{Z}_{T,ij}) = \lambda_{0T}(t)\exp(\zeta\omega_{ij} + \alpha u_i + v_{Ti}Z_{ij1} + \boldsymbol{\beta}_T\mathbf{Z}_{T,ij})) \end{cases} \tag{1}$$

with,

$$\omega_{ij} \sim N(0, \theta), \quad u_i \sim N(0, \gamma) \tag{2}$$

and

$$\begin{pmatrix} v_{Si} \\ v_{Ti} \end{pmatrix} \sim MVN\left(\mathbf{0}, \Sigma_v\right), \Sigma_v = \begin{pmatrix} \sigma^2_{v_S} & \sigma_{v_{ST}} \\ \sigma_{v_{ST}} & \sigma^2_{v_T} \end{pmatrix}. \tag{3}$$

In model (1), $\lambda_{S,ij}(.)$, $\lambda_{0S}(t)$ and $\boldsymbol{\beta}_S = (\beta_{S1}, \cdots, \beta_{Sp})$ are respectively the hazard function of failure-time S_{ij} for the j^{th} patient in trial i, the baseline hazard function and the fixed effects (or log-hazard ratio) corresponding to the covariates $\mathbf{Z}_{S,ij}$ associated with the surrogate endpoint; $\lambda_{T,ij}(.)$, $\lambda_{0T}(t)$ and $\boldsymbol{\beta}_T = (\beta_{T1}, \cdots, \beta_{Tp})$ are defined as above and are associated with the true endpoint. ω_{ij} is a shared individual level frailty that takes into account heterogeneity at the individual level; ω_{ij} serves to assess the correlation between the surrogate and the true endpoints at the individual level; $\exp(u_i)$ is a shared frailty effect associated with the baseline hazard function that takes into account the heterogeneity between trials of the baseline hazard function, associated with the fact that there are several trials in the meta-analytical design. The power parameters ζ and α distinguish both individual and trial level heterogeneities between the surrogate and the true endpoint. v_{Si} and v_{Ti} are two correlated random effects treatment-by-trial interactions (trial level frailties in interaction with the treatment) which answer the following question: does the effect of treatment on the surrogate endpoint reliably predict the effect of treatment on the true endpoint in each trial? Z_{ij1} represents the treatment arm to which the patient has been randomized. In this model, ω_{ij}, u_i and (v_{Si}, v_{Ti}) are mutually independent.

3 Methods

3.1 The joint frailty-copula model for surrogacy

For trial i, we let $\mathbf{v}_i = (u_i, v_{Si}, v_{Ti})$ be the vector of trial level random effects from models (2) and (3). We propose to define the new approach of validation based on an extension of a joint frailty-copula model (Emura et al., 2017).

Model construction

We use the bivariate copula function to define the joint survivor functions for S_{ij} and T_{ij} as follows:

$$\begin{aligned}\bar{F}(s_{ij}, t_{ij}|\mathbf{Z}_{S,ij}, \mathbf{Z}_{T,ij}, \mathbf{v}_i) &= P(S_{ij} > s_{ij}, T_{ij} > t_{ij}|\mathbf{Z}_{S,ij}, \mathbf{Z}_{T,ij}, \mathbf{v}_i) \\ &= \varphi_\theta[\varphi_\theta^{-1}(\bar{F}(s_{ij}|\mathbf{Z}_{S,ij}, u_i, v_{Si})) + \varphi_\theta^{-1}(\bar{F}(t_{ij}|\mathbf{Z}_{T,ij}, u_i, v_{Ti}))] \\ &= C_\theta\{\bar{F}(s_{ij}|\mathbf{Z}_{S,ij}, u_i, v_{Si}), \bar{F}(t_{ij}|\mathbf{Z}_{T,ij}, u_i, v_{Ti})\} \qquad (4)\end{aligned}$$

where $\bar{F}(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki}) = P(K_{ij} > k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki})$, $K \in \{S, T\}$, are the survival functions for lifetime T_{kij} given $\mathbf{Z}_{K,ij}$. The generator $\varphi_\theta : [0, \infty) \to [0, 1]$ of a parametric Archimedean copula family is a continuous strictly decreasing function with $\varphi_\theta(0) = 1$ and $\varphi_\theta(\infty) = 0$. We denote by φ_θ^{-1} the inverse function of φ_θ. For (4) to be a proper survival function, we assume that the generator is monotone. This means that all the derivatives exist and have alternating signs $(-1)^m \frac{d^m}{dt^m}\varphi_\theta(t) \geqslant 0$, $\forall t \geqslant 0$ and $m = 1, 2, \cdots$, as shown in Nelsen (2006). The generator φ_θ is the Laplace transformation of a distribution function $G_\theta(x)$ with $G_\theta(0) = 1$ (Joe, 1997; Prenen et al., 2017). C_θ is called an Archimedian copula function (Nelsen, 2006). The conditional survival functions associated with the endpoints are defined as follows:

$$\bar{F}_{S,ij}(s_{ij}|\mathbf{Z}_{S,ij}, u_i, v_{Si}) = \exp\left\{ -\int_0^{s_{ij}} \lambda_{0S}(x) \exp\left(u_i + v_{Si}Z_{ij1} + \boldsymbol{\beta}_S\mathbf{Z}_{S,ij}\right)dx\right\} \qquad (5)$$

$$\bar{F}_{T,ij}(t_{ij}|\mathbf{Z}_{T,ij}, u_i, v_{Ti}) = \exp\left\{ -\int_0^{t_{ij}} \lambda_{0T}(x) \exp\left(\alpha u_i + v_{Ti}Z_{ij1} + \boldsymbol{\beta}_T\mathbf{Z}_{T,ij}\right)dx\right\}. \qquad (6)$$

If $v_{Si} = v_{Ti} = 0$ (i.e $\sigma^2_{v_S} = \sigma^2_{v_T} = 0$), the model reduces to the joint frailty-copula (Emura et al., 2017) model that can only measure the individual level surrogacy through θ. In (5 - 6), $\lambda_{0K}(x)$ and $\boldsymbol{\beta}_k$ are respectively the baseline hazard function and the fixed effects (or log-hazard ratio) corresponding to the covariates $\mathbf{Z}_{K,ij}$ associated with the failure time endpoint, $K \in \{S, T\}$. u_i, α and v_{Ki} are defined as in section 2.

For simplicity, we focus on the Clayton and Gumbel-Hougaard copula functions. However, this model can be extended to other copula functions.

Formulation of Clayton and Gumbel-Hougaard copula models

In Clayton's model, the copula function has the form

$$C_\theta(a,b) = (a^{-\theta} + b^{-\theta} - 1)^{\frac{-1}{\theta}}, \quad \varphi_\theta(s) = (1+\theta s)^{-1/\theta}, \quad \theta > 0;$$

and in Gumbel's model, the copula function has the form

$$C_\theta(a,b) = \exp\left[-\left\{ (-\log a)^{\theta+1} + (-\log b)^{\theta+1} \right\}^{\frac{1}{\theta+1}} \right], \quad \varphi_\theta(s) = \exp\left[-s^{1/(1+\theta)}\right], \quad \theta \geq 0.$$

These formulations induce a positive association among endpoints. The strength of the association increases with increasing θ and reaches independence when $\theta \to 0$. If $v_{Si} = v_{Ti} = 0$ and the copula parameter $\theta = 0$, the models (4, 5 and 6) reduce to the joint frailty model of Rondeau *et al.* (Rondeau et al., 2015). Note that the Clayton copula produces "lower tail dependence", leading to strong dependence between large values of S_{ij} and T_{ij}. On the other hand, the Gumbel Copula produces "upper tail dependence", leading to strong dependence between small values of S_{ij} and T_{ij}. Hence, the two copulas capture different types of dependence structure between S_{ij} and T_{ij}. In practice, comparison of model performance is desirable when choosing the copula function that best fits the data.

Remarks: The parameters in the proposed models (4, 5 and 6) have different interpretation from those from the joint surrogate model (1). For instance, β_S in the model (5) is the covariate effects given (u_i, v_{S_i}) while β_S is the covariate effects given (ω_{ij}, u_i, v_{S_i}) in model (1). Hence, the latter is the covariate effects after accounting for the individual-level heterogeneity. Furthermore, integrating out ω_{ij} from the model (1) does not give the models of the forms (5-6). Therefore, β_S and all other parameters in the two models are not comparable since they refer different population quantities. The differences between frailty and copula models have been discussed elsewhere (see Section 3.3.4 of Duchateau and Janssen (2008)).

3.2 Inference in joint frailty-copula model for surrogate (S) and true (T) endpoints.

Log-likelihood construction

Let a pair of endpoints measured on subject j ($j = 1, \cdots, n_i$) in trial i ($i = 1, \cdots, G$) be the survival times associated with the surrogate (S_{ij}) and the true endpoints (T_{ij}), and C_{ij} be an independent and uninformative censoring time for subject j in trial i, where n_i is the sample size of trial i and G is the total number of trials. The observed surrogate endpoint event time T_{Sij} = min(S_{ij}, T_{ij}, C_{ij}) and the true endpoint event time T_{Tij} = min(T_{ij}, C_{ij}). Similarly, $\delta_{S,ij}$ = $\mathbb{1}(T_{Sij} = S_{ij})$ and $\delta_{T,ij}$= $\mathbb{1}(T_{Tij} = T_{ij})$ denote the progression and death indicators, where $\mathbb{1}(.)$ is the indicator function. The data consist of (T_{Tij}, T_{Sij}, $\delta_{S,ij}$, $\delta_{T,ij}$) for $i = 1 \cdots G$ and $j = 1 \cdots n_i$.

Let $d_{ij} = \sum_{K \in \{S,T\}} \delta_{k,ij}$ and $\Phi = (\theta, \sigma^2_{v_S}, \sigma^2_{v_T}, \sigma_{v_{ST}}, \gamma, \lambda_{0T}(.), \lambda_{0S}(.), \boldsymbol{\beta}_S, \boldsymbol{\beta}_T)$ be the number of uncensored event times in subject j and the vector containing all the unknown parameters. Let $\lambda_K(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki}) = \lambda_{0k}(x)\exp[\alpha^{1-\mathbb{1}_{\{k=s\}}}u_i + v_{Ki}Z_{ij1} + \boldsymbol{\beta}_k\mathbf{Z}_{K,ij}]$ be the hazards functions for the surrogate and the true endpoints, and $\Lambda_K(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki}) = \int_0^{k_{ij}} \lambda_K(x|\mathbf{Z}_{K,ij}, u_i, v_{Ki})dx$ the corresponding cumulative hazards functions. Under the proposed joint frailty-copula model (4, 5 and 6), The marginal log-likelihood can be reduced to

$$\begin{aligned} \ell(\Phi) \quad = \quad & \sum_{i=1}^{G} \log \left\{ \int_{\mathbf{v}_i} \left(\prod_{j=1}^{n_i} \varphi_\theta^{(d_{ij})} \left[\sum_{K \in \{S,T\}} \varphi_\theta^{-1}(\bar{F}(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki})) \right] \times \right. \right. \\ & \left. \left. \prod_{K \in \{S,T\}} \left[\frac{f(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki})}{\varphi'_\theta(\varphi_\theta^{-1}(\bar{F}(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki})))} \right]^{\delta_{k,ij}} \right) f_{\mathbf{V}}(\mathbf{v}_i) d\mathbf{v}_i \right\}, \end{aligned} \tag{7}$$

where $\varphi^{[l]}(x) = d^{[l]}\varphi(x)/dx^{[l]}$. The probability density function for the considered Gaussian random effects in cluster i is given by:

$$f_{\mathbf{V}}(\mathbf{v}_i) = \frac{1}{(2\pi)\sqrt{2\pi\gamma|\Sigma_v|}} \exp\left[-\frac{1}{2}(v_{Si}, v_{Ti})\Sigma_v{}^{-1}(v_{Si}, v_{Ti})' - \frac{1}{2}\frac{u_i^2}{\gamma}\right];$$

the conditional survival functions are

$$\bar{F}(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki}) = \exp\left\{ - \Lambda_K(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki})\right\};$$

and the conditional density functions are

$$f(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki}) \quad = \quad \lambda_K(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki}) \exp\left\{ - \Lambda_K(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki})\right\}.$$

The properties of the Archimedian generators are :
for the Clayton copula :

$$\varphi_\theta^{-1}(s) = \tfrac{s^{-\theta}-1}{\theta}, \quad \varphi_\theta^{'}(s) = -(1+\theta s)^{-(1+\theta)/\theta}, \quad \varphi_\theta^{''}(s) = (1+\theta)(1+\theta s)^{-(1+2\theta)/\theta},$$
$$\varphi_\theta^{'}[\varphi_\theta^{-1}(s)] = -s^{(1+\theta)} \quad \varphi_\theta^{'}[\varphi_\theta^{-1}(s) + \varphi_\theta^{-1}(t)] = -(s^{-\theta} + t^{-\theta} - 1)^{-(1+\theta)/\theta}$$

and for the Gumbel-Hougaard copula,

$$\varphi_\theta^{-1}(s) = [-\log(s)]^{(1+\theta)}, \quad \varphi_\theta^{'}(s) = -\tfrac{1}{(1+\theta)} s^{-\theta(1+\theta)} \exp\left[-s^{1/(1+\theta)}\right],$$
$$\varphi_\theta^{''}(s) = \tfrac{1}{(1+\theta)^2}\left[\theta s^{-(2\theta+1)/(\theta+1)} + s^{-2\theta/(\theta+1)}\right] \exp\left[-s^{1/(1+\theta)}\right], \quad \varphi_\theta^{'}[\varphi_\theta^{-1}(s)] =$$
$$-\tfrac{s}{(1+\theta)}\left[-\log(s)\right]^{-\theta}, \quad \varphi_\theta^{'}[\varphi_\theta^{-1}(s) + \varphi_\theta^{-1}(t)] = -\tfrac{1}{1+\theta}\left[(-\log s)^{1+\theta} + (-\log t)^{1+\theta}\right]^{-\theta/(1+\theta)} \times C_\theta(s,t)$$

The derivation of the previous expressions can be found in the Appendix A.1. and A.2.

Estimation methods

To compute the integrals over the random effects present in the likelihood (7), we considered different numerical integration strategies, given that it was quite difficult to obtain the analytical form of these integrals. Two of these methods were based on the non-adaptive Gaussian-Hermite (GH) and the pseudo-adaptive Gaussian-Hermite (PGH) quadrature (Rizopoulos, 2012). Other methods included the Monte-Carlo integration (MC) and the Laplace approximation with a second-order Taylor series expansion (Abrahantes and Burzykowski, 2005).

The baseline hazard functions $\lambda_{0S}(.)$ and $\lambda_{0T}(.)$ were approximated using cubic M-splines, which are a variant of cubic B-splines, and I-splines which are integrated M-splines (Ramsay, 1988). M-splines are non-negative and easy to integrate or differentiate. As we used a cubic spline (or of order 4), the second derivative of λ_{0S} and λ_{0T} was approximated by a linear combination of piecewise polynomial approximation of order 2. This approximation allows the hazard functions to have flexible shapes while reducing the number of parameters. If we denote $\tilde{\lambda}_{0S}$ as an approximation to the maximum penalized likelihood estimators (MPnLEs) $\hat{\lambda}_{0S}$, the approximation error can be made as small as required by increasing the number of knots, as shown by Rondeau et al. (2003). Therefore, to obtain a good estimation of the theoretical hazard function, it is necessary to use as many knots as possible to obtain a MPnlE close to the true hazard function. In our approach, although there are two different hazard functions (for surrogate and true endpoints), we use the same basis of splines for each function. However, the spline coefficients depend on the functions considered.

Semi-parametric penalized likelihood approach

The model parameters ($\sigma^2_{v_S}$, $\sigma^2_{v_T}$, $\sigma_{v_{ST}}$, θ, γ, α, $\boldsymbol{\beta}_S$, $\boldsymbol{\beta}_T$) and the hazard functions ($\lambda_{0S}(.)$ and $\lambda_{0T}(.)$) were subsequently estimated by maximizing the penalized log-likelihood. We penalized the log-likelihood by a term which has large values for rough functions (O'Sullivan, 1988; Joly et al., 1998). The penalized log-likelihood is defined as:

$$pl(\Phi) = l(\Phi) - \kappa_1 \int_0^\infty \lambda''^2_{0S}(t)dt - \kappa_2 \int_0^\infty \lambda''^2_{0T}(t)dt, \tag{8}$$

where $l(\Phi)$ is the full log-likelihood defined in (7), κ_1 and κ_2 the positive smoothing parameters which control the trade-off between the data fit and the smoothness of the functions. Maximization of (8) defines the MPnLEs $\hat{\sigma}^2_{v_S}$, $\hat{\sigma}^2_{v_T}$, $\hat{\sigma}_{v_{ST}}$, $\hat{\theta}$, $\hat{\gamma}$, $\hat{\alpha}$, $\hat{\boldsymbol{\beta}}_S$, $\hat{\boldsymbol{\beta}}_T$, $\hat{\lambda}_{0S}(t)$ and $\hat{\lambda}_{0T}(t)$. We directly use $-H^{-1}$ as a variance estimator, where H is the converged hessian matrix for the penalized log-likelihood.

The smoothing parameters were chosen by maximizing the likelihood cross-validation score, ignoring the explanatory variables as described in Joly et al. (1999). Two separate marginal models with no covariates were used to obtain κ_1 and κ_2.

Computational procedure

We implemented the proposed joint frailty-copula model in an `R` package using a couple of `R` and `Fortran` programs, with parallel computing using `OpenMP`, in order to reduce the running time. Results in this article are based on `R` version `3.5.0` and `gcc` compiler version `4.8.5` for simulation (`R_3.4.0` and `gcc_4.4.7` for the case study) . The estimated parameters were obtained by the robust Marquardt algorithm (Marquardt, 1963), which is a compromise between the steepest descent and the Newton-Raphson algorithms. This variant is more stable than the Newton-Raphson algorithm while preserving its fast convergence property near the maximum. We imposed a positivity constraint on the variance parameter, the spline coefficients and Gumbel's copula parameter. For Clayton's copula parameter, we performed an exponential transform to ensure a strictly positive value for θ. Following this change of variable, the standard errors of variance-covariance parameters were computed using the delta method (Dowd et al., 2014).

4 Criteria to evaluate surrogate endpoints

4.1 individual level surrogacy

In the context of two failure time endpoints, Burzykowski et al. (2001) suggested basing validation of the candidate surrogate endpoint at the individual level on Kendall's τ. Kendall's τ is the difference between the probability of concordance and the probability of discordance of two independent realizations of a joint distribution. It can be shown that (Genest and MacKay, 1986) Kendall's τ is solely expressed as a function of C_θ through:

$$\tau = 4 \int_0^1 \int_0^1 C_\theta(a,b) \frac{\partial^2}{\partial a \partial b} C_\theta(a,b) da db - 1.$$

For the Archimedean copula as is the case in this paper, τ can be expressed as a function of the generator function φ_θ (with $\varphi'_\theta = d\varphi_\theta(t)/dt$) by:

$$\tau = 4 \int_0^1 \frac{\varphi_\theta(t)}{\varphi'_\theta(t)} dt + 1. \tag{9}$$

Therefore, from (9), it can be shown that $\tau = \theta/(\theta+2)$ for the Clayton copula and $\tau = \theta/(\theta+1)$ for the Gumbel copula models. These formulations imply that τ is a strictly increasing function of θ, which reaches 1 in the event of perfect individual level correlation between the endpoints.

4.2 trial level surrogacy

As suggested by Buyse et al. (2000), we based validation at the trial level on the coefficient of determination R^2_{trial} obtained from the estimates of the variance covariance matrix of the random effects treatment-by-trial interaction. Thus, $R^2_{trial} = \sigma^2_{v_{ST}}/(\sigma^2_{v_S}\sigma^2_{v_T})$ and suggest a perfect trial-level association between the surrogate and the true endpoints for a value close to 1.

The standards errors of Kendall's τ and R^2_{trial} were calculated using the delta method (Dowd et al., 2014). The delta method can lead to confidence limits violating the [0,1]. However, using other methods would not change the conclusions of the article. A surrogate endpoint may be considered as valid if R^2_{trial} and Kendall's τ are sufficiently close to 1.

5 Simulations

5.1 Simulation design

A simulation study was performed to evaluate the performance of the proposed model. Let N = $\sum_{j=1}^{G} n_i$ be the sample size, with n_i the size of trial i. We considered two sample sizes with a variable number of subjects ($N = 600$ and 1000 resp·) and a variable number of trials ($G = 10$ and 30 resp·). Five hundred simulated datasets were used in each simulation design and the proposed model (1) was used, considering one covariate. The data was generated with the following algorithm:

1. For each trial i, we generated respectively a Gaussian random variable $u_i \sim N(0, \gamma)$, γ = 0.8, and a couple of Gaussian random variables $(v_{S_i}, v_{T_i}) \sim MVN(\mathbf{0}, \Sigma_v)$, with Σ_v defined as in (3), where $\sigma^2_{v_S}$=$\sigma^2_{v_T}$=0.7 and $\sigma_{v_{ST}} = 0.42$ or 0.63 for low and a high correlations.

2. We generated the binary treatment variable Z_{ij1} and another bivariate covariate Z_{ij2} from a Bernoulli distribution with P(Z=1) = 0.5.

3. A fixed right-censoring variable was considered with $C_{ij} = 349$ and $C_{ij} = 100$ for respectively $\sim$ 44% and 69% censorship.

4. We generated event times S_{ij} and T_{ij} associated with the surrogate and the true endpoints following a Weibull distribution, with a method close to that used in (Rotolo et al., 2018; Nelsen, 2006) under the Clayton copula function. For the Weibull distribution, we specify the baseline hazard functions by

$$\lambda_{0S}(t) = \rho_S \gamma_S t^{\gamma_S - 1}, \quad \rho_S > 0, \quad \gamma_S > 0; \quad \lambda_{0T}(t) = \rho_T \gamma_T t^{\gamma_T - 1}, \quad \rho_T > 0, \quad \gamma_T > 0,$$

where ρ_S and ρ_T are the scale parameters, and γ_S and γ_S are the shape parameters. Therefore, S^*_{ij} are simulated conditionally on the random effects generated before:

$$S_{ij} = \left[- \frac{1}{\rho_S \exp(u_i + v_{Si} Z_{ij1} + \beta_S Z_{ij1})} log(1 - U_{S,ij}) \right]^{1/\gamma_S},$$

with

$$U_{S,ij} \sim U(0, 1);$$

and T_{ij} are generated conditionally on the random effects generated before and on the value of S_{ij}:

$$T_{ij}|S_{ij} = \left[\frac{1}{\theta \rho_T \exp(\alpha u_i + v_{Ti} Z_{ij1} + \beta_T Z_{ij1})} \log \left\{ 1 - W_{ij} + W_{ij}(1 - U_{T,ij})^{-\frac{\theta}{1+\theta}} \right\} \right]^{1/\gamma_T},$$

with

$$W_{ij} = (1 - U_{S,ij})^{-\theta}, \quad U_{T,ij} \sim U(0, 1)$$

The copula parameter θ were fixed according to the desired value of Kendall's τ, given that $\tau = \frac{\theta}{\theta+2}$ for the Clayton copula model. The scale and shape parameters were respectively set to γ_T = 1.1 and ρ_T = 0.0025 for the true endpoint; γ_S = 1.3 and ρ_S = 0.0025 for the surrogate endpoint. The corresponding observed death times were $T_{Tij} = min(T_{ij}, C_{ij})$ and $\delta_{T,ij} = 1$ if $T_{Tij} = T_{ij}$. In addition, the observed progression times were $T_{Sij} = min(S_{ij}, T_{Tij})$ and $\delta_{S,ij} = 1$ if $T_{Sij} = S_{ij}$. When progression and death occurred the same days, we only considered the death event.

For all simulation designs, the fixed-treatment effects β_S and β_T were set to -1.25. Furthermore, we performed a simulation design by considering two covariates, with the fixed effects for the additional variable set to 0.5 for the surrogate and the true endpoints. To investigate the effect of the degree of correlation both at individual and trial levels on the performance of the estimators, we considered a strong and weak correlation by varying R^2_{trial} (between 0.36 and 0.81) and Kendall's τ (between 0.33 and 0.60). We eliminated the cases where convergences or numerical problems occurred in estimating parameters.

We extended our simulation study to investigate the impact of model misspecification on the surrogacy evaluation criteria by generating the new trials of data from the joint frailty model (Sofeu et al., 2019). We thereafter considered estimation based both on the Clayton Copula, the Gumbel copula and the joint surrogate model.

5.2 Evaluation criteria

Convergences were considered when the difference between two consecutive log likelihoods was small ($< \epsilon_b$), the estimated coefficients were stable (consecutive values ($< \epsilon_a$) and the gradient was small enough ($< \epsilon_d$). The default values were $\epsilon_a = \epsilon_b = \epsilon_d = 10^{-3}$. We considered a maximum number of 35 iterations. However, the results are mostly unchanged even if we increase the maximum number of iterations. We were mainly interested in the trial level and the individual level surrogacy. For each parameter, we report the mean, the empirical standard errors (SD), i.e. the standard error of estimates, the mean of the estimated standard errors (SE) and the coverage percentage of the 95% confidence interval estimates (CP in %). To compare the results according to the estimation method used, we also considered the absolute bias and the mean square errors (MSE).

5.3 Results

The simulation results are organised in four subsections. We first assess the performance of the proposed model in terms of bias and the CP. The integration methods are also compared in this subsection. Next, in the second subsection, we assess the robustness of the new model by varying the parameters of the data. In the third subsection, we describe the results from the sensitivity analysis by varying the parameters of the model. Finally in the last subsection, we present the comparative results between the proposed model (4) and the joint surrogate model (1), as well as the results from the study of the misspecification on the proposed model to the data.

Integration methods and joint frailty-copula model: assessment

The reference simulation design includes the following characteristics: N = 600 subjects in G = 30 trials, 44 % of censorship, high individual level and trial-level associations (Kendall's τ = 0.6 and R^2_{trial} = 0.81), estimation based on a Monte-Carlo integration (MC) with 1000 samples, 8 spline knots for the approximation of the baseline risks, model estimation and data generation using the Clayton copula model, estimation using the known values for the parameters as initial values, and the update of the smoothing parameters in the event of convergence issues.

Based on the results in Table 1, the model had good performance in terms of biases and the CP for the copula parameter (θ), the power parameter α, the fixed treatment effects β_S and β_T, and the individual-level association τ, with absolute bias $\leq 3.5\%$ and CP $\geq 91\%$. The estimation of the variance parameters

Table 1 Estimates (Mean), mean of standard errors (SE), empirical standard errors (SD) and percentage of coverage (CP). Estimation based on a Monte-Carlo (MC), Laplace, non-adaptive Gaussian Hermite quadrature (GH) and pseudo-adaptive Gaussian Hermite quadrature (PGH) integration methods†, for M = 500 samples, N = 600 subjects and G = 30 trials

Parameters	True	Mean	SD	SE	CP	Mean	SD	SE	CP
		MC				Laplace			
θ	3	3.034	0.298	0.303	96	3.005	0.293	0.304	97
γ	0.8	0.814	0.294	0.198	79	0.777	0.236	0.234	89
α	1	1.002	0.059	0.053	93	1.001	0.057	0.055	94
σ_{v_S}	0.7	0.722	0.307	0.239	84	0.649	0.253	0.246	85
σ_{v_T}	0.7	0.78	0.389	0.268	82	0.655	0.302	0.269	84
$\sigma_{v_{ST}}$	0.63	0.677	0.318	0.229	83	0.59	0.257	0.235	85
β_S	-1.25	-1.27	0.206	0.175	91	-1.276	0.188	0.179	93
β_T	-1.25	-1.261	0.209	0.187	92	-1.263	0.191	0.19	95
R^2_{trial}	0.81	0.822	0.116	0.095	80	0.825	0.121	0.114	82
τ	0.6	0.601	0.023	0.024	95	0.599	0.023	0.024	97
CI††:n(%)	-	2(0)	-	-	-	63(13)	-	-	-
		GH				PGH			
θ	3	3.045	0.305	0.303	96	3.034	0.313	0.303	95
γ	0.8	0.788	0.235	0.235	91	0.766	0.246	0.23	88
α	1	0.99	0.196	0.064	91	1.008	0.119	0.058	95
σ_{v_S}	0.7	0.702	0.266	0.26	90	0.666	0.258	0.251	86
σ_{v_T}	0.7	0.765	0.381	0.298	87	0.674	0.306	0.275	87
$\sigma_{v_{ST}}$	0.63	0.645	0.283	0.255	89	0.608	0.262	0.241	86
β_S	-1.25	-1.277	0.201	0.175	90	-1.267	0.194	0.178	93
β_T	-1.25	-1.296	0.238	0.181	89	-1.256	0.193	0.189	94
R^2_{trial}	0.81	0.781	0.122	0.123	94	0.829	0.12	0.108	82
τ	0.6	0.602	0.024	0.024	95	0.601	0.024	0.024	95
CI††:n(%)	-	205(42)	-	-	-	74(15)	-	-	-
		MC with 2 covariates							
θ	3	3.057	0.318	0.322	96				
γ	0.8	0.825	0.324	0.2	80				
α	1	1.006	0.063	0.058	92				
σ_{v_S}	0.7	0.714	0.295	0.241	87				
σ_{v_T}	0.7	0.774	0.364	0.277	84				
$\sigma_{v_{ST}}$	0.63	0.669	0.286	0.233	87				
β_{1S}	-1.25	-1.247	0.194	0.179	93				
β_{2S}	-0.5	-0.505	0.089	0.09	96				
β_{1T}	-1.25	-1.244	0.211	0.195	94				
β_{2T}	-0.5	-0.5	0.098	0.098	96				
R^2_{trial}	0.81	0.829	0.13	0.109	78				
τ	0.6	0.603	0.025	0.025	95				
CI††:n(%)	-	44(9)	-	-	-				

†update of smoothing parameters in event of convergence issues; ††CI = Convergence issues; Samples for MC = 1000; quadrature points (GH) = 20; quadrature points(PGH) = 12 points; spline knots = 8; true initial values used, generation and estimation using Clayton copula model.

of the random effects at trial level showed a slight drop in performance with CPs around 80%, compared to an expected 95 %. Consequently, we observed a good estimation of R^2_{trial}, but with only 80% of coverage percentage. In the simulation design considered, we only had two meta-analyses on which the model did not converged. Therefore, in terms of the surrogacy evaluation criteria, the proposed model behaved well.

In Table 1, we compare the simulation results with different integration methods. Using the Gaussian Hermite quadrature (GH) integration, we obtained better estimates of the variance parameters of the random effects in terms of bias, estimation of standard errors (SE) and the CP, compared to the other integration methods. Therefore, the CP for R^2_{trial} was around 94%, as expected. However, using the GH integration, we really faced convergence issues, with 42% of non-convergence cases. Unlike the MC and GH integration methods, Laplace approximation and the pseudo-adaptive Gaussian Hermite quadrature (PGH) tended to underestimate the variance parameters of the random effects treatment-by-trial interaction. Regarding the surrogacy evaluation criteria, we found comparable results with MC, Laplace approximation and PGH in terms of bias, CP and the standard errors for these parameters.

The last part of Table 1 shows a simulation design including two covariates to test the ability of the proposed model to consider more than one exploratory variable. We observed good estimation of the fixed effects in terms of bias and SE, with CPs around 95%. Therefore, by adjusting the model on potential confounders, it can be used to better estimate R^2_{trial} and Kendall's τ.

Variation in data characteristics

Table 2, shows the robustness of the new model. We considered different characteristics of the data, including six designs: increase in the number of trials, with the same number of subjects per trial (G = 50, n_i = 20), very low number of trials (G = 10), variation in the proportion of subjects per trial, reduced censorship time (Censorship = 100, for 69% of censoring rate), weak trial-level association (R^2_{trial} = 0.36), and weak individual level and trial-level association (Kendall's τ = 0.33 and R^2_{trial} = 0.36).

Regardless of the simulation design, the estimation of Kendall's τ, their SEs and CPs were comparable. In the event of a high trial-level association, we observed better estimation of R^2_{trial} both by increasing the number of trials, and varying the number of subjects per trial (bias = 0.004 and 0.003 respectively). With censorship = 100, we made the same observation as previously, with the bias on R^2_{trial} equal to 0.004. However, the coverage probability of R^2_{trial} was reduced to 72% due to bias and an increase in SE (SE = 0.22 compared to 0.096 when G = 30). Considering an extreme case with a very small number of trials (G = 10), R^2_{trial} was slightly overestimated (bias = 0.025) with a drastic decrease in the CP (CP = 56%). In addition, the CPs of σ_{v_S}, σ_{v_T} and $\sigma_{v_{ST}}$ were also reduced to 78%, 75% and 78% respectively. In the event of a weak trial-level association, we observed high overestimation of R^2_{trial} regarding the individual-level association (absolute bias = 0.047 respectively 0.069 when τ = 0.60 respectively 0.33). With the last scenario, we observed a decrease in the CP of R^2_{trial} (70% when τ = 0.60 and 77% when τ = 0.33). Overall, as in the reference design, all scenarios tended to overestimate the estimation of the variance parameters of the random effects treatment-by-trial interaction (σ_{v_S}, σ_{v_T} and $\sigma_{v_{ST}}$). The model was robust enough in estimating R^2_{trial}.

Sensitivity analysis

To assess the robustness of the model with regard to the initial values, the smoothing parameters used, the number of spline knots, the number of samples used for the Monte Carlo integration, and the choice of the copula function, we performed a sensitivity analysis by varying these parameters. Simulation results are presented in Table 3. Regardless of the scenario, the model was robust enough with well estimated R^2_{trial} and Kendall's τ, and the results comparable to those of the reference scenario.

In the scenario with default initial values, we set γ, α, σ_{v_S}, σ_{v_T}, β_S, β_T to 0.5, θ to 1 and $\sigma_{v_{ST}}$ to 0.48. Results in Table 3 for this simulation design are quite close to the reference in terms of point estimate, the SEs and the CPs. We made the same observation by using the same couple of smoothing parameters

Table 2 Variation in data characteristics. Estimates (Mean), mean of standard errors (SE), empirical standard errors (SD) and percentage of coverage (CP). Estimation based on a Monte-Carlo (MC) integration method[‡], for M = 500 samples, N = 600 subjects and G = 30 trials

Parameters	True	Mean	SD	SE	CP	Mean	SD	SE	CP
		N = 1000 and G = 50				N = 200 and G = 10			
θ	3	3.023	0.231	0.235	95	3.175	0.607	0.55	93
γ	0.8	0.802	0.232	0.15	80	0.831	0.501	0.355	79
α	1	1.003	0.043	0.041	93	1.006	0.14	0.11	93
σ_{v_S}	0.7	0.739	0.229	0.189	90	0.684	0.492	0.392	78
σ_{v_T}	0.7	0.806	0.293	0.214	88	0.768	0.759	0.475	75
$\sigma_{v_{ST}}$	0.63	0.694	0.234	0.184	90	0.643	0.535	0.384	78
β_S	-1.25	-1.241	0.147	0.138	92	-1.264	0.321	0.289	91
β_T	-1.25	-1.245	0.159	0.148	93	-1.291	0.354	0.32	91
R^2_{trial}	0.81	0.814	0.089	0.071	82	0.835	0.23	0.22	56
τ	0.6	0.601	0.018	0.019	96	0.608	0.045	0.041	93
CI[††]:n(%)	-	10(2)	-	-	-	78(16)	-	-	-
		Variation of n_i [‡‡]				Censoring rate = 69%			
θ	3	3.06	0.315	0.307	94	3.107	0.493	0.481	96
γ	0.8	0.78	0.271	0.176	78	1.071	0.586	0.304	80
α	1	1.001	0.059	0.055	92	1.003	0.096	0.094	94
σ_{v_S}	0.7	0.731	0.321	0.242	85	0.706	0.407	0.321	85
σ_{v_T}	0.7	0.765	0.37	0.275	83	0.819	0.682	0.424	83
$\sigma_{v_{ST}}$	0.63	0.673	0.313	0.233	84	0.66	0.42	0.309	83
β_S	-1.25	-1.272	0.184	0.167	92	-1.21	0.234	0.216	91
β_T	-1.25	-1.273	0.189	0.181	94	-1.212	0.274	0.261	93
R^2_{trial}	0.81	0.813	0.134	0.105	80	0.814	0.199	0.38	72
τ	0.6	0.603	0.025	0.024	94	0.605	0.037	0.037	94
CI[††]:n(%)	-	28(6)	-	-	-	18(4)	-	-	-
		R^2_{trial} = 0.36 and τ = 0.60				R^2_{trial} = 0.36 θ = 1 τ = 0.33			
θ	3	3.039	0.305	0.306	95	1.016	0.152	0.155	96
γ	0.8	0.853	0.334	0.201	77	0.808	0.306	0.212	81
α	1	1	0.061	0.055	93	1.003	0.102	0.092	92
σ_{v_S}	0.7	0.739	0.321	0.241	86	0.708	0.303	0.258	86
σ_{v_T}	0.7	0.787	0.38	0.248	79	0.759	0.403	0.303	83
$\sigma_{v_{ST}}$	0.42	0.474	0.29	0.193	81	0.464	0.282	0.21	85
β_S	-1.25	-1.262	0.214	0.175	90	-1.269	0.203	0.185	91
β_T	-1.25	-1.249	0.21	0.186	92	-1.248	0.215	0.198	93
R^2_{trial}	0.36	0.407	0.21	0.139	70	0.429	0.221	0.178	77
τ	0.6	0.602	0.024	0.024	95	0.335	0.034	0.034	96
CI[††]:n(%)	-	4(1)	-	-	-	11(2)	-	-	-

[‡]Update of smoothing parameters in event of convergence issues, samples for MC = 1000, spline knots = 8, true initial values used, generation and estimation using Clayton copula model, Censoring rate = 44%; [‡‡] n_i = number of subjects per trial, based on Ovarian cancer Meta-Analysis Project (1991); [††]CI = Convergence issues.

for all generated datasets, as well as by considering 10 spline knots. However, by increasing the number of spline knots and therefore the number of parameters, the model parameters were in general difficult to estimate. By reducing the number of samples for MC to 500, we observed a more marked overestimation

Table 3 Sensitivity analysis. Estimates (Mean), mean of standard errors (SE), empirical standard errors (SD) and percentage of coverage (CP). Estimation based on a Monte-Carlo (MC) integration method, for M = 500 samples, N = 600 subjects and G = 30 trials

Parameters	True	Mean	SD	SE	CP	Mean	SD	SE	CP
		Reference[‡]				Default initial values			
θ	3	3.034	0.298	0.303	96	3.032	0.297	0.303	96
γ	0.8	0.814	0.294	0.198	79	0.822	0.321	0.199	80
α	1	1.002	0.059	0.053	93	1.002	0.058	0.053	93
σ_{v_S}	0.7	0.722	0.307	0.239	84	0.733	0.327	0.24	84
σ_{v_S}	0.7	0.78	0.389	0.268	82	0.776	0.396	0.263	82
$\sigma_{v_{ST}}$	0.63	0.677	0.318	0.229	83	0.679	0.33	0.228	82
β_S	-1.25	-1.27	0.206	0.175	91	-1.260	0.203	0.175	92
β_T	-1.25	-1.261	0.209	0.187	92	-1.251	0.208	0.186	92
R^2_{trial}	0.81	0.822	0.116	0.095	80	0.819	0.118	0.097	80
τ	0.6	0.601	0.023	0.024	95	0.601	0.023	0.024	95
CI[††]:n(%)	-	2(0)	-	-	-	12(2)	-	-	-
		kappa used = 0[†]				Spline knots = 10			
θ	3	3.031	0.297	0.303	96	3.047	0.299	0.304	96
γ	0.8	0.783	0.261	0.19	81	0.849	0.324	0.208	81
α	1	1.002	0.058	0.053	94	1.003	0.058	0.054	93
σ_{v_S}	0.7	0.709	0.288	0.233	85	0.719	0.296	0.235	84
σ_{v_T}	0.7	0.763	0.367	0.259	83	0.776	0.369	0.264	83
$\sigma_{v_{ST}}$	0.63	0.662	0.296	0.224	83	0.674	0.302	0.226	82
β_S	-1.25	-1.279	0.198	0.175	92	-1.256	0.203	0.174	92
β_T	-1.25	-1.271	0.205	0.186	93	-1.247	0.208	0.186	93
R^2_{trial}	0.81	0.822	0.115	0.095	79	0.823	0.115	0.094	80
τ	0.6	0.601	0.023	0.024	95	0.602	0.023	0.024	95
CI[††]:n(%)	-	7(1)	-	-	-	10(2)	-	-	-
		Samples for MC = 500				Gumbel Copula			
θ	3	3.038	0.299	0.303	96	0.974	0.115	0.103	*
γ	0.8	1.032	0.421	0.246	78	0.824	0.27	0.217	86
α	1	1.001	0.059	0.053	92	1.021	0.071	0.069	93
σ_{v_S}	0.7	0.697	0.316	0.223	80	0.726	0.29	0.253	88
σ_{v_T}	0.7	0.801	0.423	0.261	80	0.791	0.38	0.304	88
$\sigma_{v_{ST}}$	0.63	0.668	0.332	0.215	79	0.68	0.302	0.247	88
β_S	-1.25	-1.21	0.208	0.169	88	-1.285	0.203	0.183	92
β_T	-1.25	-1.223	0.219	0.184	90	-1.276	0.21	0.198	94
R^2_{trial}	0.81	0.81	0.121	0.096	80	0.817	0.127	0.115	84
τ	0.6	0.602	0.023	0.024	95	0.492	0.029	0.027	*
CI[††]:n(%)	-	1(0)	-	-	-	0(0)	-	-	-

[‡]Update of smoothing parameters in event of convergence issues, samples for MC = 1000, spline knots = 8, true initial values used, generation and estimation using Clayton copula model; [†]First smoothing parameter used for all generated datasets; [††]CI = Convergence issues; *true values for θ and τ are unknown given that data were generated using Clayton Copula model and were estimated using Gumbel Copula model.

in the estimates of γ (bias = 0.232) and σ_{v_T} (bias = 0.101), as well as a decrease in the CP of γ, σ_{v_S}, σ_{v_T} and $\sigma_{v_{ST}}$. However, we observed better estimation of R^2_{trial} with a bias < 10^{-3}.

We observed an improvement in the CP for R^2_{trial} and all estimated parameters when the estimates were based on the Gumbel Copula model. Given that data were generated using the Clayton Copula, we have not shown the CP for θ and Kendall's τ in the estimation with the Gumbel Copula. In addition, in extreme scenarios such as those with strong individual level and trial-level associations (Kendall τ = 0.9, R^2_{trial} = 0.95), the Gumbel Copula was more stable in terms of convergence issues (not converged dataset = 36, 7%) than the Clayton Copula (not converged dataset = 406, 81%).

Comparison between proposed one-step joint frailty-copula model and one-step joint surrogate approaches (Sofeu et al., 2019), for M = 500 samples, N = 600 subjects and G = 30 trials

Table 4 shows comparisons between the joint frailty-copula model and the joint surrogate model of Sofeu et al. (2019). We considered two main scenarios depending on the model used to generate the meta-analysis: the proposed model or the joint surrogate model.

Overall, regardless of the generated model, results between the Gumbel copula and the joint surrogate models were comparable concerning R^2_{trial} in terms of absolute bias and CP. The previous models had better properties in the estimation of R^2_{trial} compared to the Clayton copula model, which led to stronger overestimation, mainly when data were generated based on the joint frailty model (bias = 0.053 and CP = 60%). Regardless of the generated model, we observed better CPs for Kendall's τ (around 95% versus 80%) when they were estimated using the Clayton Copula. However, the Clayton Copula and the joint surrogate models showed comparable results for Kendall's τ in terms of point estimate and comparable biases. All models were comparable in terms of MSE for R^2_{trial} and Kendall's τ.

Regarding the convergence issues, the Copula models were more stable with less than 1% of rejection compared to the joint surrogate model, which had 9% of rejection when data generation was based on the Clayton Copula model. Using the joint surrogate model for the estimation, the non-convergence rate drastically increased with a drastic increase in the individual-level associations as shown in the last part of Table 4. With Kendall's τ = 0.81, the non-convergence rate was 58 %. In addition, the bias on τ was 0.084 and the CP was 1 % compared to the Clayton Copula with a bias equal to 0.002 and CP equal to 94%.

6 Application to advanced ovarian cancer meta-analysis dataset

6.1 Data collection and endpoints

The `dataOvarian` dataset (Ovarian cancer Meta-Analysis Project, 1991) combines data that were collected in four double-blind randomized clinical trials in advanced ovarian cancer. In the first two trials of this study, data were available in the centers in which patients were treated, and each of the other two trials were considered as an homogeneous group according to the investigators. Finally, the statistical unit in the first two trials was center, and it was trial in the other two trials. Therefore, total of 50 units was available for surrogacy evaluation. The objective in these studies was to examine the efficacy of cyclophosphamide plus cisplatin (CP) versus cyclophosphamide plus adriamycin plus cisplatin (CAP) to treat advanced ovarian cancer. The candidate surrogate endpoint **S** was progression-free survival time (PFS), defined as the time (in years) from randomization to clinical progression of the disease or death. The true endpoint **T** was survival time, defined as the time (in years) from randomization to death from any cause.

6.2 Data description

The dataset includes 1192 subjects with 82% of PFS-related events and 79.8% of deaths. The median survival time was 2.59 months [Interquartile range (IQR): 1.20 - 6.66] for PFS and 3.66 months [IQR: 1.84 - 9.08] for OS. As shown in Figure 1, 79% (750/951) of deceased patients had progressed before.

Table 4 Estimates (Mean), Bias, and Mean square errors (MSE): Comparison between joint frailty-copula model and joint surrogate model (Sofeu et al., 2019), and study of misspecification; for M = 500 samples, N = 600 subjects and G = 30 trials.

		Model used for data generation								
		Clayton copula				Joint surrogate				
Parameters	True	Mean	Bias	MSE	CP	True	Mean	Bias	MSE	CP
	θ = 3 and Kendall's τ = 0.6									
Clayton copula[†]										
R^2_{trial}	0.81	0.822	-0.012	0.013	80	0.81	0.863	-0.053	0.053	60
Kendall's τ	0.6	0.601	-0.001	0.001	95	0.537	0.519	0.018	0.002	93
CI[††]:n(%)			2(0)					2(0)		
Gumbel Copula[†]										
R^2_{trial}	0.81	0.817	-0.007	0.016	84	0.81	0.833	-0.023	0.065	67
CI[††]:n(%)			0(0)					1(0)		
joint surrogate[‡]										
R^2_{trial}	0.81	0.802	0.008	0.022	85	0.81	0.819	-0.009	0.064	69
Kendall's τ	0.6	0.632	-0.032	0.002	82	0.537	0.532	0.005	0.002	79
CI[††]:n(%)			45(9)					2(0)		
	θ = 8 and Kendall's τ = 0.8[‡‡]									
Clayton copula[†]										
R^2_{trial}	0.81	0.802	0.008	0.013	76					
Kendall's τ	0.8	0.802	-0.002	0	94					
CI[††]			2(0)							
Gumbel Copula[†]										
R^2_{trial}	0.81	0.774	0.036	0.018	86					
CI[††]			2(0)							
joint surrogate[‡]										
R^2_{trial}	0.81	0.816	-0.006	0.018	79					
Kendall's τ	0.8	0.716	0.084	0.009	1					
CI[††]			290(58)							

[†]Estimation using Monte Carlo integration method with 1000 replications, [‡] Estimation based on a combination of Monte-Carlo with (300 replications) and non-adaptive Gaussian Hermite quadrature (with 20 knots) integration, [††]CI = Convergence issues, [‡‡]Very large individual-level association

The median delay between progression and death was 0.8 [IQR: 0.3-1.7] months. There was a median of 11.5 (IQR: 6.0-18.8)% of subjects included per center, and almost 50 % of randomized patients in each treatment arm. The Kaplan Meier curves for both the surrogate and the true endpoints were close enough, i.e. proof of a high individual-level association (see Figure 2).

6.3 Surrogacy evaluation using the proposed joint frailty-copula surrogate model

We studied the validity of PFS as a good surrogate for OS in advanced ovarian cancer meta-analysis using the Clayton Copula, the Gumbel Copula and the joint surrogate model. The choice of the best model was based on both the surrogate threshold effect (STE) and the approximate likelihood cross-validation criteria (LCV_a). STE is the minimum treatment effect on the surrogate necessary to predict a non-zero (significant) effect on the true endpoint (Burzykowski and Buyse, 2006). LCV_a is the equivalent of AIC in the penalized likelihood framework (Commenges et al., 2007). Results for the application are shown in Table 5.

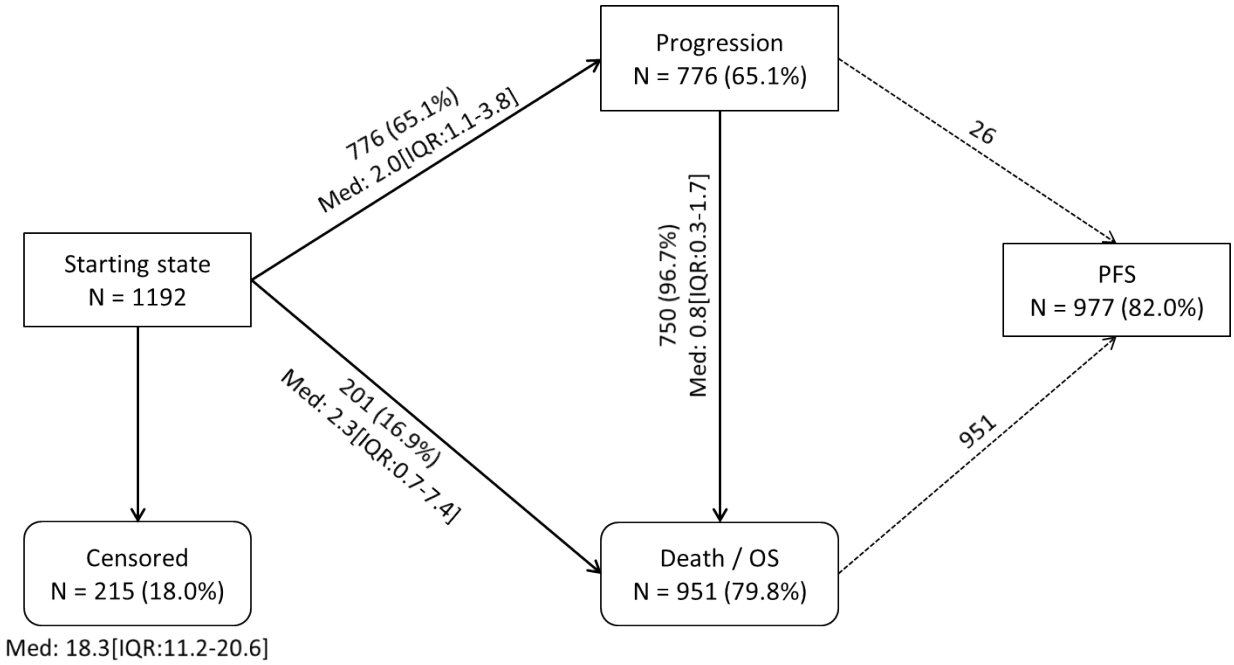


Figure 1 Flowchart of included individuals and description of survival times (in months) according to endpoint. OS = overall survival, PFS = progression-free survival, Med = median, IQR = Interquartile range. Observation from the Ovarian cancer Meta-Analysis Project (1991)

.

Table 5 Evaluation of individual and trial level surrogacy of **progression-free survival** for **overall survival** , based on the advanced ovarian cancer meta-analysis dataset.

Estimation model	R^2_{trial}	Kendall's τ	STE	LCV$_a$
Joint frailty-copula				
Clayton copula	0.997 (0.934 - 1.06)	0.84 (0.82 - 0.85)	-0.004 (HR = 0.996)	6.68
Gumbel Copula	1.00 (0.993 - 1.01)	0.79 (0.77 - 0.81)	-0.101 (HR = 0.904)	6.71
Joint surrogate†	1.00 (0.998 - 1.00)	0.68 (0.66 - 0.70)	-0.274 (HR = 0.761)	9.16

†Model including shared random effects to account for heterogeneity at individual level, STE = Surrogate threshold effect, LCV$_a$ = approximate likelihood cross-validation criterion

We observed a strong trial-level association with R^2_{trial} close to 1, whatever the estimation method used. The Clayton Copula model showed the highest Kendall's τ for surrogacy evaluation at the individual level (τ = 0.84 (0.82 - 0.85)). On the other hand, the joint surrogate model showed the lowest Kendall's τ with a value 0.68 (0.66 - 0.70). The weak value of Kendall's τ corroborates the observation made in the simulation studies that τ was underestimated by the joint surrogate model in the event of a strong individual-level association. The Clayton Copula model also showed the smallest absolute value (0.004) for log STE and therefore is the best model to use for this dataset. This observation is also supported by LCV$_a$ which is the smallest of the three observed values. This result suggests that, even with a slight significance effect of treatment on PFS (log STE = -0.004), one can predict a significant effect of treatment on OS by using the joint frailty Clayton Copula model. Therefore, we confirmed that PFS is a good surrogate endpoint for OS in advanced ovarian cancer, based on the dataset from the Ovarian cancer Meta-Analysis Project (1991)

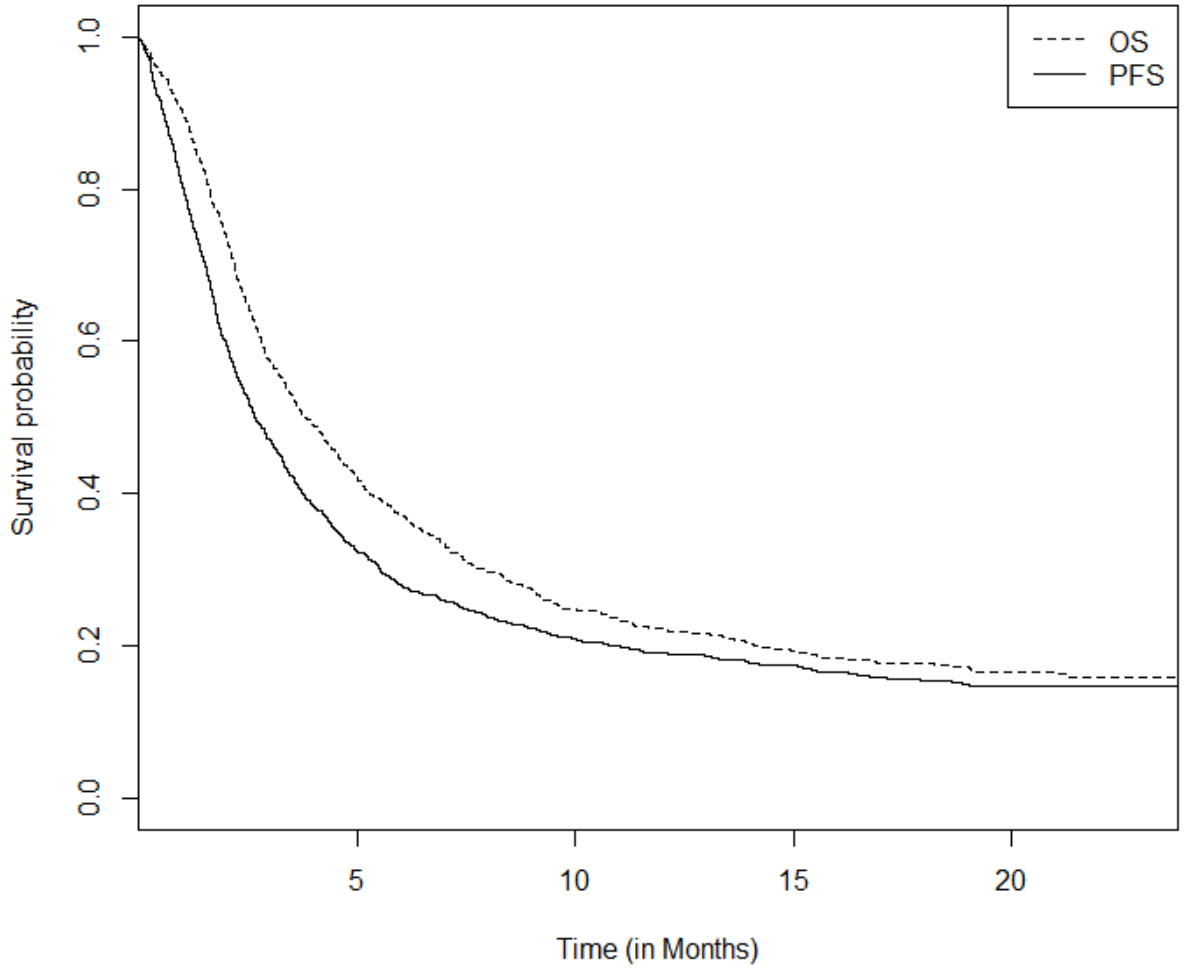


Figure 2 Kaplan Meier curves for survival functions associated with overall survival (OS) and progression-free survival (PFS), Observation from the Ovarian cancer Meta-Analysis Project (1991)

.

To assess the efficiency of the prediction with the different models, we performed leave-on-out cross-validation (loocv). Results suggest a prediction error of 8.3% (4/48) for the Clayton copula, 34.7% (17/49) for the Gumbel copula and 47.8% (22/46) for joint surrogate models. Figure 3 shows results for trials in which models reached convergence. We excluded from the graph 14% (7/50) of trials with outliers on the observed or predicted treatment effect on the true endpoint.

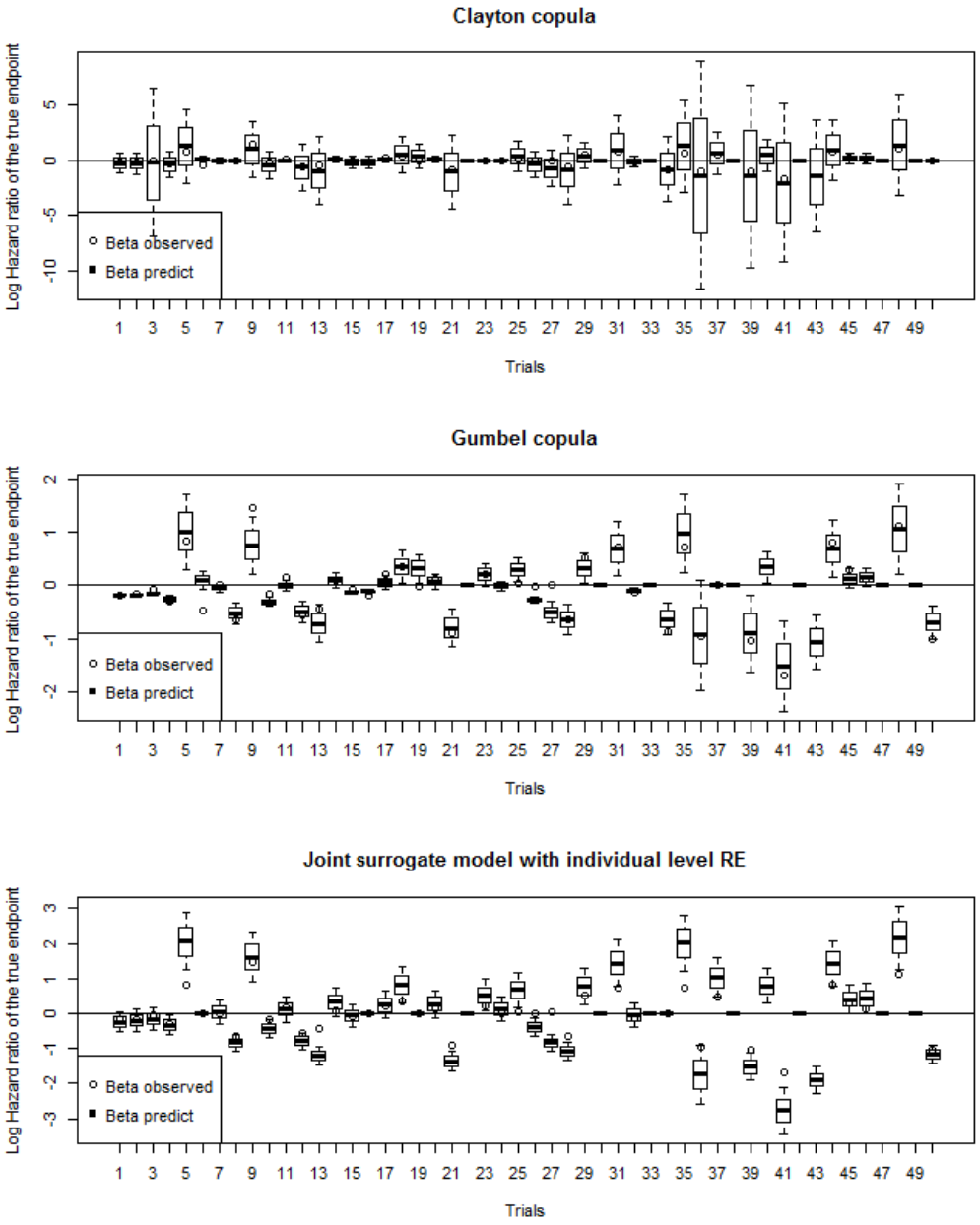


Figure 3 Leave-one-out cross-validation results associated with advanced ovarian cancer meta-analysis for the assessment of the Clayton copula model, the Gumbel copula model and the joint surrogate model with individual level random effects (RE). Boxplots represent the predicted treatment effects on the true endpoint (T) with the prediction intervals, the circles for the observed treatment effect on T and the dashes for unused trials.

7 Discussion

In this article, we developed a new model for the one-step validation of time-to-event surrogate endpoints based on individual patients data, from a meta-analysis of randomized controlled trials. The model is a variant from the previous one-step joint surrogate model (Sofeu et al., 2019), which included shared random effects ω_{ij} in combination with a power parameter ζ to take into account heterogeneity in the data at the individual level. The variance of individual level random effects ω_{ij} (θ) in the previous model served as a proxy for the individual-level association measurement between endpoints and was used for a novel definition of Kendall's τ, in combination with ζ, the variance of the shared random effects at trial level associated with the baseline risk γ and the power parameter α. In the new approach, we propose copula models to investigate the association between the surrogate and true endpoints based on bivariate copula models in which estimates for θ and α are replaced by the estimate of the copula parameter θ. A great advantage of the copula is the simplicity in the definition of Kendall's τ, as a function of θ.

The fundamental difference between the proposed joint frailty-copula model and the two-step copula approach of Burzykowski et al. (2001) in addition to the specification of the baseline risks is that we included two correlated random effects treatment-by-trial interaction σ_{v_S} and σ_{v_T} in the survivor functions for surrogate and true endpoints, so all model parameters are estimated in one step. This avoids to take into account estimation errors from the first step, which usually causes convergence issues or model estimation in the two-step copula model (Renfro et al., 2012; Sofeu et al., 2019; Shi et al., 2011; Rotolo et al., 2019). Moreover, in the proposed approach, heterogeneity on the baseline risk is taken into account using a shared random effect, instead of stratification across trials. This allows the number of model parameters to be reduced.

Results obtained from the simulation studies show that the joint frailty-copula model has good properties in estimating of Kendall's τ and R^2_{trial}. Moreover, regarding the integration method used, we found comparable accuracy when estimating the surrogacy evaluation criteria with MC, Laplace and PGH integration methods. However, estimates of the variance parameters for the random effects treatment-by-trial interaction (Σ_v) and the CP for R^2_{trial} were better using the GH integration. This suggests using the GH first and then choosing between the other integration methods in the event of convergence issues in order to obtain comparable results. The drawback of GH integration is at the level of convergence issues and the fact that regardless of parallel computing, it requires more computer resources for fast estimation. The use of the pseudo-adaptive Gaussian-Hermite quadrature with fewer quadrature nodes did not improve the estimates of Σ_v as suggested by Rizopoulos (2012). Sofeu et al. (2019) also observed satisfactory results by approximating the integrals over the trial level random effects using the MC with 300 samples when fitting the joint surrogate model.

By varying the characteristics of the meta-analysis and therefore the number of trials, the proportion of subjects per trials, the censorship, and the individual level and the trial level surrogacy, we found that the proposed model was robust for estimating Kendall's τ with satisfactory CP. We observed an underestimation of R^2_{trial} in the event of a very small number of trials and a weak trial-level association. These results call for cautious when the meta-analysis includes a low number of trials and therefore to take into account other analysis units such as the center.

In the sensitivity analysis, the model was robust for the initial values, the smoothing parameters, the number of knots for splines, the number of samples for the MC integration and the choice of the copula function. This represents different parametrization of the estimation model to manage the convergence issues. Therefore, in the event of convergence and regardless of the parametrisation used, the user can be confident in the quality of the fit. However, we recommand using at least 300 samples for MC integration for a good estimate of R^2_{trial}. Using the Gumbel copula family, we more often obtained few convergence issues with better CP for all estimates of model parameters, although the data were generated with the Clayton copula model. However, the approximate likelihood cross-validation criterion can be used to choose the correct parametrization of the model. The robustness of the Gumbel copula has also been discussed by Rotolo et al. (2019) in the assessment of the two-step copula model when meta-analyses were

generated with the mixed proportional hazard model, the Clayton copula and the Poisson models. Due to convergence issues inherent in the two-step approach, results were quite controversial.

In the comparative analysis, the joint surrogate models were robust with individual level random effects (Sofeu et al., 2019) and with the Gumbel copula family compared to the joint surrogate model with the Clayton copula family, in terms of misspecification. The model with individual level random effects showed similar results with the Gumbel copula for estimation of R^2_{trial} and with the Clayton copula for estimation of moderate Kendall's τ. These results suggest that the Clayton copula family is more appropriate for estimating individual level surrogacy compared to the model with individual level random effects. In addition, when the true Kendall's τ is very high, we observed an increase in the underestimation of τ with CP close to 1% using the individual level random effects model. Similar results were observed in Sofeu et al. (2019) for the assessment of disease-free survival as a surrogate for OS in gastric cancer, in which the previous model showed the smallest τ compared to the approaches based on one-step Poisson models, and two-step Clayton and Plackett copula models. Therefore, for $\tau > 0.6$, the interpretation of individual level surrogacy should take into account the potential bias in the event of estimation based on the individual level random effects model. This drawback can be avoided by using the proposed joint frailty-copula model.

In the application, we confirmed that PFS is a good surrogate for OS in advanced ovarian cancer. Compared to the two step-copula approach in which the adjusted R^2_{trial} was not available when the method was been applied to the advanced ovarian cancer meta-analysis (Burzykowski et al., 2001), the new model converged. The value of τ for the Clayton model was close to the 0.871[0.860-0.883] and 0.853[0.842-0.863] observed by Burzykowski et al. (2001) for Clayton's and Gumbel's two-step copula model.

To conclude, we proposed additional tools for validating the failure-time surrogate endpoint based on a novel joint frailty-copula model. The proposed model seems more robust than the existing approaches in the assessment of very high individual level and trial level surrogacy, with excellent properties for estimating of individual level surrogacy. The most important advantages of this model beyond its robustness is the possibility to consider two copula families and the flexibility in the parametrization of the model to manage convergence issues. Although we just considered the Archimedian Clayton and Gumbel copula, the model can be extended to other copula families. In addition, the new model makes it possible to consider potential confounders and therefore to model jointly any pair of failure-time endpoints in a setting different from that of surrogacy evaluation.

Acknowledgements The authors thank the ovarian cancer meta-analysis project for permission to use their data. Computer times were provided by the computing facilities of MCIA (Mésocentre de Calcul Intensif Aquitain) at the Université de Bordeaux and the Université de Pau et des Pays de l'Adour. This work was supported by the Association pour la Recherche sur le Cancer, Grant/Award Number: PJA20161205147; Institut National du Cancer, Grant/Award Number: 2017-125; Institut National de la Santé et de la Recherche Médicale; Région Aquitaine

Conflict of Interest

The authors have declared no conflict of interest.

Appendix

A.1. Log-likelihood construction

Assume a pair of endpoints measured on subject j ($j = 1, \cdots, n_i$) in trial i ($i = 1, \cdots, G$), corresponding to the survival times associated with the surrogate (S_{ij}) and the true endpoints (T_{ij}). Let $\mathbf{v}_i = (u_i, v_{Si}, v_{Ti})$ the vector of trial level random effects from models (5-6), and $\mathbf{Z}_{K,ij} = (\mathbf{Z}_{S,ij}, \mathbf{Z}_{T,ij})$ the vector of covariates for the surrogate and the true endpoints. The conditional joint survival function for the pair of

endpoints from (4) is:

$$\begin{aligned}\bar{F}(s_{ij},t_{ij}|\mathbf{Z}_{S,ij},\mathbf{Z}_{T,ij},\mathbf{v}_i) &= P(S_{ij}>s_{ij},T_{ij}>t_{ij}|\mathbf{Z}_{S,ij},\mathbf{Z}_{T,ij},\mathbf{v}_i) \\ &= \varphi_\theta[\varphi_\theta^{-1}(\bar{F}(s_{ij}|\mathbf{Z}_{S,ij},u_i,v_{Si}))+\varphi_\theta^{-1}(\bar{F}(t_{ij}|\mathbf{Z}_{T,ij},u_i,v_{Ti}))] \\ &= \bar{F}_{ST}\end{aligned} \tag{10}$$

The generator φ_θ can be rewritten as a Laplace transformation of a positive distribution function $G_\theta(x)$, with $G_\theta(0)=1$ as:

$$\varphi_\theta(t)=\int_0^{+\infty}\exp(-tx)dG_\theta(x),\quad t\geqslant 0 \tag{11}$$

By assuming $\bar{F}(s_{ij},t_{ij}|\mathbf{Z}_{S,ij},\mathbf{Z}_{T,ij},\mathbf{v}_i)=\bar{F}_{ST}$, the conditional joint survival function can be rewritten as follows:

$$\begin{aligned}\bar{F}_{ST} &= \int_0^{+\infty}\exp\Big(-x[\varphi_\theta^{-1}(\bar{F}(s_{ij}|\mathbf{Z}_{S,ij},u_i,v_{Si}))+\varphi_\theta^{-1}(\bar{F}(t_{ij}|\mathbf{Z}_{T,ij},u_i,v_{Ti}))]\Big) \\ &\quad dG_\theta(x) \\ &= \int_0^{+\infty}\exp\Big(-x[\varphi_\theta^{-1}(\bar{F}(s_{ij}|\mathbf{Z}_{S,ij},u_i,v_{Si}))\Big)\times \\ &\quad exp\Big(-x[\varphi_\theta^{-1}(\bar{F}(t_{ij}|\mathbf{Z}_{T,ij},u_i,v_{Ti}))]\Big)dG_\theta(x) \\ &= \int_0^{+\infty}\prod_{K\in\{S,T\}}\exp\Big(-x[\varphi_\theta^{-1}(\bar{F}(k_{ij}|\mathbf{Z}_{K,ij},u_i,v_{Ki}))\Big)dG_\theta(x),\end{aligned} \tag{12}$$

Let $\Phi=(\theta,\sigma^2_{v_S},\sigma^2_{v_T},\sigma_{v_{ST}},\gamma,\lambda_{0T}(.),\lambda_{0S}(.),\boldsymbol{\beta}_S,\boldsymbol{\beta}_T)$ be the vector containing all the unknown parameters of the model (4, 5 and 6). The contribution of subject j to the likelihood function is obtained from the derivatives of the joint survival function with respect to uncensored survival time in the pair. Then this contribution is equivalent to:

$$L_{ij}(\Phi|\mathbf{Z}_{K,ij},\mathbf{v}_i)=(-1)^{d_{ij}}\frac{\partial^{d_{ij}}}{(\partial x_{S,ij})^{\delta_{S,ij}}(\partial x_{T,ij})^{\delta_{T,ij}}}\bar{F}(x_{S,ij},x_{T,ij}|\mathbf{Z}_{S,ij},\mathbf{Z}_{T,ij},\mathbf{v}_i),$$

where $d_{ij}=\sum_{K\in\{S,T\}}\delta_{k,ij}$ is the number of censored event times in subject j. From (12), this derivative is given by:

$$\begin{aligned}L_{ij}(\Phi|\mathbf{Z}_{K,ij},\mathbf{v}_i) &= (-1)^{d_{ij}}\frac{\partial^{d_{ij}}}{(\partial x_{S,ij})^{\delta_{S,ij}}(\partial x_{T,ij})^{\delta_{T,ij}}}\int_0^{+\infty} \\ &\quad \prod_{K\in\{S,T\}}\exp\Big(-x[\varphi_\theta^{-1}(\bar{F}(k_{ij}|\mathbf{Z}_{K,ij},u_i,v_{Ki}))\Big) \\ &\quad dG_\theta(x) \\ &= \int_0^{+\infty}\exp\Big(-x\sum_{K\in\{S,T\}}[\varphi_\theta^{-1}(\bar{F}(k_{ij}|\mathbf{Z}_{K,ij},u_i,v_{Ki}))]\Big) \\ &\quad \prod_{K\in\{S,T\}}\left[\frac{-xf(k_{ij}|\mathbf{Z}_{K,ij},u_i,v_{Ki})}{\varphi'_\theta(\varphi_\theta^{-1}(\bar{F}(k_{ij}|\mathbf{Z}_{K,ij},u_i,v_{Ki})))}\right]^{\delta_{k,ij}}dG_\theta(x),\end{aligned}$$

where $f = -d\bar{F}/dt$ is the conditional density of the lifetime k_{ij}. Given $\mathbf{v}_i$, all subjects are independent. Therefore, the conditional contribution to the likelihood of the i^{th} trial is given by:

$$\begin{aligned} L_i(\Phi|\mathbf{Z}_{K,ij},\mathbf{v}_i) &= \prod_{j=1}^{n_i}\int_0^{+\infty}\exp\Big(-x\sum_{K\in\{S,T\}}[\varphi_\theta^{-1}(\bar{F}(k_{ij}|\mathbf{Z}_{K,ij},u_i,v_{Ki}))]\Big) \\ &\quad \prod_{K\in\{S,T\}}\left[\frac{-xf(k_{ij}|\mathbf{Z}_{K,ij},u_i,v_{Ki})}{\varphi_\theta'(\varphi_\theta^{-1}(\bar{F}(k_{ij}|\mathbf{Z}_{K,ij},u_i,v_{Ki})))}\right]^{\delta_{k,ij}} dG_\theta(x) \\ &= \prod_{j=1}^{n_i}\int_0^{+\infty}\prod_{K\in\{S,T\}}\exp\Big(-x[\varphi_\theta^{-1}(\bar{F}(k_{ij}|\mathbf{Z}_{K,ij},u_i,v_{Ki}))]\Big) \\ &\quad \left[\frac{-xf(k_{ij}|\mathbf{Z}_{K,ij},u_i,v_{Ki})}{\varphi_\theta'(\varphi_\theta^{-1}(\bar{F}(k_{ij}|\mathbf{Z}_{K,ij},u_i,v_{Ki})))}\right]^{\delta_{k,ij}} dG_\theta(x). \end{aligned}$$

Hence, by integrating L_i over the random effects $\mathbf{v}_i$, the marginal contribution of trial i is:

$$\begin{aligned} L_i(\Phi) &= \int_{\mathbf{v}_i}\Big\{\prod_{j=1}^{n_i}\int_0^{+\infty}\prod_{K\in\{S,T\}} exp\Big(-x[\varphi_\theta^{-1}(\bar{F}(k_{ij}|\mathbf{Z}_{K,ij},u_i,v_{Ki}))]\Big) \\ &\quad \left[\frac{-xf(k_{ij}|\mathbf{Z}_{K,ij},u_i,v_{Ki})}{\varphi_\theta'(\varphi_\theta^{-1}(\bar{F}(k_{ij}|\mathbf{Z}_{K,ij},u_i,v_{Ki})))}\right]^{\delta_{k,ij}} dG_\theta(x)\Big\} f_{\mathbf{V}}(\mathbf{v}_i)d\mathbf{v}_i. \end{aligned}$$

Furthermore, since all the studies are independent, the marginal likelihood function for all trials is:

$$\begin{aligned} L(\Phi) &= \prod_{i=1}^{G} L_i(\Phi) \\ &= \prod_{i=1}^{G}\int_{\mathbf{v}_i}\Big\{\prod_{j=1}^{n_i}\int_0^{+\infty}\prod_{K\in\{S,T\}}\exp\Big(-x[\varphi_\theta^{-1}(\bar{F}(k_{ij}|\mathbf{Z}_{K,ij},u_i,v_{Ki}))]\Big) \\ &\quad \left[\frac{-xf(k_{ij}|\mathbf{Z}_{K,ij},u_i,v_{Ki})}{\varphi_\theta'(\varphi_\theta^{-1}(\bar{F}(k_{ij}|\mathbf{Z}_{K,ij},u_i,v_{Ki})))}\right]^{\delta_{k,ij}} dG_\theta(x)\Big\} f_{\mathbf{V}}(\mathbf{v}_i)d\mathbf{v}_i \end{aligned} \tag{13}$$

In general, it is difficult to evaluate expression (13), except for a very specific choice of the distribution G_θ. Given that the generator φ_θ is the Laplace transform of G_θ, there is an alternative expression for this likelihood function, which is found by using derivatives of this generator. i.e.

$$\begin{aligned} \varphi_\theta^{(m)}(t) &\equiv \frac{\partial^m}{(\partial t)^m}\varphi_\theta(t) \\ &= \int_0^{+\infty}(-x)^m\exp(-tx)dG_\theta(x), \quad m=0,1,\cdots. \end{aligned}$$

Thereby, the marginal likelihood (13) can be rewritten as :

$$\begin{aligned} L(\Phi) &= \prod_{i=1}^{G} L_i(\Phi) \\ &= \prod_{i=1}^{G} \int_{\mathbf{v}_i} \Bigg\{ \prod_{j=1}^{n_i} \int_0^{+\infty} (-x)^{d_{ij}} \prod_{K\in\{S,T\}} exp\Big(- x[\varphi_\theta^{-1}(\bar{F}(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki}))]\Big) \\ &\quad \left[\frac{f(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki})}{\varphi_\theta^{'}(\varphi_\theta^{-1}(\bar{F}(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki})))}\right]^{\delta_{k,ij}} dG_\theta(x)\Bigg\} f_{\mathbf{V}}(\mathbf{v}_i) d\mathbf{v}_i \\ &= \prod_{i=1}^{G} \int_{\mathbf{v}_i} \Bigg\{ \prod_{j=1}^{n_i} \varphi_\theta^{(d_{ij})}\Big(\sum_{K\in\{S,T\}} \varphi_\theta^{-1}(\bar{F}(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki}))\Big) \\ &\quad \prod_{K\in\{S,T\}} \left[\frac{f(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki})}{\varphi_\theta^{'}(\varphi_\theta^{-1}(\bar{F}(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki})))}\right]^{\delta_{k,ij}} \Bigg\} f_{\mathbf{V}}(\mathbf{v}_i) d\mathbf{v}_i \end{aligned} \tag{14}$$

Thus, we found the log-likelihood function defined in equation (7)

A.2. Computation of expressions in the likelihood (7)

From equations (5 - 6), the conditional survival function in equation (7) can be defined as:

$$\bar{F}(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki}) = \exp\Big\{ - \Lambda_K(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki})\Big\} \tag{15}$$

where $K \in \{S,T\}$, $\Lambda_K(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki})$ are the conditional cumulative hazard functions associated with the failure-time endpoints, defined as follows:

$$\begin{aligned} \Lambda_K(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki}) &= \int_0^{k_{ij}} \lambda_{0k}(x) \exp\Big(\alpha^{1-\mathbf{1}_{\{k=s\}}} u_i + v_{Ki} Z_{ij1} + \boldsymbol{\beta}_k \mathbf{Z}_{K,ij}\Big) dx \\ &= \int_0^{k_{ij}} \lambda_K(x|\mathbf{Z}_{K,ij}, u_i, v_{Ki}) dx, \end{aligned} \tag{16}$$

with $\lambda_K(x|Z_{ijq}, u_i, v_{k,i})$ the conditional instantaneous risk function associated with the even-times k.

The conditional distribution function $f(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki})$ for the event time k observed on subject j in trial i is given by:

$$\begin{aligned} f(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki}) &= -\frac{d\bar{F}(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki})}{dk_{ij}} \\ &= -\frac{d}{dk_{ij}}\left[\exp\Big\{ - \Lambda_K(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki})\Big\}\right] \\ &= \lambda_K(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki}) \exp\Big\{ - \Lambda_K(k_{ij}|\mathbf{Z}_{K,ij}, u_i, v_{Ki})\Big\} \end{aligned} \tag{17}$$

The generators associated with the Archimedian copulas are defined as shown by Prenen *et al.* (Prenen et al., 2017) for the Clayton copula as:

$$\varphi_\theta(s) = (1+\theta s)^{-1/\theta},$$

with the associated inverse, the first and the second order derivatives :

$$\varphi_\theta^{-1}(s) = \frac{s^{-\theta}-1}{\theta}, \quad \varphi_\theta^{'}(s) = -(1+\theta s)^{-(1+\theta)/\theta}, \quad \varphi_\theta^{''}(s) = (1+\theta)(1+\theta s)^{-(1+2\theta)/\theta}.$$

Therefore,

$$\begin{aligned}\varphi_\theta^{'}[\varphi_\theta^{-1}(s)] &= -(1+\theta(\frac{s^{-\theta}-1}{\theta}))^{-(1+\theta)/\theta} \\ &= -(1+(s^{-\theta}-1))^{-(1+\theta)/\theta} \\ &= -s^{(1+\theta)}\end{aligned}$$

For the Gumbel-Hougaard copula the generator is :

$$\varphi_\theta(s) = \exp\left[-s^{1/(1+\theta)}\right]$$

with the associated inverse, the first and the second order derivatives :

$$\varphi_\theta^{-1}(s) = [-\log(s)]^{(1+\theta)}, \quad \varphi_\theta^{'}(s) = -\tfrac{1}{(1+\theta)} s^{-\theta(1+\theta)} \exp\left[-s^{1/(1+\theta)}\right],$$
$$\varphi_\theta^{''}(s) = \tfrac{1}{(1+\theta)^2}\left[\theta s^{-(2\theta+1)/(\theta+1)} + s^{-2\theta/(\theta+1)}\right] \exp\left[-s^{1/(1+\theta)}\right].$$

Therefore,

$$\begin{aligned}\varphi_\theta^{'}[\varphi_\theta^{-1}(s)] &= -1/(1+\theta)[[-\log(s)]^{(1+\theta)}]^{-\theta(1+\theta)} \exp\left[-[[-\log(s)]^{(1+\theta)}]^{1/(1+\theta)}\right] \\ &= -1/(1+\theta)[-\log(s)]^{-\theta} s \\ &= -\frac{s}{(1+\theta)}\left[-\log(s)\right]^{-\theta}\end{aligned}$$

References

Abrahantes, J. C. and Burzykowski, T. (2005). A version of the em algorithm for proportional hazard model with random effects. *Biometrical Journal* **47,** 847–862.

Booth, C. M. and Eisenhauer, E. A. (2012). Progression-free survival: Meaningful or simply measurable? *Journal of Clinical Oncology* **30,** 1030–1033.

Branchoux, S., Bellera, C., Italiano, A., Rustand, D., Gaudin, A.-F., and Rondeau, V. (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.

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

Burzykowski, T., Molenberghs, G., Buyse, M., Geys, H., and Renard, D. (2001). Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *Journal of the Royal Statistical Society C (Applied Statistics)* **50,** 405–422.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* **1,** 49–67.

Commenges, D., Joly, P., Gégout-petit, A., and Liquet, 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.

Dowd, B. E., Greene, W. H., and Norton, E. C. (2014). Computation of standard errors. *Health Services Research* **49,** 731–750.

Duchateau, L. and Janssen, P. (2008). *The Frailty Model*. Springer-Verlag.

Emura, T., Nakatochi, M., Murotani, K., and Rondeau, V. (2017). A joint frailty-copula model between tumour progression and death for meta-analysis. *Statistical Methods in Medical Research* **26,** 2649–2666.

Genest, C. and MacKay, J. (1986). The joy of copulas: Bivariate distributions with uniform marginals. *The American Statistician* **40,** 280–283.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall.

Joly, P., Commenges, D., and Letenneur, L. (1998). A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia. *Biometrics* **54,** 185–194.

Joly, P., Letenneur, L., Alioum, A., and Commenges, D. (1999). Phmpl: a computer program for hazard estimation using a penalized likelihood method with interval-censored and left-truncated data. *Computer Methods and Programs in Biomedicine* **60,** 225 – 231.

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.

Matulonis, U. A., Oza, A. M., Ho, T. W., and Ledermann, J. A. (2015). Intermediate clinical endpoints: A bridge between progression-free survival and overall survival in ovarian cancer trials. *Cancer* **121,** 1737–1746.

Nelsen, R. B. (2006). *An introduction to Copulas*. Springer.

Ovarian cancer Meta-Analysis Project (1991). Cyclophosphamide plus cisplatin plus adriamycin persus cyclophosphamide, doxorubicin, and cisplatin chemotherapy of ovarian carcinoma: A meta-analysis. *Classic Papers and Current Comments* **3,** 237–234.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing* **9,** 363–379.

Prenen, L., Braekers, R., and Duchateau, L. (2017). Extending the archimedean copula methodology to model multivariate survival data grouped in clusters of variable size. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79,** 483–505.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statist. Sci.* **3,** 425–441.

Renfro, L. A., Shi, Q., Sargent, D. J., and Carlin, B. P. (2012). Bayesian adjusted r2 for the meta-analytic evaluation of surrogate time-to-event endpoints in clinical trials. *Statistics in Medicine* **31,** 743–761.

Rizopoulos, D. (2012). Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive gaussian quadrature rule. *Computational Statistics & Data Analysis* **56,** 491 – 501.

Rondeau, V., Commenges, D., and Joly, P. (2003). Maximum penalized likelihood estimation in frailty models. *Lifetime Data Analysis* **9,** 139–153.

Rondeau, V., Gonzalez, J. R., Mazroui, Y., Mauguen, A., Diakite, A., Laurent, A., Lopez, M., Krol, A., and Sofeu, C. L. (2019). `frailtypack`: *General Frailty Models: Shared, Joint and Nested Frailty Models with Prediction; Evaluation of Failure-Time Surrogate Endpoints*. `R` package version 3.0.3.

Rondeau, V., Pignon, J. P., and Michiels, S. (2015). A joint model for the dependence between clustered times to tumour progression and deaths: A meta-analysis of chemotherapy in head and neck cancer. *Statistical Methods in Medical Research* **24,** 711–729.

Rotolo, F., Paoletti, X., Burzykowski, T., Buyse, M., and Michiels, S. (2019). A poisson approach to the validation of failure time surrogate endpoints in individual patient data meta-analyses. *Statistical Methods in Medical Research* **28,** 170–183.

Rotolo, F., Paoletti, X., and Michiels, S. (2018). `surrosurv`: An `R` package for the evaluation of failure time surrogate endpoints in individual patient data meta-analyses of randomized clinical trials. *Computer Methods and Programs in Biomedicine* **155,** 189 – 198.

Savina, M., Gourgou, S., Italiano, A., Dinart, D., Rondeau, V., Penel, N., Mathoulin-Pelissier, S., and Bellera, C. (2018). Meta-analyses evaluating surrogate endpoints for overall survival in cancer randomized trials: A critical review. *Critical Reviews in Oncology/Hematology* **123,** 21 – 41.

Shi, Q., Renfro, L. A., Bot, B. M., Burzykowski, T., Buyse, M., and Sargent, D. J. (2011). Comparative assessment of trial-level surrogacy measures for candidate time-to-event surrogate endpoints in clinical trials. *Computational Statistics & Data Analysis* **55,** 2748 – 2757.

Sofeu, C. L., Emura, T., and Rondeau, V. (2019). One-step validation method for surrogate endpoints using data from multiple randomized cancer clinical trials with failure-time endpoints. *Statistics in Medicine* **38,** 2928–2942.

Supporting Information

All the models were estimated using an extension of the `R` package `frailtypack` (Rondeau et al., 2019). The new version of `frailtypack` implementing the new joint surrogate model will soon be available on the `Comprehensive R Archive Network (CRAN)`. However, an ongoing version can be found on github at `https://github.com/socale/frailtypack/tree/surrogacy_submitted_3-0-4_sansMPI`. Additional supporting information including the source code to reproduce the results may be found online in the Supporting Information section at the end of the article.

5.2 Code R pour l'application de ce nouveau modèle

Les résultats présentés dans la partie application de cet article peuvent être reproduits à partir du programme R qui suit :

```
library(frailtypack)
# Make sure to have the R file "plot.jointSurroPenalloocv.R" in your work directory
source("plot.jointSurroPenalloocv.R")
# Data description
data(dataOvarian, package = "frailtypack")
# number of subjects
nrow(dataOvarian)
# proportion PFS
round((prop.table(table(dataOvarian$statusS))["1"]*100),1)
# proportion Death
round((prop.table(table(dataOvarian$statusT))["1"]*100),1)
# median survivall times
round(quantile(dataOvarian$timeS*12, c(.50, .25, .75)),2)
round(quantile(dataOvarian$timeT*12, c(.50, .25, .75)),2)
# number of subjects include per trial round(quantile(table(dataOvarian$trialID), c(.50,
.25, .75)),1)
# randomized subjects per trial round((prop.table(table(dataOvarian$trt))["1"]*100))
# Figure 1: DAG for data description # # progression
attach(dataOvarian)
# patient with pregression time < death time
progress.before.death <- dataOvarian[statusS==1 & (statusT == 0) | (timeS < timeT),
c("timeS","timeT","statusS","statusT")] nrow(progress.before.death)
round(100 * nrow(progress.before.death) / nrow(dataOvarian),1)
round(quantile(progress.before.death$timeS*12, c(.50, .25, .75)),1)
# patient who death without progression
death.without.progression <- dataOvarian[statusT==1 & (timeS >= timeT),
c("timeS","timeT","statusS","statusT")] nrow(death.without.progression)
round(100 * nrow(death.without.progression)/nrow(dataOvarian),1)
round(quantile(death.without.progression$timeT*12, c(.50, .25, .75)),1)
# censored patients
without.event <- dataOvarian[statusT==0 & statusS==0, c("timeS","timeT","statusS","statusT")]
nrow(without.event)
```

```
round(100 * nrow(without.event)/nrow(dataOvarian),1) round(quantile(without.event$timeT*12,
c(.50, .25, .75)),1)
# death after progression
death.after.progress <- dataOvarian[statusT==1 & timeS < timeT,
c("timeS","timeT","statusS","statusT")]
nrow(death.after.progress)
round(100 * nrow(death.after.progress)/nrow(progress.before.death),1)
# delay between progression and death
round(quantile((death.after.progress$timeT - death.after.progress$timeS)*12, c(.50, .25, .75)),1)
detach(dataOvarian)
# Figure 2: Kaplan Meier curves
survS <- survfit(Surv(timeS*12, statusS) 1, data = dataOvarian)
survT <- survfit(Surv(timeT*12, statusT) 1, data = dataOvarian)
xlim1 <- range(c(survS$time, survT$time))
ylim1 <- c(0,1)
png(file = "kaplanMeier.png", width = 600, height = 500)
plot.new()
plot(survS, conf.int = FALSE, xlim = xlim1, ylim = ylim1, panel.first = abline(h = c(0,1)),
xlab = "Time (in Months)", ylab = "Survival probability", lty = 1) lines(survT,conf.int=FALSE,
lty = 2)
legend("topright", c("OS", "PFS"), lty = c(2,1))
dev.off()
# Table 5: Evaluation of PFS as a good surrogate for OS in the advanced ovariant cancer
# The values for smoothing parameters (init.kappa) have been choosen using the crossval-
idation as describes in the main manuscript.
# Estimation using the joint frailty Clayton copula model
joint.surro.Clayton <- jointSurroCopPenal(data = dataOvarian, int.method = 0,
n.knots = 8, maxit = 35, kappa.use = 4, nb.mc = 1000, typecopula = 1, print.iter = F,
init.kappa = NULL, scale = 1/365)
summary(joint.surro.Clayton)
save.image("clayton.RData")
# Estimation using the joint frailty Gumbel copula model
joint.surro.Gumbel <- jointSurroCopPenal(data = dataOvarian, int.method = 0,
n.knots = 8, maxit = 35, kappa.use = 4, nb.mc = 1000, typecopula = 2, print.iter = F,
init.kappa = NULL, scale = 1/365)
summary(joint.surro.Gumbel)
```

```
save.image("Gumbel.RData")
# Estimation using the joint surrogate model (Sofeu et al., 2019)
joint.surro.jointSurro <- jointSurroPenal(data = dataOvarian, n.knots = 8, kappa.use = 4,
nb.mc = 200, print.iter = F, scale = 1/365, init.kappa = NULL, indicator.alpha = 0)
summary(joint.surro.jointSurro)
save.image("Joinsurro.RData")
# ====Figure 3: Leave-one-out cross-validation (loocv) results====
# loocv from the Clayton copula
loocv.result.Clayton <- loocv(joint.surro.Clayton)
save.image("loocvclayton.RData")
# Gumbel copula
loocv.result.Gumbel <- loocv(joint.surro.Gumbel)
save.image("loocvGumbel.RData")
# joint surrogate
loocv.result.jointsurro<- loocv(joint.surro.jointSurro)
save.image("loocvJointSurr.RData")
# Plot of the loocv
dev.off()
par(mfrow=c(2,2))
par(mfrow=c(3,1))
plot.jointSurroPenalloocv(object = loocv.result.Clayton, unusedtrial = c(22, 30, 33, 38, 42,
47, 49), x = "bottomleft", y = NULL, main = "Clayton copula")
plot.jointSurroPenalloocv(object = loocv.result.Gumbel, unusedtrial = c(22, 30, 33, 38, 42,
47, 49), x = "bottomleft", y = NULL, main = "Gumbel copula")
plot.jointSurroPenalloocv(object = loocv.result.jointsurro, unusedtrial = c(22, 30, 33, 38,
42, 47, 49), x = "bottomleft", y = NULL, main = "Joint surrogate model with individual
level RE")
dev.print(device = png, file = "loocvOvarian.png", width = 650)
```

Chapter 6

Discussion générale

6.1 Conclusion sur le travail de thèse

Il était question dans cette thèse de proposer une approche de validation en une étape des critères de substitution à temps d'évènements applicable aux données individuelles des patients inclus dans des méta-analyses d'essais cliniques randomisés. Cet objectif s'inscrivait dans une démarche d'amélioration du processus de développement de nouveaux traitements pour la prise en charge des cancers. Nous avons ainsi proposé des approches statistiques basées sur les modèles conjoints à fragilités et à copules pour répondre à notre objectif de recherche.

Dans la première partie de cette thèse, nous avons développé un modèle conjoint à fragilités pour étudier le risque de survenue d'un évènement intermédiaire et le risque d'observer un évènement terminal. En plus d'ajuster le modèle sur l'indicatrice de traitement, nous y avons inclus : des effets aléatoires au niveau essai associés aux fonctions de risque de base, pour la prise en compte de l'hétérogénéité entre les essais; des effets aléatoires au niveau essai en interaction avec le traitement, pour la prise en compte de l'hétérogénéité liée aux traitements; et des effets aléatoires au niveau individuel permettant de prendre en compte l'hétérogénéité entre les sujets, due aux variables explicatives non observées dans le modèle. Introduire conjointement les effets aléatoires au niveau individuel et au niveau essai nous a permis à partir d'un seul modèle, et en une étape d'estimer les paramètres utilisés pour définir les quantités statistiques visant à évaluer les critères de substitution candidats. Ainsi, contrairement à l'approche standard (Burzykowski et al. 2001) qui tient compte des erreurs d'estimation dans son modèle de deuxième étape, la mesure d'association au niveau essai n'a plus besoin d'ajustement. Cette stratégie de modélisation a permis de réduire considérablement les problèmes de convergence souvent rencontrés dans la deuxième étape de l'approche de Burzykowski et al. (2001). Nous avons considéré plusieurs méthodes d'intégrations numériques pour approcher les intégrales

présentes dans la formulation de la log-vraisemblance marginale

Les propriétés et la robustesse du modèle ont été étudiées en simulation. En ce qui concerne les mesures d'association pour l'évaluation des critères de substitution (R^2_{trial} et τ), les résultats obtenus ont montré une robustesse face aux méthodes d'intégration et à la variation des caractéristiques des données. Le modèle était plus robuste à la mauvaise spécification, comparé à l'approche en deux étapes de Burzykowski et al. (2001) et l'approche en une étape de Rotolo et al. (2019). Globalement, les paramètres étaient estimés avec de bonnes précisions et des taux de couverture acceptables, à l'exception du R^2_{trial} pour lequel le modèle avait du mal à atteindre 90%. En revanche, les erreurs quadratiques moyennes (MSE) étaient plus faibles que celles obtenues à partir des approches existantes. Toutefois, le problème avec le taux de couverture du R^2_{trial} viendrait d'après les résultats présentés dans le Tableau 4 du Chapitre 5 du modèle de génération des données. En effet, indépendamment du modèle utilisé pour l'estimation des paramètres, lorsque les temps d'évènements sont générées à partir des copules de Clayton on a une augmentation de près de 30% du taux de couverture du R^2_{trial}. Afin d'améliorer la convergence du modèle, nous avons implémenté un algorithme permettant la gestion en interne (sans nécessiter l'intervention de l'utilisateur) des cas de non convergence. Cette stratégie opère principalement sur les paramètres de lissage, le nombre de points de quadrature lorsque la méthode d'intégration inclut la quadrature de Gauss-Hermite, et les valeurs initiales des paramètres.

Concernant le choix de la méthode d'intégration, elle doit être orientée en même temps par la convergence du modèle et le temps de calcul. Dans un premier temps, nous encourageons les utilisateurs à choisir la combinaison Monte-Carlo (pour intégrer sur les effets aléatoires au niveau essai) et quadrature de Gauss-Hermite (pour l'intégration sur les effets aléatoires au niveau individuel). Cette approche prend moins de temps que les approches globales avec seulement du Monte Carlo ou de la quadrature de Gauss-Hermite. Toutefois, les utilisateurs peuvent changer de méthode d'intégration en cas de problèmes de convergence. Lorsque plusieurs méthodes d'intégration permettent d'avoir la convergence du modèle, l'utilisateur peut baser son choix sur la qualité de l'ajustement du modèle, évaluée par le aLCV (approximate likelihood cross-validation criteria)(Commenges et al. 2007).

Afin d'obtenir la formulation du τ proposé pour le modèle, nous avons émis l'hypothèse d'homogénéité des effets du traitement entre les essais ($\sigma_{v_S} = \sigma_{v_T} = 0$). Nous avons ajusté le modèle sur les effets aléatoires au niveau individuel. Cette définition du τ est différente de celle proposée par Burzykowski et al. (2001), car ils considèrent des effets du traitement spécifiques aux essais. Par conséquent, τ ne s'interprète pas de la même façon d'une approche à l'autre. Dans l'interprétation du τ que nous proposons, les utilisateurs doivent tenir compte du fait qu'il est conditionnel aux effets aléatoires au niveau individuel. Ceci pourrait expliquer les différences observées sur les τ à partir des deux approches, aussi bien dans les études de

simulations que dans l'application.

Nous avons émis l'hypothèse d'indépendance entre les effets aléatoires u_i et (v_{S_i}, v_{T_i}), mais également entre ω_{ij} et (u_i, v_{S_i}, v_{T_i}). On peut envisager d'assouplir ces hypothèses, principalement entre les effets aléatoires au niveau essai associés au risque de base (u_i) et les effets aléatoires au niveau essai en interaction avec le traitement (v_{S_i}, v_{T_i}). A partir du modèle, nous avons défini la validation au niveau essai en utilisant la forme réduite de R^2_{trial} (voir l'équation 2.30), qui n'inclut pas les paramètres de covariances entre u_i et (v_{S_i}, v_{T_i}). Par conséquent, en assouplissant les hypothèses soulignées relatives aux effets aléatoires, R^2_{trial} ne pourrait pas être affecté «de manière significative». En outre, cela induirait une complexité du modèle pouvant aboutir à un problème d'identifiabilité, voire une augmentation des temps de calculs, sans pour autant modifier les conclusions sur l'évaluation du critère de substitution. En ce qui concerne l'hypothèse d'indépendance entre les effets aléatoires au niveau individuel et au niveau essai, de par la définition du modèle, cette hypothèse ne peut pas être assouplie, pour les mêmes raisons que précédemment.

En plus des hypothèses émises précédemment, la façon avec laquelle nous spécifions le modèle nous a permis d'éviter les problèmes d'identifiabilité. En effet, plutôt que de considérer des effets du traitement ainsi que des fonctions de risque de base spécifiques aux essais, nous prenons en compte l'hétérogénéité à l'aide des effets aléatoires. Ainsi, nous estimons les paramètres de variances des effets aléatoires plutôt que les fonctions de risque de bases et les effets fixes pour chaque essai, comme dans l'approche classique (Burzykowski et al. 2001). Cette considération nous évite d'avoir à imposer des contraintes sur le nombre de sujets par centre (il fallait au moins 3 sujets par centre dans l'application avec l'approche de Burzykowski et al. (2001) pour être en mesure d'estimer les paramètres λ_{Ti}, λ_{Si}, r_{Ti}, r_{Si}, α_i, β_i; avec au moins 1 sujet par bras de traitement et 1 sujet chez qui on a observé un évènement). En revanche, avec le modèle que nous proposons, nous pouvons relâcher les contraintes liées au nombre minimal de sujet par essai. Par contre au moins 5 essais sont requis pour garantir l'identifiabilité de notre modèle, avec au moins un sujet par bras de traitement et un évènement par critère de jugement. De plus, les résultats des simulations montrent qu'il faut plus de 10 essais pour assurer une propriété acceptable des estimateurs, en termes de taux de couverture.

Dans la deuxième partie de cette thèse, nous avons développé un package `R` convivial pour la mise en œuvre du modèle proposé précédemment. L'objectif visé était double: (a) vulgariser le modèle proposé auprès de la communauté scientifique et (b) développer des outils complémentaires pour la prédiction et l'évaluation des critères de substitution. Ainsi, nous avons mis à disposition dans le package `frailtypack` des outils d'évaluation des critères de substitution à temps d'évènements à l'aide des données issues des patients inclus dans des méta-analyse d'essais cliniques randomisés. Ces outils incluent l'effet minimum d'un critère de substitution

(STE). Le STE a été proposé par Burzykowski and Buyse (2006) pour orienter les prises de décisions quant à la validité d'un critère de substitution, au-delà de la disponibilité du R^2_{trial}. Il représente l'effet minimum du traitement à observer sur le critère de substitution pour prédire un effet significatif du traitement sur le vrai critère de jugement. Il est également possible, à partir de `frailtypack`, de prédire dans un nouvel essai l'effet du traitement sur le critère principal en utilisant l'effet du traitement observé sur le critère de substitution. Nous avons implémenté la méthode de validation croisée (loocv) pour évaluer la précision de la prédiction à partir du modèle. Nous avons également proposé des fonctions pour générer des données et conduire des études de simulation. L'idée pour cette dernière fonctionnalité étant d'orienter le choix des valeurs des arguments des fonctions en cas de problème de convergence, et d'orienter la planification de nouveaux essais cliniques (Jurgen et al. 2015) ou de nouvelles méta-analyses pour la validation des critères de substitution. Nous avons illustré nos développements à la fois sur des données réelles et sur des données simulées.

A l'issue de ces développements, nous avons rédigé un article scientifique, sous forme d'un tutoriel pour orienter les utilisateurs du package. Dans cet article, nous présentons les nouvelles fonctions du package avec une description complète de leurs arguments; nous discutons le choix des couples arguments/valeurs, et la gestion des problèmes de convergence; enfin nous accompagnons les utilisateurs dans l'interprétation des sorties. Un des avantages de notre package comparé aux programmes existants (Rotolo et al. 2018; Alonso et al. 2017) vient de sa flexibilité dans la gestion des problèmes de convergence. Toutefois, cet avantage peut très vite être limité par les temps de calculs qui peuvent être relativement longs, surtout si l'utilisateur ne dispose pas d'un ordinateur avec une puissance de calcul acceptable.

En effet, notre modèle présente l'inconvénient d'être assez gourmand en temps de calcul. Sur une machine multiprocesseurs avec `40 coeurs` et `378 Go de RAM`, l'exemple sur le cancer gastrique avec 3288 sujets et 14 essais mettait `9 minutes` pour l'estimation des paramètres tandisque le même exemple sur une machine avec `4 coeurs` et `16 Go de RAM` tournait pendant environ `une heure`. Ce phénomène est d'autant plus important que la taille des essais (ou de la population) est élevée et la puissance de calcul de l'ordinateur utilisé est faible. Cette difficulté vient de la présence dans la log-vraisemblance du modèle des intégrales multiples qui ne pouvaient pas être calculées analytiquement. Par conséquent, la gestion des cas de non convergences par l'utilisateur demanderait d'attendre beaucoup de temps avant de relancer le modèle. Ce qui risquerait de ne pas convenir à un utilisateur habitué aux sorties du modèle quelques secondes après soumission de la commande. En revanche, nous avons entrepris des mesures permettant de minimiser au mieux les temps de calculs. (a) nos programmes sont écrits en `Fortran`, qui est un langage de haut niveau (proche du langage naturel) adapté au calcul scientifique et qui dispose des compilateurs performants, produisant des exécutables

très rapides. (b) nos algorithmes pour le calcul des intégrales ont été parallélisés en utilisant l'interface de programmation `OpenMP`. Ce paradigme de calcul parallèle est à mémoire partagée et permet d'accélérer les calculs, suivant les ressources (mémoires, cœurs de calculs) disponibles de l'ordinateur. Ainsi, plus élevé sera le nombre de cœurs plus rapidement les calculs pourront s'effectuer.

Une autre façon de mieux optimiser nos codes de calculs serait d'implémenter au sein de `frailtypack` le paradigme de calcul parallèle `MPI` (Message Passing Interface). En effet, contrairement aux programmes `OpenMP` qui ne peuvent s'exécuter que sur une machine dans laquelle toutes les tâches sont gérées par un seul processus, une application `MPI` est un ensemble de processus autonomes exécutant chacun leur propre code et communiquant via des appels à des sous-programmes de la bibliothèque `MPI`. La communication peut se faire entre des ordinateurs distants ou dans un ordinateur multiprocesseur. De plus, pour des utilisateurs ayant accès à des grappes de machines indépendantes multi-cœur à mémoire partagée (clusters de calcul), la programmation hybride combinant `MPI` et `OpenMP` permettrait de mieux tirer profit des ressources disponibles pouvant aller jusqu'à rendre instantanée l'exécution des programmes.

Dans la troisième partie de cette thèse, nous avons proposé un nouveau modèle conjoint, combinant les effets aléatoires et les copules, pour la validation des critères de substitution. En effet, la complexité des temps de calcul discutée précédemment émane en grande partie de la prise en compte de l'hétérogénéité au niveau individuel par les effets aléatoires ω_{ij}. De ce fait, la corrélation entre le critère de substitution et le critère de jugement principal est mesurée à l'aide de la variance des effets aléatoires, θ et du paramètre de puissance ζ. Cette spécification du modèle exige non seulement d'estimer deux paramètres associés aux effets aléatoires mais également d'intégrer sur ω_{ij}. Dans ce nouveau travail, nous avons proposé de prendre en compte la dépendance entre le critère de substitution et le critère de jugement principal par une fonction copule (Nelsen 2006; Prenen et al. 2017). Par conséquent seul le paramètre de copule est estimé, sans plus avoir besoin d'intégrales au niveau individuel. Un autre avantage avec ce modèle vient de sa proximité avec le modèle en deux étapes de Burzykowski et al. (2001), et précisément avec la définition du τ de Kendall qui a une interprétation semblable à celui proposé par les auteurs. Comme précédemment, ce modèle a été inclue dans `frailtypack` afin de vulgariser son utilisation. La nouvelle version du package incluant le modèle sera disponible très prochainement sur le CRAN. En revanche, elle est disponible sur le compte Git Hub de Casimir Sofeu : https://github.com/socale/frailtypack/tree/surrogacy_ submitted_3-0-4_sansMPI

L'évaluation des performances des estimateurs a révélé des résultats satisfaisants et une meilleure stabilité du nouveau modèle comparé au modèle précédent, surtout dans des situations de très forte association au niveau individuel. Indépendamment des mesures d'association

pour l'évaluation des critères de substitution, le modèle était assez robuste en ce qui concerne les méthodes d'intégration, la variation des caractéristiques des données et la variation des valeurs assignées aux arguments de la fonction de simulation. Par ailleurs, comparé au modèle précédent, le modèle basé sur les copules de Clayton avait tendance à mieux estimer le τ de Kendall. Avec ces résultats, nous recommandons en première intention de baser la validation des critères de substitution sur le modèle conjoint à fragilités et à copules. Afin d'avoir une formulation simplifiée de la log-vraisemblance telle que décrite en appendice de l'article présenté dans le Chapitre 5, nous avons considéré deux types de copules, celle de Clayton et celle de Gumbel. Il serait intéressant d'étendre le modèle à d'autres types de copules, afin de retenir suivant la situation qui se présente la fonction de copule qui s'ajusterait le mieux aux données.

6.2 Autres avantages et limites

Globalement, que l'on soit dans un contexte de modèle conjoint à fragilités ou de modèle conjoint à fragilités et à copule, les approches développées ont pour avantage de faire de la validation des critères de substitution en une étape, d'être assez robustes et de considérer des fonctions de risque de base flexibles utilisant des splines. Cette considération évite de faire une hypothèse supplémentaire sur la forme paramétrique des fonctions de risque de base. Nous avons proposé deux modèles différents, pouvant venir en complément des approches existantes (Burzykowski et al. 2001; Rotolo et al. 2019; Renfro et al. 2012; Buyse et al. 2016) pour améliorer la validation des critères de substitution. De plus, tous nos programmes sont Open source, ce qui facilite la reproductibilité de nos résultats et donne la possibilité aux chercheurs d'écrire des programmes dérivés afin d'étendre le champ de la recherche sur les méthodes développées.

En revanche, en plus du temps de calcul, nos modèles présentent l'inconvénient d'être sensibles aux méta-analyses avec un nombre réduit d'essais. Il s'agit d'un problème réel, souvent rencontré en pratique. On observe parfois des essais avec de faibles effectifs, pendant que d'autres excellent en nombre de sujets. Par exemple, dans la méta-analyse sur le cancer gastrique adjuvant (The GASTRIC Group 2010) qui incluait 14 essais et 3288 patients, 25% des centres avaient inclus plus de 270 sujets chacun pendant que le premier quartile était à 181 sujets, et le plus petit centre recensait 88 sujets. Face à ce déséquilibre, on peut s'interroger sur une approche de reconstitution aléatoire des centres, de façon à conserver non seulement la randomisation, mais également de maintenir une certaine homogénéité intra et bien entendu une hétérogénéité inter essais. L'idée étant d'augmenter le nombre d'essais ou de centres pour une meilleure estimation du R^2_{trial}. En perspective à cette thèse, nous étudierons la plausibilité d'une telle approche en simulation.

6.3 Discussions supplémentaires et perspectives

Kassaï et al. (2007) ont remis en cause les approches de validation purement descriptives (comme celles disponibles à ce jour) suite à leur impossibilité à prédire la toxicité des médicaments et par conséquent d'estimer le rapport bénéfice/risque avec les critères intermédiaires. Par ailleurs, Matulonis et al. (2015) discutent la difficulté d'utiliser certains critères intermédiaires à l'exemple de la PFS, comme critère de substitution. En effet, avec la survie post-progression qui peut être relativement longue, l'effet du traitement observé sur la PFS peut être compromis par l'introduction de nouveaux traitement à la suite d'un échec thérapeutique ou d'une toxicité. De plus, la toxicité à long terme liée à un traitement pris sur le court terme peut affecter l'OS et d'autres critères de substitution considérés après la progression, mais pas la PFS. Cette critique interpelle les chercheurs sur la nécessité de prendre en compte la toxicité des traitements (à travers les échecs thérapeutiques) au moment de la validation des critères de substitution. Une façon élégante d'intégrer ces informations avec notre approche serait de considérer un modèle trivarié (ou multivarié) à fragilités et à copule. Dans un tel modèle, la troisième équation serait destinée au risque de survenue d'un échec thérapeutique ou d'un évènement intermédiaire comme une récidive après une progression ou une toxicité. La prédiction de l'effet du traitement sur le critère de jugement principal serait de ce fait conditionnelle à l'effet du traitement sur le critère intermédiaire. Un tel modèle permettrait également de prédire sous certaines conditions la toxicité du traitement à partir de son effet sur le critère de substitution. La fonction de copule devrait prendre en compte la dépendance entre le critère de jugement principal et chaque critère de substitution potentiel. Cela nécessite donc un modèle de copule multivarié, avec au moins 2 paramètres de copule. A ce jour, la paramétrisation qui remplit ces critères est la copule de vine (Barthel et al. 2018). Nous travaillerons très prochainement sur le développement de ce modèle. Une autre piste serait également de combiner l'évolution de la taille tumorale et l'apparition de nouvelles lésions pour définir des nouveaux critères de substitution associés à la survie globale. Il s'agit ici de développer des méthodes de validation de critères de substitution pour données récurrentes ou des marqueurs longitudinaux.

Dans nos modèles, nous avons considéré exclusivement la censure à droite, ce qui suppose connues les dates exactes de survenue des évènements d'intérêt. Généralement dans les études, les visites de suivi sont planifiées à l'avance. Par conséquent les dates exactes de survenue des évènements intermédiaires ne sont pas toujours connues, nous plaçant dans un contexte de censure par intervalle. Il serait intéressant de considérer la censure par intervalle, dans les évolutions du package. Nous prévoyons également d'étendre le modèle conjoint à fragilités présenté dans le Chapitre 3 pour prendre en compte des facteurs pronostiques autres que le traitement.

La définition de nos modèles s'appuie sur les modèles à risque proportionnel. Il peut arriver en pratique que l'hypothèse des risques proportionnels ne soit pas vérifiée pour la variable traitement, sur au moins un des critères de jugement. Ce qui est conseillé dans ce cas, compte tenu de la nécessité d'estimer les effets du traitement sur les critères de jugement (Commenges and Jacqmin-Gadda 2015) (a) c'est de modéliser les variations du risque relatif en fonction du temps en introduisant une interaction avec le temps dans le modèle. Avec la présence de l'interaction entre l'indicatrice de traitement et les effets aléatoires au niveau essai dans nos modèles, cette solution pourrait engendrer des problèmes d'identifiabilité du modèle; en plus elle n'est pas adaptée à la validation d'un critère de substitution et par conséquent n'est pas envisageable. (b) La deuxième solution serait de considérer la variable telle quelle au cas où la présence de l'interaction avec le temps ne modifie pas considérablement la relation entre la variable et le critère de jugement considéré. (c) Dans le cas où la violation de l'hypothèse de proportionnalité des risques s'avère évidente, plutôt que le modèle àrisque proportionnel, considérer d'autres type de modèles à tel que le modèle à risque additif, décrit à la section `4.7` du livre de Commenges and Jacqmin-Gadda (2015). Le modèle additif a l'avantage d'être complètement non paramétrique. Ce modèle serait donc une alternative à explorer dans le cadre de la validation des critères de substitution.

Une autre perspective pour ce travail est de développer les modèles proposés dans un cadre bayésien. On pourra s'inspirer des travaux de Li et al. (2019) qui ont repris le modèle conjoint à fragilités et à copule pour données récurrentes et un évènement terminal en bayésien. Cette considération permettra par exemple de réduire les temps de calcul en cas de tailles d'échantillons élevées tout en garantissant la convergence des modèles. Nous pensons également à étendre le concept de validation en une étape des critères de substitution à d'autres type de critères de jugement, comme par exemple des marqueurs longitudinaux ou encore des variables binaires. Cette considération permettra d'éviter des problèmes comme la non disponibilité des erreurs d'estimation des effets du traitement à l'issue du modèle de première étape pour ajustement dans le modèle de deuxième étape comme discuté par Burzykowski et al. (2019) lors que la validation de la taille tumorale comme critère de substitution pour l'OS dans le cancer colorectal. Une autre perspective serait d'étendre le concept de validation en une étape aux modèles d'inférence causale sur données de méta-analyses, afin d'adopter une interprétation causale de la relation entre le critère de substitution et le critère de jugement principal au-delà des essais considérés.

6.4 Démarche pour valider un critère de substitution

Afin de valider un critère de substitution de type survie, nous recommandons la démarche suivante qui s'appuie sur les approches développées dans cette thèse et les recommandations de Institute for Quality and Efficiency in Health Care (2011):

1. Disposer des données individuelles de méta-analyses issues des patients suivis dans le cadre de la pathologie étudiée
2. Estimer les critères d'évaluation du critère de substitution à l'aide des modèles conjoints à fragilités et à copules : tester les deux fonctions de copules.
3. Faire de même pour le modèle conjoint à fragilités.
4. En cas de problème de convergence dans l'un des 3 modèles, traiter le problème en jouant sur les arguments des fonctions comme décrit dans le Chapitre 4.
5. Pour tous les modèles qui auraient convergé, les comparer à l'aide du critère $aLCV$, puis retenir celui qui s'ajuste le mieux aux données.
6. Utiliser la valeur estimée du τ de Kendall, du R^2_{trial} et du STE pour décider si le critère de substitution peut être considérer comme valide ou non suivant l'algorithme :
   - Pour un τ de Kendall $>= 0,6$ accepter la validité au niveau individuel ;
   - pour $R^2_{trial} > 0,72$, accepter la validité au niveau essai si l'effet du traitement sur le critère de substitution est statistiquement significatif ;
   - si $0,49 < R^2_{trial} < 0,72$ alors
     - si l'IC à 90% ou à 80% du STE $<$ à l'effet du traitement sur le critère de substitution, accepter la validité au niveau essai si l'effet du traitement sur le critère de substitution est statistiquement significatif ;
     - sinon, rejeter la validité au niveau essai.
   - Pour tous les autre cas, rejeter la validité au niveau essai ou au niveau individuel
   - Si la validité au niveau individuel et au niveau essai sont acceptées, alors le critère de substitution peut être considéré comme valide.

6.5 Conclusion générale

Dans ce travail de thèse, nous avons développé des méthodes statistiques pour la validation en une étape des critères de substitution à temps d'évènement, en utilisant des données issues

des méta-analyses d'essais cliniques randomisés. La première méthode s'appuyait sur un modèle conjoint à fragilités partagées, pendant que le second s'appuyait sur un modèle à fragilités partagées et à copules. La validation des critères de substitution s'est faite à partir du τ de Kendall pour le niveau individuel et du R^2_{trial} pour le niveau essai. Il s'agit d'outils statistiques supplémentaires de validation permettant de réduire les problèmes de convergence et d'optimisation rencontrés dans l'approche standard. Tous les modèles ont été implémentés dans un package `R` pour ainsi faciliter la reproductibilité des résultats, favoriser la vulgarisation des méthodes auprès de la communauté scientifique et faciliter les développements futurs. Nous avons donné la possibilité d'ajuster le modèle conjoint à fragilités partagées et à copules sur des facteurs pronostiques autres que le traitement. Par conséquent, ce modèle peut être utilisé dans un contexte autre que la validation des critères de substitution. On pourra par exemple l'utiliser dans un contexte de risque compétitif pour étudier les facteurs associés à un critère de jugement intermédiaire, en prenant en compte la censure informative par un évènement terminal. Les modèles développés peuvent également être utilisés dans un domaine autre que le cancer pour la validation des critères de substitution. Ce travail est motivé par le besoin d'accélérer la mise en place des essais cliniques et par conséquent de raccourcir les durées de mise sur le marché des traitements jugés efficaces.

Bibliography

A. Alonso, W. V. der Elst, G. Molenberghs, M. Buyse, and T. Burzykowski. An information-theoretic approach for the evaluation of surrogate endpoints based on causal inference. *Biometrics*, 72(3):669–677, 2016. doi: 10.1111/biom.12483. URL http://dx.doi.org/10.1111/biom.12483.

A. Alonso, T. Bigirumurame, T. Burzykowski, M. Buyse, G. Molenberghs, L. Muchene, N. J. Perualila, Z. Shkedy, and W. V. d. Elst. *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. Chapman and Hall/CRC, 2017.

A. Alonso, P. Meyvisch, W. V. der Elst, G. Molenberghs, and G. Verbeke. A reflection on the possibility of finding a good surrogate. *Journal of Biopharmaceutical Statistics*, 29(3): 468–477, 2019. doi: 10.1080/10543406.2018.1559854. URL https://doi.org/10.1080/10543406.2018.1559854. PMID: 30686082.

S. G. Baker. Five criteria for using a surrogate endpoint to predict treatment effect based on data from multiple previous trials. *Statistics in Medicine*, 37(4):507–518, 2 2018. doi: 10.1002/sim.7561. URL http:https://doi.org/10.1002/sim.7561.

S. G. Baker and B. S. Kramer. A perfect correlate does not a surrogate make. *BMC Medical Research Methodology*, 3(1):16, Sep 2003. doi: 10.1186/1471-2288-3-16. URL https://doi.org/10.1186/1471-2288-3-16.

N. Barthel, C. Geerdens, M. Killiches, P. Janssen, and C. Czado. Vine copula based likelihood estimation of dependence patterns in multivariate event time data. *Computational Statistics & Data Analysis*, 117:109 – 127, 2018. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2017.07.010. URL http://www.sciencedirect.com/science/article/pii/S0167947317301688.

C. A. Bellera, M. Pulido, S. Gourgou, L. Collette, A. Doussau, A. Kramar, T. Dabakuyo, M. Ouali, A. Auperin, T. Filleron, C. Fortpied, C. Le Tourneau, X. Paoletti, M. Mauer, S. Mathoulin-Pélissier, and F. Bonnetain. Protocol of the definition for the assessment of

time-to-event endpoints in cancer trials (datecan) project: formal consensus method for the development of guidelines for standardised time-to-event endpoints' definitions in cancer clinical trials. *European Journal of Cancer*, 49(4):769 – 781, 2013. doi: 10.1016/j.ejca.2012.09.035.

C. A. Bellera, N. Penel, M. Ouali, S. Bonvalot, P. G. Casali, O. S. Nielsen, M. Delannes, S. Litière, F. Bonnetain, T. S. Dabakuyo, R. S. Benjamin, J.-Y. Blay, B. N. Bui, F. Collin, T. F. Delaney, F. Duffaud, T. Filleron, M. Fiore, H. Gelderblom, S. George, R. Grimer, P. Grosclaude, A. Gronchi, R. Haas, P. Hohenberger, R. Issels, A. Italiano, V. Jooste, A. Krarup-Hansen, C. Le Péchoux, C. Mussi, O. Oberlin, S. Patel, S. Piperno-Neumann, C. Raut, I. Ray-Coquard, P. Rutkowski, S. Schuetze, S. Sleijfer, E. Stoeckle, M. Van Glabbeke, P. Woll, S. Gourgou-Bourgade, and S. Mathoulin-Pélissier. Guidelines for time-to-event end point definitions in sarcomas and gastrointestinal stromal tumors (GIST) trials: results of the DATECAN initiative (Definition for the Assessment of Time-to-event Endpoints in CANcer trials)†. *Annals of Oncology*, 26(5):865–872, 07 2014. ISSN 0923-7534. doi: 10.1093/annonc/mdu360. URL `https://doi.org/10.1093/annonc/mdu360`.

S. Branchoux, C. Bellera, A. Italiano, D. Rustand, A.-F. Gaudin, and V. Rondeau. 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, 2019. ISSN 1040-8428. doi: https://doi.org/10.1016/j.critrevonc.2019.02.013. URL `http://www.sciencedirect.com/science/article/pii/S1040842819300447`.

T. Burzykowski and M. Buyse. Surrogate threshold effect: An alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics*, 5(3):173–186, 2006. doi: 10.1002/pst.207. URL `http://dx.doi.org/10.1002/pst.207`.

T. Burzykowski, G. Molenberghs, M. Buyse, H. Geys, and D. Renard. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *Journal of the Royal Statistical Society C (Applied Statistics)*, 50(4):405–422, 2001. doi: 10.1111/1467-9876.00244. URL `http://dx.doi.org/10.1111/1467-9876.00244`.

T. Burzykowski, G. Molenberghs, M. Buyse, and H. Geys. *The Evaluation of Surrogate Endpoints*. Springer-Verlag, New-york, NK, 2005.

T. Burzykowski, E. Coart, E. D. Saad, Q. Shi, D. W. Sommeijer, C. Bokemeyer, E. Díaz-Rubio, J.-Y. Douillard, A. Falcone, C. S. Fuchs, R. M. Goldberg, J. R. Hecht, P. M. Hoff, H. Hurwitz, F. F. Kabbinavar, M. Koopman, T. S. Maughan, C. J. A. Punt, L. Saltz, H.-J. Schmoll, M. T. Seymour, N. C. Tebbutt, C. Tournigand, E. Van Cutsem, A. de Gramont,

J. R. Zalcberg, M. Buyse, and for the Aide et Recherche en Cancerologie Digestive Group. Evaluation of Continuous Tumor-Size–Based End Points as Surrogates for Overall Survival in Randomized Clinical Trials in Metastatic Colorectal Cancer. *JAMA Network Open*, 2(9): e1911750–e1911750, 09 2019. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2019.11750. URL `https://doi.org/10.1001/jamanetworkopen.2019.11750`.

M. Buyse and G. Molenberghs. The validation of surrogate endpoints in randomized experiments. *Biometrics*, 54:1014–1029, 1998.

M. Buyse, G. Molenberghs, T. Burzykowski, D. Renard, and H. Geys. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, 1(1):49–67, 2000.

M. Buyse, G. Molenberghs, X. Paoletti, K. Oba, A. Alonso, W. V. der Elst, and T. Burzykowski. Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. *Biometrical Journal*, 58(1):104–132, 2016. doi: 10.1002/bimj.201400049. URL `http://dx.doi.org/10.1002/bimj.201400049`.

CIRC. Dernières données mondiales sur le cancer : le fardeau du cancer atteint 18,1 millions de nouveaux cas et 9,6 millions de décès par cancer en 2018. *COMMUNIQUE DE PRESSE Numéro 263*, 2018. URL `https://www.iarc.fr/wp-content/uploads/2018/09/pr263_F.pdf`.

D. Commenges and H. Jacqmin-Gadda. *Modèles biostatistiques pour l'épidémiologie*. de Boeck, Oct. 2015. URL `https://hal.inria.fr/hal-01580144`.

D. Commenges, P. Joly, A. Gégout-petit, and B. Liquet. Choice between semi-parametric estimators of markov and non-markov multi-state models from coarsened observations. *Scandinavian Journal of Statistics*, 34(1):33–52, 2007. ISSN 03036898, 14679469. URL `http://www.jstor.org/stable/41548537`.

L. Duchateau and P. Janssen. *The Frailty Model*. Springer-Verlag, 2008.

E. Eisenhauer, P. Therasse, J. Bogaerts, L. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European Journal of Cancer*, 45(2):228 – 247, 2009. doi: 10.1016/j.ejca.2008.10.026.

T. Emura, M. Nakatochi, K. Murotani, and V. Rondeau. A joint frailty-copula model between tumour progression and death for meta-analysis. *Statistical Methods in Medical Research*,

26(6):2649–2666, 2017. doi: 10.1177/0962280215604510. URL https://doi.org/10.1177/0962280215604510.

J. Ferlay, E. Steliarova-Foucher, J. Lortet-Tieulent, S. Rosso, J. Coebergh, H. Comber, D. Forman, and F. Bray. Cancer incidence and mortality patterns in europe: estimates for 40 countries in 2012. *European Journal of Cancer*, 49(6):1374 – 1403, 2013. doi: 10.1016/j.ejca.2012.12.027.

F. Fiteni, V. Westeel, X. Pivot, C. Borg, D. Vernerey, and F. Bonnetain. Endpoints in cancer clinical trials. *Journal of Visceral Surgery*, 151(1):17 – 22, 2014. ISSN 1878-7886. doi: https://doi.org/10.1016/j.jviscsurg.2013.10.001. URL http://www.sciencedirect.com/science/article/pii/S1878788613001318.

T. R. Fleming, R. L. Prentice, M. S. Pepe, and D. Glidden. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and aids research. *Statistics in Medicine*, 13(9):955–968, 1994. doi: 10.1002/sim.4780130906. URL http://dx.doi.org/10.1002/sim.4780130906.

Fondation contre le cancer. Traitements du cancer. *Ressource informatique [consulter le 26/09/2019]*, 2019. URL https://www.cancer.be/le-cancer/traitements-du-cancer.

G. A. Fredricks and R. B. Nelsen. On the relationship between spearman's rho and kendall's tau for pairs of continuous random variables. *Journal of Statistical Planning and Inference*, 137(7):2143 – 2150, 2007. doi: http://dx.doi.org/10.1016/j.jspi.2006.06.045. URL http://www.sciencedirect.com/science/article/pii/S0378375806002588.

L. S. Freedman, B. I. Graubard, and A. Schatzkin. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, 11(2):167–178, 1992. doi: 10.1002/sim.4780110204. URL http://dx.doi.org/10.1002/sim.4780110204.

Global Burden of Disease Cancer Collaboration. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study. *JAMA Oncology*, 3(4):524–548, 04 2017. ISSN 2374-2437. doi: 10.1001/jamaoncol.2016.5688. URL https://doi.org/10.1001/jamaoncol.2016.5688.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. ISSN 01621459. URL http://www.jstor.org/stable/2289064.

Institute for Quality and Efficiency in Health Care. Validity of surrogate endpoints in oncology: Executive summary, 2011. URL `www.iqwig.de/download/A10-05_Executive_Summary_v1-1_Surrogate_endpoints_in_oncology.pdf`.

P. Joly, D. Commenges, and L. Letenneur. A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia. *Biometrics*, 54(1):185–194, 1998. URL `http://www.jstor.org/stable/2534006`.

H. Jurgen, W. Song, and K. John. Using simulation to optimize adaptive trial designs: applications in learning and confirmatory phase trials. *Clinical Investigation*, 5(4):401–413, 2015. doi: 10.4155/CLI.15.14. URL `https://www.openaccessjournals.com/articles/using-simulation-to-optimize-adaptive-trial-designs-applications-in-learning\-and-confirmatory-phase-trials.pdf`.

B. Kassaï, F. Guyffier, and J.-P. Boissel. Critères intermédiaires et critères de substitution. *Médecine thérapeutique*, 13(4):279 – 286, 2007. doi: 10.1684/met.2007.0117. URL `https://www.jle.com/fr/revues/met/e-docs/criteres_intermediaires_et_criteres_de_substitution_275939/article.phtml`.

E. L. Korn, P. S. Albert, and L. M. McShane. Assessing surrogates as trial endpoints using mixed models. *Statistics in Medicine*, 24(2):163–182, 1 2005. ISSN 1097-0258. doi: 10.1002/sim.1779. URL `https://doi.org/10.1002/sim.1779`.

A. Król, A. Mauguen, Y. Mazroui, A. Laurent, S. Michiels, and V. Rondeau. Tutorial in joint modeling and prediction: A statistical software for correlated longitudinal outcomes, recurrent events and a terminal event. *Journal of Statistical Software, Articles*, 81(3):1–52, 2017. doi: 10.18637/jss.v081.i03. URL `https://www.jstatsoft.org/v081/i03`.

Z. Li, V. M. Chinchilli, and M. Wang. A bayesian joint model of recurrent events and a terminal event. *Biometrical Journal*, 61(1):187–202, 1 2019. ISSN 0323-3847. doi: 10.1002/bimj.201700326. URL `https://doi.org/10.1002/bimj.201700326`.

D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963. doi: 10.1137/0111030. URL `https://doi.org/10.1137/0111030`.

U. A. Matulonis, A. M. Oza, T. W. Ho, and J. A. Ledermann. Intermediate clinical endpoints: A bridge between progression-free survival and overall survival in ovarian cancer trials. *Cancer*, 121(11):1737–1746, 2015. doi: 10.1002/cncr.29082. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/cncr.29082`.

S. Michiels, A. L. Maître, M. Buyse, T. Burzykowski, E. Maillard, J. Bogaerts, J. B. Vermorken, W. Budach, T. F. Pajak, K. K. Ang, J. Bourhis, and J.-P. Pignon. Surrogate endpoints for overall survival in locally advanced head and neck cancer: Meta-analyses of individual patient data. *The Lancet Oncology*, 10(4):341 – 350, 2009. doi: https://doi.org/10.1016/S1470-2045(09)70023-3. URL `http://www.sciencedirect.com/science/article/pii/S1470204509700233`.

G. Molenberghs, M. Buyse, H. Geys, D. Renard, T. Burzykowski, and A. Alonso. Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials*, 23(6):607 – 625, 2002. ISSN 0197-2456. doi: https://doi.org/10.1016/S0197-2456(02)00236-2. URL `http://www.sciencedirect.com/science/article/pii/S0197245602002362`.

R. B. Nelsen. *An introduction to Copulas*. Springer, 2006.

F. O'Sullivan. Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing*, 9(2):363–379, 1988. doi: 10.1137/0909024. URL `https://doi.org/10.1137/0909024`.

X. Paoletti, F. Rotolo, and S. Michiels. Quelles exigences pour qu'un biomarqueur puisse être un critère de substitution acceptable ? *Bulletin du Cancer*, 103(6, Supplement 1):S63 – S70, 2016. ISSN 0007-4551. doi: https://doi.org/10.1016/S0007-4551(16)30147-3. URL `http://www.sciencedirect.com/science/article/pii/S0007455116301473`. 3e Congrès de la Société Française du Cancer - Décrypter et partager les avancées en oncologie - Paris Palais des congrès - 29 juin - 1er juillet 2016.

J. Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pages 411–420, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-800-1. URL `http://dl.acm.org/citation.cfm?id=2074022.2074073`.

L. Prenen, R. Braekers, and L. Duchateau. Extending the archimedean copula methodology to model multivariate survival data grouped in clusters of variable size. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):483–505, 2017. doi: 10.1111/rssb.12174. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12174`.

R. L. Prentice. Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, 8(4):431–440, 1989. doi: 10.1002/sim.4780080407. URL `http://dx.doi.org/10.1002/sim.4780080407`.

J. O. Ramsay. Monotone regression splines in action. *Statist. Sci.*, 3(4):425–441, 1988. doi: 10.1214/ss/1177012761. URL `https://doi.org/10.1214/ss/1177012761`.

L. A. Renfro, Q. Shi, D. J. Sargent, and B. P. Carlin. Bayesian adjusted r2 for the meta-analytic evaluation of surrogate time-to-event endpoints in clinical trials. *Statistics in Medicine*, 31 (8):743–761, 2012. doi: 10.1002/sim.4416. URL `http://dx.doi.org/10.1002/sim.4416`.

V. Rondeau, D. Commenges, and P. Joly. Maximum penalized likelihood estimation in frailty models. *Lifetime Data Analysis*, 9(2):139–153, 2003.

V. Rondeau, S. Mathoulin-Pelissier, H. Jacqmin-Gadda, V. Brouste, and P. Soubeyran. Joint frailty models for recurring events and death using maximum penalized likelihood estimation: Application on cancer events. *Biostatistics*, 8(4):708–721, 2007. doi: 10.1093/biostatistics/kxl043. URL `http://biostatistics.oxfordjournals.org/content/8/4/708.abstract`.

V. Rondeau, S. Michiels, B. Liquet, and J. P. Pignon. Investigating trial and treatment heterogeneity in an individual patient data meta-analysis of survival data by means of the penalized maximum likelihood approach. *Statistics in Medicine*, 27(11):1894–1910, 2008. doi: 10.1002/sim.3161. URL `http://dx.doi.org/10.1002/sim.3161`.

V. Rondeau, J. P. Pignon, and S. Michiels. A joint model for the dependence between clustered times to tumour progression and deaths: A meta-analysis of chemotherapy in head and neck cancer. *Statistical Methods in Medical Research*, 24(6):711–729, 2015. doi: 10.1177/0962280211425578. URL `http://hinarilogin.research4life.org/uniquesigdx.doi.org/uniquesig0/10.1177/0962280211425578`.

F. Rotolo, X. Paoletti, and S. Michiels. `surrosurv`: An `R` package for the evaluation of failure time surrogate endpoints in individual patient data meta-analyses of randomized clinical trials. *Computer Methods and Programs in Biomedicine*, 155:189 – 198, 2018. doi: https://doi.org/10.1016/j.cmpb.2017.12.005. URL `http://www.sciencedirect.com/science/article/pii/S0169260717302316`.

F. Rotolo, X. Paoletti, T. Burzykowski, M. Buyse, and S. Michiels. A poisson approach to the validation of failure time surrogate endpoints in individual patient data meta-analyses. *Statistical Methods in Medical Research*, 28(1):170–183, 2019. doi: 10.1177/0962280217718582. URL `https://doi.org/10.1177/0962280217718582`.

M. Savina, S. Gourgou, A. Italiano, D. Dinart, V. Rondeau, N. Penel, S. Mathoulin-Pelissier, and C. Bellera. Meta-analyses evaluating surrogate endpoints for overall survival in cancer randomized trials: A critical review. *Critical Reviews in Oncology/Hematology*, 123:21 – 41,

2018. ISSN 1040-8428. doi: https://doi.org/10.1016/j.critrevonc.2017.11.014. URL http://www.sciencedirect.com/science/article/pii/S1040842817302184.

Q. Shi, L. A. Renfro, B. M. Bot, T. Burzykowski, M. Buyse, and D. J. Sargent. Comparative assessment of trial-level surrogacy measures for candidate time-to-event surrogate endpoints in clinical trials. *Computational Statistics & Data Analysis*, 55(9):2748 – 2757, 2011. doi: https://doi.org/10.1016/j.csda.2011.03.014. URL http://www.sciencedirect.com/science/article/pii/S0167947311001058.

C. L. Sofeu, T. Emura, and V. Rondeau. 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, 2019. doi: 10.1002/sim.8162. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8162.

J. M. G. Taylor, Y. Wang, and R. Thiébaut. Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics*, 61(4):1102–1111, 2005. doi: 10.1111/j.1541-0420.2005.00380.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2005.00380.x.

The GASTRIC Group. Benefit of adjuvant chemotherapy for resectable gastric cancer: A meta-analysis. *JAMA*, 303(17):1729–1737, 2010. doi: 10.1001/jama.2010.534. URL +http://dx.doi.org/10.1001/jama.2010.534.

P. Therasse, S. G. Arbuck, E. A. Eisenhauer, J. Wanders, R. S. Kaplan, L. Rubinstein, J. Verweij, M. Van Glabbeke, A. T. van Oosterom, M. C. Christian, and S. G. Gwyther. New Guidelines to Evaluate the Response to Treatment in Solid Tumors. *JNCI: Journal of the National Cancer Institute*, 92(3):205–216, 02 2000. ISSN 0027-8874. doi: 10.1093/jnci/92.3.205. URL https://doi.org/10.1093/jnci/92.3.205.

H. C. van Houwelingen, L. R. Arends, and T. Stijnen. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in medicine*, 21 4:589–624, 2002.

S. Vandenberghe, L. Duchateau, L. Slaets, J. Bogaerts, and S. Vansteelandt. Surrogate marker analysis in cancer clinical trials through time-to-event mediation techniques. *Statistical Methods in Medical Research*, 27(11):3367–3385, 2018. doi: 10.1177/0962280217702179. URL https://doi.org/10.1177/0962280217702179. PMID: 28425345.

WHO. Who handbook for reporting results of cancer treatment. world health organization. 1979. URL https://apps.who.int/iris/handle/10665/37200.

Annexe

Annexe Démonstration de la présence des biais d'estimations sur R^2_{trial} non ajusté

$$
\begin{aligned}
Corr(\hat{\alpha_i}, \hat{\beta_i}) &= Corr(\alpha_i + \epsilon_{ai}, \beta_i + \epsilon_{bi}) \\
&= \frac{Cov(\alpha_i + \epsilon_{ai}, \beta_i + \epsilon_{bi})}{\sqrt{Var(\alpha_i + \epsilon_{ai})}\sqrt{Var(\beta_i + \epsilon_{bi})}} \\
&= \frac{Cov(\alpha_i, \beta_i) + Cov(\alpha_i, \epsilon_{bi}) + Cov(\epsilon_{ai}, \beta_i) + Cov(\epsilon_{ai}, \epsilon_{bi})}{\sqrt{Var(\alpha + a_i) + Var(\epsilon_{ai})}\sqrt{Var(\beta + b_i) + Var(\epsilon_{bi})}} \\
&= \frac{Cov(\alpha_i, \beta_i) + Cov(\epsilon_{ai}, \epsilon_{bi})}{\sqrt{d_{aa} + \sigma_{aa}}\sqrt{d_{bb} + \sigma_{bb}}} \\
&= Corr(\alpha_i, \beta_i) \times \frac{\sqrt{d_{aa}}}{\sqrt{d_{aa} + \sigma_{aa}}} \frac{\sqrt{d_{bb}}}{\sqrt{d_{bb} + \sigma_{bb}}} + \frac{\sqrt{\sigma_{ab}}}{\sqrt{d_{aa} + \sigma_{aa}}\sqrt{d_{bb} + \sigma_{bb}}} \\
&= Corr(\alpha_i, \beta_i) \times \frac{1}{\sqrt{1 + \frac{\sigma_{aa}}{d_{aa}}}} \frac{1}{\sqrt{1 + \frac{\sigma_{bb}}{d_{bb}}}} + \frac{\sqrt{\sigma_{ab}}}{\sqrt{d_{aa} + \sigma_{aa}}\sqrt{d_{bb} + \sigma_{bb}}} \\
&= \frac{Corr(\alpha_i, \beta_i)}{\sqrt{1 + k_a}\sqrt{1 + k_b}} + \frac{\rho\sqrt{\sigma_{aa}}\sqrt{\sigma_{bb}}}{\sqrt{d_{aa} + \sigma_{aa}}\sqrt{d_{bb} + \sigma_{bb}}} \\
&= \frac{Corr(\alpha_i, \beta_i)}{\sqrt{1 + k_a}\sqrt{1 + k_b}} + \frac{\rho}{\sqrt{1 + \frac{d_{aa}}{\sigma_{aa}}}\sqrt{1 + \frac{d_{bb}}{\sigma_{bb}}}} \\
&= \frac{Corr(\alpha_i, \beta_i)}{\sqrt{1 + k_a}\sqrt{1 + k_b}} + \frac{\rho}{\sqrt{1 + k_a^{-1}}\sqrt{1 + k_b^{-1}}}
\end{aligned}
$$