Modèles Conjoints à Classes Latentes : étude empirique des propriétés du modèle et application dans le domaine de la sclérose latérale amyotrophique

Maéva Kyheng

Université de Lille

THÈSE

pour obtenir le grade de

Docteur de l'Université de Lille

Spécialité : **Biostatistiques**

préparée au laboratoire **ULR 2694 - METRICS - Évaluation des Technologies de Santé et des Pratiques Médicales**

dans le cadre de l'École Doctorale **Biologie Santé de Lille**

présentée et soutenue publiquement
par

Maéva KYHENG

le 1er mars 2023

Titre:

Modèles Conjoints à Classes Latentes : étude empirique des propriétés du modèle et application dans le domaine de la *sclérose latérale amyotrophique*

Directeur de thèse: **Alain DUHAMEL**
Encadrante: **Génia BABYKINA**

Jury

Pr. Eric VICAUT,	Examinateur
Pr. Jacques BENICHOU,	Rapporteur
Pr. Anne GEGOUT-PETIT,	Rapportrice
Pr. David DEVOS,	Président
Dr. Génia BABYKINA,	Encadrante
Pr. Alain DUHAMEL,	Directeur de thèse

Remerciements

Á mes directeurs de thèse

Á Alain pour m'avoir permis de réaliser cette thèse qui me tenait à cœur et de m'avoir dégagé le temps nécessaire à sa réalisation. Merci pour vos conseils, votre disponibilité dans le cadre de ce travail mais également pendant les 6 années passées à la plateforme sous votre direction.

Á Génia, sans qui ce travail n'aurait pas pu voir le jour. Tu m'as appris tellement. Je te remercie pour ta patience, ta bienveillance, ta pédagogie et ta disponibilité sans faille. Je suis fière d'avoir été jusqu'au bout à tes côtés et te remercie infiniment de m'avoir épaulé.

Aux membres de mon CSI et de mon jury

Á Cecile Proust-Lima pour m'avoir donné le goût de la recherche durant tes heures de cours à l'ISPED. Merci de m'avoir fourni les clés indispensables à la réalisation de cette thèse et de nous avoir conseillé tout au long de ce travail. Tes travaux ont été une inspiration pour nos recherches.

Á Mr Eric Vicaux dont la réputation n'est plus à démontrer. C'est un honneur pour moi de vous compter parmi les membres de mon jury.

Á Mme Anne Gegout-Petit et à Mr Jacques Benichou de me faire le plaisir d'être rapporteurs de ma thèse.

Á David Devos pour m'avoir inspiré ce travail de thèse. Ta détermination sans relâche, ta passion pour ton métier et l'envie de sauver tes patients ont été une source d'inspiration. J'espère avoir été à la hauteur et que ce travail pourra t'aider dans tes futurs projets.

Á mes anciens collègues et amis

Á Julien, mon collègue, mon ami. Merci de m'avoir permis de dépasser mes limites et de croire autant en moi. Ta franchise et ton manque de tact légendaire me permettent d'être meilleure chaque jour.

Á Camille et Adeline. Notre incroyable trio m'aura permis de tenir bon même dans les moments difficiles. Merci Camille pour ta patience et ton aide pour mes débuts sur R. Merci Adeline de m'avoir encouragé et soutenu chaque jour dans le bureau. Vous avez été géniales. Cette expérience à Lille n'aurait jamais été la même sans vous ! !

Á Valérie, Hélène, Nassima, Élodie, Émilie, Émeline, Claire, Hajar, Alexis, d'avoir été des collègues et amis hors-pair ! Je garde d'incroyables souvenirs de mon expé-

rience à la plateforme grâce à vous. Vous m'avez tout appris. Vous me manquez énormément !

Et bien sûr ... à ma famille

Á ma maman, la femme la plus forte et combative que je connaisse ! Merci d'être et d'avoir été la maman la plus parfaite de l'univers.

Á ma sœur et mon frère, mes neveux et nièces, les piliers de ma vie. Notre amour invincible nous permet chaque jour de soulever des montagnes.

Á ma famille du nord et plus particulièrement à ma tatie pour m'avoir tellement bien accueilli à Lille.

Et enfin MERCI aux hommes de ma vie. Merci Nam d'être celui sur qui je peux toujours compter. Tu m'as fait devenir celle que j'ai toujours rêvée d'être. Merci de me soutenir dans tous mes délires, et de me pousser à toujours être meilleure. Merci à notre petit bonhomme, Elliot, qui aura rendu cette fin de thèse mouvementée mais tellement belle. Je vous aime.

Résumé

Les modèles conjoints à classes latentes permettent de modéliser conjointement l'évolution d'un biomarqueur et le délai de survenue d'un évènement. Ces modèles présentent l'avantage de tenir compte d'une hétérogénéité (non expliquée par les covariables) pouvant exister pour ces données dans la population étudiée. Les applications dans le domaine médical sont importantes pour : stratifier les patients en sous-groupes différents (classes latentes) relativement à l'évolution du biomarqueur et au délai de survenue de l'évènement, identifier les facteurs associés à l'évolution du biomarqueur en ajustant sur le risque d'évènement ou, inversement, étudier les facteurs associés au risque d'évènement en ajustant sur l'évolution du biomarqueur, tout en tenant compte des différentes classes latentes. Bien que l'implémentation du modèle conjoint à classes latentes soit rendu accessible pour la recherche clinique grâce au package **lcmm** du logiciel **R**, il reste très peu utilisé, de par sa complexité. Dans cette thèse, nous nous intéressons au comportement du modèle vis-à-vis de la taille de l'échantillon : nombre d'individus, nombre de mesures répétées du biomarqueurs et nombre d'évènements. Le comportement du modèle vis-à-vis de la taille de l'échantillon en termes de normalité des paramètres estimés, de biais, de taux de couverture et de capacité à séparer les classes identifiées est étudié empiriquement par une étude de simulation de Monte-Carlo. Des mises en garde quant à la taille d'échantillon en lien avec son impact sur le comportement du modele sont formulées. Une application dans le domaine de la neurologie sur des patients atteints de *sclérose latérale amyotrophique* est ensuite réalisée pour confronter les résultats des simulations aux données réelles. Dans cette application, deux objectifs distincts sont visés : (1) trouver des profils de patients différents vis-a-vis de l'évolution de leur maladie et déterminer les caractéristiques d'inclusion des patients associées á ces profils ; (2) tester l'apport du modele conjoint à classes latentes dans la recherche de covariables associées à l'évolution du biomarqueur ou à la survenue de l'évènement, comparativement aux autres modèles existants. Ce travail pourrait contribuer à développer l'utilisation du modèle conjoint à classes latentes dans le domaine de la recherche clinique.

Mots clés : Modèles conjoints à classes latentes, modèles mixtes, classes latentes, modèles de survie, maximum de vraisemblance, propriétés asymptotiques, maladies neurodégénératives, sclérose latérale amyotrophique.

Abstract

Joint latent class models allow the joint modeling of the evolution of a biomarker and the time to event. These models have the advantage of taking into account the heterogeneity (not explained by the covariates) that may exist for these data in the studied population. Applications in the medical field are important for : stratifying patients into different subgroups (latent classes) with respect to the evolution of the biomarker and the time to event, identifying factors associated with the evolution of the biomarker by adjusting on the risk of event or, conversely, studying factors associated with the risk of event by adjusting on the evolution of the biomarker, while taking into account the different latent classes. Although the implementation of the joint latent class model has been made accessible for clinical research thanks to the **lcmm** package of the **R** software, it remains rarely used, due to its complexity. In this thesis, we are interested in the behavior of the model with respect to the sample size : number of patients, number of repeated biomarker measurements and number of events. The behavior of the model with respect to sample size in terms of normality of the estimated parameters, bias, coverage rate and ability to separate the identified classes is studied empirically by a Monte-Carlo simulation study. Cautions about the sample size in relation to its impact on the behavior of the model are formulated. An application in the field of neurology on patients suffering from *amyotrophic lateral sclerosis* is then carried out to compare the results of the simulations with real data. In this application, two distinct objectives are targeted : (1) to find different patient profiles with respect to the evolution of their disease and to determine the patient baseline characteristics associated with these profiles ; (2) to test the contribution of the joint latent class model in the search for covariates associated with the evolution of biomarker or the risk of event, compared to other models. This work could contribute to develop the use of the joint latent class model in clinical research.

Key words : Joint latent class models, mixed models, latent classes, survival models, maximum likelihood, asymptotic properties, neurodegenerative diseases, amyotrophic lateral sclerosis.

Table des matières

Introduction

Dans le domaine de la recherche clinique, les données longitudinales sont des données précieuses permettant d'étudier l'évolution de l'état du patient en prenant en compte la dynamique d'un biomarqueur au cours du temps ou la survenue d'un évènement particulier.
Pour étudier et prédire l'évolution d'une pathologie au cours du temps, à ce jour, l'utilisation des modèles linéaires mixtes pour analyser l'évolution d'un biomarqueur et des modèles de durée pour analyser la survenue d'un évènement sont très utilisés par les cliniciens.

Toutefois, ces modèles sont parfois utilisés à mauvais escient ou ne permettent pas de répondre aux objectifs posés.
En effet, l'hypothèse fondamentale du modèle linéaire mixte est le caractère aléatoire des données manquantes. Il est également supposé que ces dernières ne reflètent pas directement l'état du patient. Or, dans de nombreux cas, les données longitudinales peuvent être manquantes dues à la survenue d'un évènement, comme par exemple le décès, qui stoppe définitivement les mesures d'un biomarqueur sur un patient. Dans ce cas, les estimations du modèle linéaire mixte sont biaisées. De plus, les modèles linéaires mixtes ne permettent pas d'étudier le lien entre l'évolution d'un biomarqueur et la survenue d'un évènement.
Les modèles de durée eux, ne permettent pas d'étudier l'impact de l'évolution d'un biomarqueur sur le risque de survenue d'un évènement. L'évolution du biomarqueur pourrait être introduite dans le modèle de survie via une variable dépendante du temps. Cependant, comme la valeur du biomarqueur est influencée par la survenue de l'évènement, les estimations du modèle avec variables temps dépendantes sont biaisées. De plus, l'estimation par la méthode du maximum de vraisemblance nécessite de connaître la valeur de la variable à chaque temps d'évènement, ce qui n'est pas souvent le cas pour une variable temps dépendante, mesurée en des temps discrets.

L'analyse conjointe de l'évolution d'un biomarqueur et de la survenue d'un évènement permet de répondre de manière mathématiquement correcte à plusieurs questions essentielles en recherche clinique :

- prédire la survenue d'un évènement en fonction de l'évolution d'un biomarqueur ;
- analyser l'évolution d'un biomarqueur en présence de censures informatives lorsque les données du biomarqueur sont censurées par un évènement d'intérêt ;
- étudier les facteurs associés à l'évolution d'une maladie, décrite par l'évolution du biomarqueur et par la survenue d'un évènement ;

— trouver des profils de patients en terme d'évolutions du biomarqueur et des risques de survenue d'un évènement.

Les modèles statistiques adaptés pour répondre à ces objectifs sont **les modèles conjoints**. Ces modèles regroupent deux modèles distincts : **les modèles conjoints à effets aléatoires partagés** et les **modèles conjoints à classes latentes**. La principale différence entre ces deux types de modèles réside dans l'hypothèse faite sur la structure de la population. La population est considérée homogène pour le cas du modèle à effets aléatoires partagés et hétérogène dans le cas du modèle à classes latentes.

Cette thèse est centrée sur les modèles conjoints à classes latentes.
Les éléments qui ont motivé ce travail de thèse et le choix du modèle à étudier sont listés ci-dessous :
La *sclérose latérale amyotrophique* (SLA) est une maladie grave, neurodégénérative et qui à ce jour n'a pas de traitement efficace permettant de ralentir ou de stopper l'évolution du handicap et la survenue du décès.

Lors d'une discussion méthodologique au sein de l'unité biostatistiques du CHU de Lille, un clinicien neurologue a fait l'hypothèse que les décennies d'échecs thérapeutiques liées à cette maladie pourraient être expliquées par une forte hétérogénéité des patients vis-à-vis de la dégradation de leur état de santé.

Il souhaitait donc (1) identifier différents profils d'évolution de la SLA, en prenant en compte l'évolution du handicap corrélée à la survenue d'un évènement composite ; (2) chercher quelles étaient les caractéristiques des patients à l'inclusion associées à ces différents profils ; (3) en déduire de nouveaux critères d'inclusion à utiliser dans les prochains essais thérapeutiques afin de pouvoir sélectionner les patients chez qui les traitements auraient plus de chances d'être efficaces.
Ayant étudié les modèles conjoints à classes latentes lors de mon Master de recherche à l'ISPED (Bordeaux), j'ai proposé à ce chercheur de mettre en œuvre ces modèles dans le domaine de la SLA pour trouver des profils "latents" de patients en prenant en compte l'évolution du handicap et l'évènement composite (décès, mise sous ventilation non-invasive ou trachéotomie).
Cependant, lors de la modélisation, nous nous sommes retrouvés confrontés à de nombreuses interrogations concernant l'utilisation pratique du modèle conjoint, telles que le nombre de mesures répétées adéquat, le nombre d'évènements optimal ainsi que des règles d'inclusion de covariables dans le modèle.

Une revue de la littérature, sur l'utilisation des modèles conjoints à classes latentes, a révélé que les articles méthodologiques publiés ne répondaient pas à ces questions.
Dans ce contexte, nous avons décidé d'étudier en profondeur ce modèle afin de pallier ces lacunes et de fournir des recommandations aux cliniciens quant à son utilisation.

Les objectifs de cette thèse étaient les suivants :

— synthétiser l'utilisation du modèle conjoint à classes latentes en pratique ;
— étudier le comportement du modèle vis-à-vis de la taille d'échantillon (nombre d'individus, de mesures du biomarqueur, d'évènements) ;
— appliquer le modèle à des données réelles pour répondre à l'objectif du clini-

cien développé ci-dessus ;
— étudier l'apport des modèles conjoints à classes latentes à l'analyse des données réelles comparé aux autres modèles existants.

La structure de la thèse est la suivante. Le **Chapitre 1** fourni des généralités sur les modèles utilisés dans cette thèse tels que les modèles linéaires mixtes, les modèles à classes latentes mixtes, les modèles de durées et les modèles conjoints (à effets aléatoires partagés et à classes latentes). Leur spécification, la méthode d'estimation des paramètres et leur adéquation seront détaillées.

Dans le **Chapitre 2**, nous faisons le bilan de l'utilisation des modèles conjoints à classes latentes en recherche clinique ; des freins à son utilisation sont mis en évidence. Nous proposons des règles quant à l'utilisation de covariables dans les sous-modèles selon les objectifs de l'étude. Nous nous concentrons ensuite sur les propriétés du modèle quant à la taille d'échantillon (nombre de patients, nombre de mesures du biomarqueur, nombre d'évènements) via une étude de simulations.

Le **Chapitre 3** concerne l'application de ces modèles sur des données des patients atteints de *sclérose latérale amyotrophique* à partir de l'étude TROPHOS (essai clinique visant à tester l'impact de l'*olesoxime* sur l'évolution de la maladie) afin de répondre à la question initiale du clinicien : "identifier des caractéristiques initiales des patients atteints de *sclérose latérale amyotrophique* afin de prédire différents profils d'évolution de la maladie".

Dans le **Chapitre 4**, nous évaluons l'impact de covariables préspécifiées sur l'évolution du handicap et sur la survenue de l'évènement composite. Pour cela, nous comparons les résultats du modèle linéaire mixte et du modèle de survie classique aux résultats des modèles conjoints (à classes latentes et à effets aléatoires partagés) afin de déterminer les avantages et les limites de chacun d'entre eux.
Enfin, la dernière partie concerne **la conclusion** et **les perspectives** des travaux futurs.

Chapitre 1

Généralités sur les modèles utilisés

Ce chapitre présente un état des lieux des modèles statistiques utilisés dans cette thèse. Nous présentons :

- l'analyse des données longitudinales d'une population homogène via **les modèles linéaires mixtes** ;
- l'analyse des données longitudinales d'une population hétérogène via **les modèles linéaires mixtes à classes latentes** ;
- l'analyse des **données de durées** via le modèle de survie ;
- les principes de la modélisation conjointe de biomarqueurs longitudinaux et de la survenue d'un évènement d'intérêt d'une population homogène via **les modèles à effets aléatoires partagés** ;
- les principes de la modélisation conjointe de biomarqueurs longitudinaux et de la survenue d'un évènement d'intérêt d'une population hétérogène via **les modèles conjoints à classes latentes**.

Pour chaque modèle, nous développerons leur spécification, la méthode d'estimation ainsi que leur adéquation.

1 Le modèle linéaire mixte

Dans de nombreuses études cliniques, il est courant d'avoir comme critère de jugement l'évolution d'un marqueur prédictif de l'évolution clinique de la maladie, comme par exemple le score de handicap pour la *sclérose latérale amyotrophique* ou le taux de lymphocyte T CD4+ pour le VIH (1). Ces données longitudinales au sens de données répétées sont analysées de manière usuelle à l'aide d'un modèle linéaire mixte classique introduit par Laird and Ware (2) en 1982. Ce modèle permet de prendre en compte l'évolution moyenne du biomarqueur de la population via des effets fixes et les différentes évolutions individuelles à travers des effets aléatoires. Ces effets aléatoires, spécifiques à chaque sujet, représentent les déviations individuelles par rapport à l'évolution moyenne de la population. En effet, les corrélations intra-individu induites par les répétitions au sein d'un même sujet sont prises en compte par ces effets aléatoires offrant ainsi plus de souplesse au modèle.

1.1 Le modèle

Les modèles linéaires mixtes sont une extension directe des modèles linéaires. Notons Y_{ij} la valeur d'une mesure longitudinale Gaussienne observée pour un individu i, $i = 1, ..., n$ au temps t_j, $j = 1, ..., n_i$ et X_{ij} le vecteur de p variables explicatives observées à l'instant t_j. Notons que nous omettons les notations en gras pour les vecteurs et les matrices dans les écritures pour simplifier la présentation. Le modèle linéaire sans effet aléatoire s'écrit alors de la manière suivante :

$$Y_{ij} = X_i(t_{ij})^\intercal \beta + \epsilon_{ij}, \tag{1.1}$$

avec $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ où ϵ_{ij} est l'erreur résiduelle de l'individu i au temps j qui suit une loi gaussienne de moyenne 0 et d'écart type σ^2. Ces erreurs doivent être indépendantes entre elles.

Le modèle linéaire décrit ci-dessus fait l'hypothèse d'une évolution moyenne sur les individus du paramètre longitudinal. Cependant, il est intéressant de pouvoir prendre en compte les variations inter individuelles à l'inclusion et au cours du temps. De plus, les mesures répétées d'un même individu sont plus corrélées entre elles que les mesures d'individus différents. Ces notions peuvent être prises en compte à l'aide d'effets aléatoires ajoutés au modèle précédent. Le modèle linéaire devient alors le modèle linéaire mixte standard qui s'écrit de la manière suivante :

$$Y_{ij} = X_i(t_{ij})^\intercal \beta + Z_i(t_{ij})^\intercal b_i + \epsilon_{ij}, \tag{1.2}$$

avec $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ et $b_i \sim \mathcal{N}(0, B)$.
Dans ce modèle Y_{ij} dépend de :

- un vecteur d'effets fixes β de dimension p associé au vecteur de variables explicatives $X_i(t_{ij})$, éventuellement dépendant du temps. Ce vecteur est de dimension p et correspond aux variables explicatives du sujet i mesurées au temps j ;
- un vecteur d'effets aléatoires b_i de dimension q associé au vecteur des effets aléatoires $Z_i(t_{ij})$ de dimension q qui est un sous vecteur de X_{ij}. Les effets aléatoires sont spécifiques à chaque sujet et sont supposés suivre une loi normale multivariée de moyenne 0 et de matrice variance-covariance B de dimension $q \times q$. Cette matrice B peut être structurée ou non c'est-à-dire que l'on peut spécifier ou non la structure de corrélation. En pratique, dans les modèles longitudinaux, les effets aléatoires traduisent l'hétérogénéité des individus vis-à-vis de leur niveau de base du marqueur longitudinal (*intercept aléatoire*) et/ou vis-à-vis de leur évolution (*pente aléatoire*) ;
- ϵ_{ij} est l'erreur pour l'individu i au temps j. Le vecteur d'erreurs pour un individu i, $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, ..., \epsilon_{in_i})$, est supposé Gaussien, multivarié, centré, de matrice de variance covariance Σ_i de dimension $n_i \times n_i$ et indépendant du vecteur des effets aléatoires. Ces erreurs sont le plus souvent supposées indépendantes entre elles, pour chaque sujet, et de même variance ($\Sigma_i = \sigma^2 I_{n_i}$). Cependant ces erreurs peuvent aussi être auto-corrélées et ainsi tenir compte d'une variabilité supplémentaire à chaque individu en tout temps.

Ce modèle peut également être écrit sous forme matricielle :

$$Y_i = X_i^\intercal \beta + Z_i^\intercal b_i + \epsilon_i, \tag{1.3}$$

avec $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})^\intercal$ le vecteur de réponses du sujet i, $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, ..., \epsilon_{in_i})$ le vecteur des erreurs, $X_i(n_i \times p)$ et $Z_i(n_i \times q)$ les matrices de variables explicatives pour l'ensemble des temps d'observation du sujet i, dont les lignes sont les vecteurs $X_i(t_{ij})^\intercal$ et $Z_i(t_{ij})^\intercal$. Les éléments de ces derniers sont les valeurs des p variables explicatives du sujet i au temps t_{ij}.

1.2 Estimations et inférence sur les paramètres

L'estimation des paramètres du modèle linaire mixte (Eq.1.3) s'effectue à partir de la formulation marginale du modèle (Eq.1.4), c'est-à-dire dépourvue d'effet aléatoire. En définissant le vecteur aléatoire $\varepsilon_i = Z_i b_i + \epsilon_i$, le modèle marginal s'écrit alors comme :

$$Y_i = X_i\beta + \varepsilon_i, \tag{1.4}$$

avec $\varepsilon_i \sim \mathcal{N}(0, V_i = Z_i B Z_i^\intercal + \Sigma_i)$.
Ainsi, le modèle linéaire mixte décrit en Eq.(1.3) suppose que le vecteur des réponses pour un individu soit une variable multivariée gaussienne $Y_i \sim \mathcal{N}(X_i\beta, V_i)$. Tous les paramètres associés aux variables explicatives (notés β) et aux matrices de variance-covariance V_i (notés ϕ) peuvent donc être estimés par la méthode du maximum de vraisemblance. La vraisemblance de ce modèle a la forme suivante :

$$L(\beta, \phi) = \prod_{i=1}^{N} \left(\frac{1}{2\pi}\right)^{n_i/2} |V_i|^{-1/2} \exp\left(-\frac{1}{2}(Y_i - X_i\beta)^\intercal V_i^{-1}(Y_i - X_i\beta)\right). \tag{1.5}$$

La log-vraisemblance du modèle s'écrit donc :

$$\begin{aligned} L(\beta, \phi) &= \sum_{i=1}^{N} \log(f_{Yi}(Y_i)) \\ &= -1/2 \sum_{i=1}^{N} \left(n_i \log(2\pi) + \log|V_i| + (Y_i - X_i\beta)^\intercal V_i^{-1}(Y_i - X_i\beta)\right), \end{aligned} \tag{1.6}$$

où $|V_i|$ est le déterminant de V_i.

Afin d'estimer les paramètres du modèle à partir de cette log-vraisemblance, il suffit alors de résoudre l'équation du score par rapport aux paramètres β :

$$\frac{\partial L(\beta, \phi)}{\partial \beta} = \sum_{i=1}^{N} X_i^T V_i^{-1}(Y_i - X_i\beta) = 0. \tag{1.7}$$

Lorsque les paramètres ϕ sont connus, on peut alors estimer les β par :

$$\hat{\beta} = \left(\sum_{i=1}^{N} X_i^\intercal V_i^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i^{^{-1}X_i}\right). \tag{1.8}$$

En pratique les paramètres de variance ne sont pas connus et doivent être estimés. L'estimation est obtenue en maximisant la log-vraisemblance (Eq.1.6), en remplaçant le β par son estimation $\hat{\beta}$ obtenue par l'équation Eq.(1.8) en utilisant des algorithmes itératifs (Newton-Raphson ou Quasi-Newton). Une fois ces estimations

réalisées, l'inférence statistique pour les effets fixes s'effectue grâce à un test de Wald, du rapport de vraisemblance ou de score permettant de déterminer la significativité des paramètres du modèle. Nous ne développerons pas ici les détails de cette inférence, qui est basée sur les propriétés asymptotiques d'estimateur de maximum de vraisemblance. Ces détails peuvent être trouvés, par exemple, dans l'ouvrage de Daniel Commenge et Hélène Jacqmin-Gadda (3).

1.3 Adéquation du modèle

L'adéquation du modèle linéaire mixte peut être évaluée à l'aide d'analyse des résidus, comme dans le modèle linéaire classique. Les résidus sont définis de la manière suivante :

$$Y_i - X_i\hat{\beta}, \tag{1.9}$$

et sont appelés les *résidus marginaux*. Elle peut également être étudiée à l'aide des *résidus spécifiques aux sujets*, définis par

$$Y_i - X_i\hat{\beta} - Z_i\hat{b}_i. \tag{1.10}$$

Dans ce cas, $\hat{b}_i$ est estimé à partir de :

$$\hat{b}_i = \hat{B}Z_i^\intercal \hat{V}_i^{-1}\left(Y_i - X_i\hat{\beta}\right). \tag{1.11}$$

En pratique les méthodes graphiques sont souvent privilégiées pour l'étude d'adéquation. Nous pouvons les résumer par les approches suivantes :

- le graphique présentant les valeurs prédites selon les valeurs observées, donnant un indice sur le pouvoir prédictif du modèle ;
- le graphique des résidus standardisés (résidus divisés pas l'écart-type) versus le temps ou versus une variable explicative quantitative permettant d'évaluer si l'effet de cette variable est correctement spécifié ou de repérer des observations extrêmes et éventuellement une homoscédasticité ;
- le quantile-quantile plot permettant de vérifier la normalité des résidus.

2 Le modèle linéaire mixte à classes latentes

Le modèle linéaire mixte tient compte d'une hétérogénéité individuelle en termes d'évolution de marqueur longitudinal, mais suppose un seul profil (en termes de vecteur X) d'évolution de Y avec des déviations individuelles gaussiennes centrées autour de ce profil moyen. Cependant dans de nombreux cas, une hétérogénéité de la population vis-à-vis de cette évolution peut être suspectée. Lorsque l'on ne connaît pas la variable qui explique les différents profils d'évolution, on parle d'une variable latente. Cette variable s'oppose à une variable observée comme par exemple, un traitement dans les essais thérapeutiques. Ce phénomène latent peut être pris en compte dans le cadre de modèles spécifiques (4; 5) développés à la fin des années 1990. Le principe de ces modèles est de supposer qu'il existe un nombre défini G de sous-populations qui caractérisent les profils différents, appelés *classes latentes*. Ce modèle est introduit ci-dessous.

2.1 Le modèle

Le modèle linéaire mixte à classes latentes est composé de 2 sous-modèles : un modèle logistique multinomial qui définit l'appartenance aux classes et un modèle mixte spécifique à chaque classe, définissant l'évolution du marqueur longitudinal pour chaque profil retrouvé. Chaque sujet i appartient à une seule classe latente qui se traduit par la variable latente c_i valant g si l'individu i appartient à la classe g.

Le sous-modèle logistique multinomial

Le sous-modèle logistique multinomial pour un individu i définit la probabilité d'appartenir à la classe latente g sachant les variables explicatives X_{1i} :

$$\pi_{ig} = P(c_i = g|X_{1i}) = \frac{\exp(\xi_{0g} + X_{1i}^{\intercal}\xi_{1g})}{\sum_{l=1}^{G}\exp(\xi_{0l} + X_{1i}^{\intercal}\xi_{1l})}, \tag{1.12}$$

où l'on suppose que $\xi_{0G} = 0$ et $\xi_{1G} = 0$ afin que la classe G devienne la référence. En général, les variables explicatives X_{1i} sont incluses dans le modèle logistique multinomial lorsque l'on souhaite créer ou interpréter les classes en fonction de ces variables. N'ayant le plus souvent aucun a priori sur la constitution des classes, aucune variable explicative n'est ajoutée à ce modèle en pratique. Il devient donc :

$$\pi_{ig} = \frac{\exp(\xi_{0g})}{\sum_{l=1}^{G}\exp(\xi_{ol})}. \tag{1.13}$$

Le sous-modèle linéaire mixte

Le sous-modèle linéaire mixte est issu du modèle développé dans l'Eq.(1.2) et est spécifique à chaque classe latente g. Il est définit de la façon suivante :

$$Y_{ij} = X_{2i}(t_{ij})^{\intercal}\beta + X_{3i}(t_{ij})^{\intercal}\beta_g + Z_i(t_{ij})^{\intercal}b_i + \epsilon_{ij}, \tag{1.14}$$

avec $b_i \sim \mathcal{N}(\mu_g, B_g)$.

Dans ce modèle, $X_{2i}(t_{ij}), X_{3i}(t_{ij})$ et $Z_i(t_{ij})$ sont des vecteurs de variables explicatives sans aucune intersection entre eux pour assurer l'identifiabilité du modèle.

- $X_{2i}(t_{ij})$ est associé au vecteur d'effets fixes β qui sont communs aux classes latentes c'est-à-dire, l'effet des covariables est le même dans chacune des classes. Ces variables sont inclues dans le modèle lorsque l'on souhaite s'affranchir de leurs effets sur la création des classes.
- $X_{3i}(t_{ij})$ est associé au vecteur d'effets fixes β_g spécifiques aux classes latentes. Lorsque le temps est inséré dans ce vecteur, cela permet de modéliser une évolution temporelle (pente) différente selon les classes. Nous pouvons également ajouter d'autres covariables si l'on souhaite créer les classes qui en dépendent.
- $Z_i(t_{ij})$ est un vecteur associé aux effets aléatoires b_i distribués identiquement et indépendamment suivant une loi normale spécifique aux classes. A l'inverse de l'Eq.(1.2), les effets aléatoires ne sont pas centrés sur 0 mais ont une moyenne μ_g différente dans chaque classe latente et une variance-covariance possiblement différente B_g. En effet la variance des effets aléatoires peut être égale ou proportionnelle entre les classes latentes.
- L'erreur ϵ_{ij} est similaire à celle définie dans l'Eq.(1.2).

2.2 Estimations et inférence sur les paramètres

Les paramètres du modèle sont le plus souvent estimés par le maximum de vraisemblance (4; 5; 6) avec un nombre de classes latentes fixé. Une approche Bayesienne est également possible mais plus compliquée à mettre en œuvre (7) et ne sera donc pas abordée dans cette thèse. Les paramètres à estimer sont ceux de l'Eq.(1.12) et les paramètres du sous-modèle linéaire mixte (l'Eq.(1.14)) (β, β_g, les éléments de B et les éléments de Σ_i). Nous regroupons ces paramètres dans le vecteur θ_G. La distribution marginale de Y_{ij} par rapport aux effets aléatoires et conditionnelles aux classes latentes utilisée pour l'estimation est comme suit :

$$Y_{ij}|c_i = g \sim \mathcal{N}(X_{2i}\beta + X_{3i}\beta_g + Z_i\mu_g, Z_iB_gZ_i^\intercal + \Sigma_i), \tag{1.15}$$

où X_{2i}, X_{3i} et Z_i sont les matrices ayant respectivement comme vecteur ligne $X_{2ij}^\intercal$, $X_{3ij}^\intercal$ et $Z_{ij}^\intercal$ avec $j = 1, ..., n_i$.
La vraisemblance pour les individus i s'écrit donc :

$$L(X_iY_i|\theta_G) = \prod_{i=1}^{N}\left(\sum_{g=1}^{G} P(c_i = g|X_{1i}, \theta_G) \times \phi_{ig}(Y_i|c_i = g; X_{2i}, X_{3i}, Z_i, \theta_G)\right), \tag{1.16}$$

où ϕ_{ig} est la densité d'une loi normale multivariée d'espérance $X_{2i}\beta + X_{3i}\beta_g + Z_i\mu_g$ et de variance $Z_iB_gZ_i^\intercal + \Sigma_i$. La log vraisemblance est maximisée par un algorithme itératif de type EM (espérance-maximisation) ou Newton-Raphson.

2.3 Classification à posteriori

Le modèle mixte à classes latentes permet, pour un nombre de classes fixé, de prédire la probabilité *a posteriori* qu'un individu appartienne à une classe selon ses caractéristiques. Cette probabilité notée $\hat{\pi}_{ig}^Y$ est obtenue par la formule de Bayes :

$$\begin{aligned}\hat{\pi}_{ig}^Y &= P(c_i = g|X_i, Y_i, \hat{\theta}_G) \\ &= \frac{P(c_i = g|X_{1i}, \hat{\theta}_G)\phi_{ig}(Y_i|c_i = g, X_{2i}, X_{3i}, Z_i, \hat{\theta}_G)}{\sum_{l=1}^{G} P(c_i = l|X_{1i}; \hat{\theta}_G)\phi_{il}(Y_i|c_i = l, X_{2i}, X_{3i}, Z_i, \hat{\theta}_G)}.\end{aligned} \tag{1.17}$$

Chaque sujet peut donc être affecté à une classe latente à laquelle il a la plus grande probabilité d'appartenir.

2.4 Le choix du nombre de classes latentes

Pour choisir le nombre optimal de classes latentes il faut faire un compromis entre :

- la vraisemblance maximale, via le critère d'information bayésien basée sur la vraisemblance ($BIC = -2L(\theta_G) + p_G \log(N)$) où p_G est le nombre de paramètres estimés et N est le nombre d'individu ;
- le nombre d'individu par classe ; une classe ne doit pas être représentée par trop peu d'individus ;
- l'intérêt clinique ; le nombre de classes doit rester interprétable cliniquement ;

— la classification *a posteriori* : la probabilité moyenne d'appartenir à la bonne classe permet d'obtenir le niveau d'ambiguïté de la classification. Par exemple, peut être définie la proportion de sujet ayant une probabilité au-dessus d'un certain seuil (0.8 par exemple) ou inversement en dessous d'un certain seuil (0.6 ou 0.5 par exemple) pour chaque classe obtenue *a posteriori*. Si la classification est parfaite, chaque sujet devrait avoir une probabilité 1 d'appartenir à la classe latente affectée *a posteriori* et une probabilité de 0 d'appartenir aux autres classes latentes.

2.5 Adéquation du modèle

L'adéquation du modèle linéaire mixte à classes latentes peut être évaluée de plusieurs façons. Comme dans les modèles linéaires mixtes classiques, des prédictions individuelles (marginales ou conditionnelles aux effets aléatoires) peuvent être calculées et comparées aux observations pour évaluer le pouvoir prédictif du modèle (les prédictions seront détaillées dans la Chapitre 1, Section 5.3). Ces prédictions sont spécifiques aux classes latentes. Á partir de ces prédictions individuelles dans chaque classe latente deux prédictions moyennes sont calculables :

— une unique prédiction par sujet, en les moyennant sur les classes et en les pondérant par la probabilité d'appartenance aux classes. Ces dernières peuvent être comparées aux moyennes pondérées des observations ;
— une prédiction unique par classe, en calculant la moyenne de toutes les mesures observées à un temps donné et pondérées par la probabilité d'appartenance aux classes. Nous pouvons alors évaluer l'adéquation à l'aide du graphique des prédictions moyennées sur chaque classe *vs.* observations moyennes pondérées, à différents temps de mesures. Nous pouvons également représenter les résidus conditionnels aux effets aléatoires et moyennés sur les classes sur la distribution normale théorique de ces résidus.

Enfin, il est important de vérifier la qualité de la classification issue du modèle mixte à classes latentes à partir de la classification *a posteriori* définit dans l'Eq.(1.17) du Chapitre 1, Section 2.3.

3 Le modèle de durée

L'étude des facteurs associés au délai jusqu'à un évènement précis (par exemple le décès) s'effectue grâce aux méthodes d'analyses de durées, appelées couramment «analyses de survie». Ces méthodes sont utilisées pour des données qui contiennent des informations concernant l'évènement à étudier : la date ou délai jusqu'à l'évènement, type d'évènement et des variables explicatives qui influencent les délais de suivis. Notons que le terme «survie» est utilisé également lorsqu'il s'agit d'autres types d'évènement que le décès ; dans ce cas «survivre» signifie «ne pas avoir un évènement». La principale difficulté dans ce type d'analyse réside dans le fait que les données sont souvent incomplètes ; les temps d'évènements ne sont pas observés pour tous les individus. Par exemple, les évènements peuvent intervenir avant la date d'inclusion et dans ce cas nous parlerons de données *tronquées à gauche* ; ils peuvent intervenir entre deux temps d'observation et nous parlerons alors de *cen-*

sures par intervalle ; ils peuvent intervenir après la fin du suivi si l'individu sort de l'étude sans avoir subi l'évènement, nous parlons en pratique de *censures à droite*. Nous nous limiterons à l'analyse de données censurées (à droite et par intervalle).

Les fonctions caractérisant le modèle de survie sont décrites ci-après.

3.1 Définitions

Notons T^* la variable aléatoire continue et positive qui représente la durée de survie depuis l'entrée dans l'étude. Cette variable peut être décrite par différentes fonctions liées entre elles :

— la fonction de survie qui est la probabilité que l'évènement arrive après le temps t, elle peut être interprétée comme la probabilité de "survivre" jusqu'à un temps t fixé :

$$S(t) = \mathbb{P}(T^* > t), t \geq 0 \tag{1.18}$$

— la fonction de répartition qui est la probabilité d'observer l'évènement avant t :

$$F(t) = \mathbb{P}(T^* \leq t) = 1 - S(t) \tag{1.19}$$

— la fonction de densité de probabilité, notée $f(t)$, est la probabilité d'observer l'évènement dans un petit intervalle de temps $[t, t + \Delta_t]$. Elle est définie par :

$$f(t) = \lim_{\Delta_t \to 0^+} \frac{\mathbb{P}(t \leq T^* < t + \Delta_t)}{\Delta_t} = F'(t) \tag{1.20}$$

— la fonction de risque instantané, noté $\lambda(t)$, qui est le risque d'observer l'évènement dans un petit intervalle de temps $[t, t + \Delta_t]$, sachant que le sujet a survécu jusqu'au temps t (c'est-à-dire il est dans le groupe à risque de survenue d'évènement) :

$$\begin{aligned} \lambda(t) &= \lim_{\Delta_t \to 0^+} \frac{\mathbb{P}\left(t \leq T^* < t + \Delta_t | T^* \geq t\right)}{\Delta_t} \\ &= \lim_{\Delta_t \to 0^+} \frac{\mathbb{P}\left(t \leq T^* < t + \Delta_t\right)}{\mathbb{P}\left(T^* \geq t\right) \Delta_t} \\ &= \frac{f(t)}{S(t)} = \frac{F'(t)}{S(t)} \\ &= \frac{-S'(t)}{S(t)} = -\ln(S(t))' \end{aligned} \tag{1.21}$$

— la fonction du risque cumulé, $\Lambda(t)$ est l'intégral du risque instantané :

$$\Lambda(t) = \int_0^t \lambda(u)du = -\ln(S(t)); \tag{1.22}$$

— la fonction indicatrice de l'évènement, δ. Si l'évènement n'a pas eu lieu avant la fin d'étude, la réelle durée T^* n'est pas connue. Ainsi, la durée qui est considérée dans la modélisation est définie par :

$$T = \min(T^*, C), \tag{1.23}$$

où C est le temps de censure. $\delta = \begin{cases} 0, & \text{si } T^* > C. \\ 1, & \text{sinon.} \end{cases}$

Plusieurs méthodes d'analyses de survie peuvent être effectuées en fonction du but recherché.

— **L'estimation de la fonction de survie** $S(t)$. Cette estimation se fait principalement grâce à un estimateur non paramétrique : l'estimateur de Kaplan Meier. Cet estimateur est basé sur le maximum de vraisemblance non-paramétrique de la fonction de survie. D'autres estimateurs, paramétriques cette fois, tels que le modèle exponentiel, de Weibull ou de Gompertz, permettent d'estimer cette fonction. Ils supposent que la distribution des temps de survie appartient à une famille dont le vecteur de paramètres réels est de dimension finie.
Dans ce mémoire, nous développerons certains modèles de survie paramétriques (exponentiel et Weibull) qui peuvent être utilisés dans le modèle conjoint.

— **La comparaison simple de courbes de survie selon les modalités d'une variable explicative qualitative**. Cette comparaison s'effectue à l'aide du *test du logrank* dont le principe est de comparer la fréquence d'évènements entre les groupes au cours du temps.

— **L'analyse multidimensionnelle**. Dans cette analyse l'influence de variables explicatives (quantitatives ou qualitatives) sur la survie des individus se fait à l'aide de modèles de régression. Le modèle le plus utilisé est le modèle à *risques proportionnels* (ou modèle de Cox). Dans ce modèle la fonction de survie est estimée d'une manière semi-paramétrique. Neanmoins, ce modèle peut être facilement généralisé pour le cas paramétrique (fonction de survie exponentielle, Weibull, Gompertz, *etc.*).

3.2 Le modèle à risques proportionnels

Le modèle à risques proportionnels est le modèle de survie le plus utilisé pour étudier l'effet de covariables sur la durée de survenue d'un événement d'intérêt avec des données censurées. Pour le sujet $i, i = 1, \cdots, N$, nous notons $T_i^\star$ le temps jusqu'à l'événement d'intérêt et C_i le temps de censure. C_i peut représenter le temps de suivi des sujets encore à risque à la fin de l'étude ou le temps jusqu'à leur sortie d'étude. La fonction de risque instantané définie dans l'Eq.(1.21) du sujet i au temps t est la suivante :

$$\lambda_i(t|W_i) = \lambda_0(t)\exp(W_i^T\gamma), \tag{1.24}$$

où $\lambda_0(t)$ est la fonction de *risque de base* commune à tous les sujets et γ est le vecteur de coefficients mesurant l'association entre le vecteur de covariable W_i et le risque instantané d'évènement. L'impact de chaque covariable W_k sur le risque instantané est exprimés en risque relatif, défini par :

$$RR(W_k) = \frac{R(W_k = w_k)}{R(W_k = w_k^{\text{ref}})} = \exp(\gamma_k), \tag{1.25}$$

où w_k^{ref} est la modalité de référence. Pour les variables quantitatives, le risque relatif est exprimé en évolution de risque pour l'augmentation d'une unité de la variable.

Le modèles à risques proportionnels peut-être spécifié d'une manière **paramétrique**, dans lequel λ_0 est modélisé par une fonction de risque de type Weibull,

exponentiel *etc.*, ou d'une manière **semi-paramétrique** où λ_0 reste non spécifié. Ce dernier modèle, plus communément appelé modèle de COX à risques proportionnels (8), est le plus souvent utilisé.

3.3 Estimations et inférence sur les paramètres

L'estimation des paramètres du modèle de survie dépend de son caractère paramétrique ou semi-paramétrique.

Dans le cas d'un **modèle paramétrique**, l'estimation est réalisée via la maximisation directe de la vraisemblance basée sur les fonctions définies dans les Eq.(1.18), Eq.(1.21) et Eq.(1.23). Pour l'ensemble des individus cette vraisemblance s'écrit comme :

$$\prod_{i=1}^{N} \lambda_i(T_i)^{\delta_i} S(T_i). \tag{1.26}$$

Dans le cas d'un modèle exponentiel, la fonction de risque est constante au cours du temps $\lambda_0(t) = \lambda_0$. Si cette hypothèse de risque constant est trop forte, il est possible d'utiliser une loi exponentielle par morceaux, ou par une distribution plus flexible, par exemple distribution de Weibull. Pour $T^* \sim Weibull\,(\xi_1, \xi_2)$, le risque instantané de base a la forme suivante :

$$\lambda_0(t) = \xi_1 \xi_2 (\xi_2 t)^{\xi_1^{-1}}. \tag{1.27}$$

L'estimation du modèle **semi-paramétrique** de Cox est basée sur la maximisation de la vraisemblance partielle, où λ_0 est un paramètre de nuisance. Notons que le modèle paramétrique de Weibull sera utilisé dans la suite de ce mémoire, ainsi, les détails d'estimations dans le cadre du modèle semi-paramétriques ne sont pas fournis.

3.4 Adéquation du modèle

Il existe plusieurs types de résidus dans le cadre de modélisation des durées.
Les résidus de **Martingale** sont définis pour l'individu i à l'instant T_i comme :

$$\widehat{M_i} = \delta_i - \widehat{\Lambda}(T_i). \tag{1.28}$$

Ces résidus représentent la différence entre l'indicatrice d'évènement et le risque d'évènement cumulé par l'individu au cours du temps, estimé par le modèle. Ces résidus permettent de vérifier la bonne spécification du modèle et l'hypothèse de log-linéarité pour les covariables quantitatives. L'analyse de ces résidus peut être effectuée par des tests ou graphiquement.

Les résidus de **Schoenfeld** permettent de vérifier l'hypothèse de proportionnalité des risque c'est-à-dire qu'il n'y ait pas de corrélation entre le temps et les résidus pour toutes les variables explicatives. Ces résidus sont définis pour les individus qui ont subi l'évènement et pour chaque variable explicative comme :

$$\widehat{\mathrm{Sc}}_i = \delta_i (W_{ik} - \widehat{W}_{ik}), \tag{1.29}$$

où W_{ik} est la valeur de covariable W_k de l'individu i au temps T_i et $\widehat{W}_{ik}$ est la moyenne de cette covariable sur l'ensemble d'individus à risque au temps T_i, pondérée par leur probabilité de subir un évènement à T_i, estimée par le modèle.

Ces résidus peuvent également être analysés à l'aide des tests ou graphiquement.

4 Le modèle conjoint à effets aléatoires partagés

Dans les Sections précédentes nous avons mis en avant l'utilisation de données longitudinales en recherche clinique. Initialement analysés de manière séparée, les modèles linéaires mixtes pour les marqueurs continus longitudinaux et les modèles de durée pour les évènements d'intérêts permettent aux chercheurs d'étudier la progression d'une maladie et de trouver les facteurs qui y sont associés.
Récemment, l'analyse conjointe de l'évolution d'un biomarqueur et de la survenue d'un évènement est devenue un sujet essentiel en recherche clinique. Elle permet entre autres d'analyser **la relation entre un biomarqueur longitudinal et un temps d'événement**. Il est possible d'étudier cette relation en intégrant un marqueur longitudinal comme une variable dépendante du temps dans un modèle de durée, mais cette approche n'est pas adaptée pour deux raisons. La première provient du fait que le modèle de survie avec variables dépendantes du temps nécessite de connaître la valeur de la variable explicative à tous les temps d'évènement. Or le marqueur longitudinal est souvent mesuré à des temps fixes ce qui nécessiterait une imputation des ces données, ce qui peut induire des erreurs. La deuxième raison est que le marqueur longitudinal est souvent une variable endogène ; c'est-à-dire qu'elle peut être modifiée par la survenue de l'évènement. Seules les variables exogènes peuvent être utilisés dans ces modèles de survie, comme par exemple les niveaux de pollution, les saisons, les températures extérieures. Ce sont des variables qui restent inchangées après la survenue de l'évènement et qui ne sont généralement pas mesurées sur le sujet (9). D'autre part, l'analyse conjointe permet d'analyser **l'évolution d'un biomarqueur** lorsque celui-ci est **censuré par un évènement**. En effet, la sortie d'étude informative (liée à l'évolution longitudinale) dans le modèle linéaire mixte qui engendre des données manquantes du biomarqueur, est une source de biais dans les estimations du modèle. Le modèle linéaire mixte n'est donc plus adapté dans cette situation et l'utilisation d'un modèle conjoint devient la solution.

Les modèles statistiques qui étudient conjointement l'évolution d'un paramètre longitudinal et la survenue d'un évènement sont apparus dans les années 90 avec l'approche "$Two - stage$" développé par Tsiatis et al. (10) en 1995. Ces modèles consistaient à estimer les prédictions individuelles de chaque sujet à tous les temps à l'aide du modèle linéaire mixte puis de les introduire dans le modèle de survie comme covariables dépendantes du temps. La principale limite de ces modèles repose sur le fait que les modèles linéaires mixtes sont estimés en présence de données manquantes informatives (*missing not at random* ou MNAR).
Les modèles conjoints à effets aléatoires partagés ont donc ensuite été développés pour pallier à ce problème (11; 12). Ils ont seulement été utilisés de manière régulière à la fin des années 2000 après la diffusion de plusieurs packages R dédiés à ces modèles (JM, $JMbayes$, $joineR$ et surtout $lcmm$) et d'autres fonctions (notamment $stjm$ sous Stata et $JMFit$ sous SAS). Ce sont ces modèles qui seront développés dans la

suite de la Section.

4.1 Le modèle

Le modèle à effets aléatoires partagés combine un modèle linéaire mixte pour l'évolution du marqueur longitudinal et un modèle de survie pour le temps d'évènement. La structure latente qui définit l'association entre le marqueur et l'évènement est étudiée en introduisant au sein du modèle de survie des effets aléatoires ou fonctions d'effets aléatoires comme variables explicatives. Nous développerons ici le modèle pour une variable longitudinale Gaussienne et un évènement censuré à droite bien que d'autres modèles aient été développés pour différentes situations, comme par exemple, plusieurs marqueurs longitudinaux (13; 14; 15)

Notons $Y(t)$ l'évolution d'un marqueur continu au cours du temps. Ce marqueur est mesuré aux temps t_{ij} et Y_{ij} désigne la variable mesurée au temps t_{ij} sur le sujet $i, i = 1, ...N$. T_i^* est le temps de survenue de l'évènement et C_i le temps de censure à droite comme définit dans l'Eq.(1.23). Le modèle à effets aléatoires partagés combine un modèle linéaire mixte et un modèle à risques proportionnels. Il est défini de la manière suivante :

$$\begin{cases} Y_{ij} = X_{Yij}^\intercal \beta + Z_{ij}^\intercal b_i + \epsilon_{ij} = Y_{ij}^* + \epsilon_{ij} \\ \text{où } \epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2) \\ \text{et } b_i \sim \mathcal{N}(0, B) \\ \lambda_i(t) = \lambda_0(t) \exp(X_{Ti}^\intercal \gamma + g_i(b_i, t)^\intercal \eta), \end{cases} \tag{1.30}$$

où λ_i est la fonction de risque, X_{Yij} et X_{Ti} sont des vecteurs de variables explicatives associés à l'évolution du marqueur et à la survenue de l'évènement respectivement et Z_{ij} un sous-vecteur de X_{Yij} représentant les variables aléatoires du modèle linéaire mixte. La fonction du risque de base λ_0 peut prendre les mêmes formes que dans un modèle de survie classique (voir Chapitre 1, Section 3).
L'association entre le modèle linéaire mixte et le modèle de survie est représentée par la fonction $g_i(b_i, t)$, qui peut être uni- ou multidimensionnelle, et par le vecteur de paramètres η. La fonction des effets aléatoires $g_i(b_i, t)$ doit être choisie en fonction de l'objectif de l'étude.

— Si l'on souhaite tester si le marqueur Y est un bon marqueur de substitution d'un évènement (comme le décès par exemple), on teste si le risque instantané de l'évènement au temps t dépend de la valeur du marqueur en t débarrassée de l'erreur de mesure (16; 11). Ainsi on définit :

$$g_i(b_i, t) = Y_i^*(t) = X_{Yi}^\intercal(t)\beta + Z_i^\intercal(t) b_i, \tag{1.31}$$

et le modèle de survie devient :

$$\lambda(t) = \lambda_0(t) \exp(X_i \gamma + Y_i^*(t)\eta). \tag{1.32}$$

Si X correspond à une variable indicatrice d'un traitement pour un individu i, ce modèle évalue la qualité de Y en tant que marqueur de substitution de l'évènement. Le test de nullité est alors effectué sur γ.

— Si l'on souhaite savoir si le risque d'évènement dépend de la dynamique du marqueur, la fonction des effets aléatoires devient $g_i(b_i, t)^\intercal = (Y_i^*(t), Y_i^*(t)')$

(17). Le risque instantané de l'évènement en t dépend à la fois de la vraie valeur en t du paramètre et de sa pente en t. Le modèle de survie devient alors :

$$\lambda(t) = \lambda_0(t) \exp(X_i\gamma + Y_i^*(t)\eta_1 + Y_i^*(t)'\eta_2). \tag{1.33}$$

Si $\eta_2 = 0$ cela indique qu'après ajustement sur la valeur courante du marqueur le risque ne dépend pas de sa dynamique.

— La fonction de dépendance peut également prendre la forme d'une dépendance directe à travers les effets aléatoires individuels (18). Cette dernière suppose que le risque d'évènement en t dépend de la déviation individuelle du marqueur au temps t plutôt que de l'espérance du marqueur en t. On utilise ce modèle pour distinguer l'impact des variables explicatives X_{Yi} sur le risque d'évènement, de l'impact des variations individuelles du marqueur longitudinal sur ce risque. La fonction des effets aléatoires s'écrit : $g_i(b_i, t) = Y_i^*(t) = Z_i^\intercal(t)b_i$ et le modèle de survie devient :

$$\lambda(t) = \lambda_0(t) \exp(X_{Yi}^\intercal\gamma + Z_i^\intercal(t)b_i\eta). \tag{1.34}$$

— D'autres fonctions de dépendances que nous ne développerons pas dans ce rapport peuvent être utilisées en fonction de l'objectif : mesurer l'association entre le risque d'évènement et les déviations individuelles sur le niveau initial et sur la pente de Y, ou sur le niveau initial et le changement du marqueur entre le temps 0 et le temps t par exemple.

4.2 Estimations et inférence sur les paramètres

L'estimation des paramètres d'un modèle conjoint à effets aléatoires partagés peut se faire de différentes façons. La première consiste à estimer le modèle conjoint en deux étapes, c'est-à-dire d'estimer d'abord le modèle linéaire mixte afin d'en déduire les estimations de $g_i(b_i, t)$ pour chaque sujet et d'ensuite estimer le modèle de survie ajusté sur $g_i(b_i, t)$. Le principal problème de cette méthode d'estimation repose sur le fait que l'estimation du modèle linéaire mixte se fait en présence de données manquantes informatives qui peuvent induire une erreur dans la maximisation de la vraisemblance. Tisatis *et al.* (10) ont proposé de contourner ce problème en estimant un modèle linéaire mixte pour chaque temps d'évènement en n'utilisant que les données collectées avant la survenue de l'évènement sur les sujets à risque. Cependant cette méthode est bien plus lourde numériquement et également moins robuste en raison de variations de la taille de l'échantillon et du nombre de mesures disponibles à chaque temps. La meilleure façon d'estimer les paramètres de ces modèles est par la vraisemblance conjointe de Y_i, T_i, et δ_i pour chaque sujet.

La vraisemblance conjointe

Soit θ le vecteur contenant l'ensemble des paramètres des sous-modèles linéaires mixtes (effets fixes θ_Y et effets aléatoires θ_b) et de survie (θ_e). Tous les paramètres sont estimés simultanément à travers la log-vraisemblance conjointe qui s'écrit grâce à l'hypothèse d'indépendance conditionnelle entre l'évolution du marqueur Y et le

temps d'évènement T :

$$l(\theta) = \sum_{i=1}^{N} \log \left(\int f(Y_i|b_i;\theta_y)\lambda_i(T_i|b_i;\theta_e)^{\delta_i} S_i(T_i|b_i,\theta_e) f(b_i;\theta_b) db_i \right), \tag{1.35}$$

où $f(Y_i|b_i;\theta_y)$ et $f(b_i;\theta_b)$ sont les fonctions de densité de lois normales multivariées d'espérance $X_{Yi}\beta + Z_i b_i$ et 0 respectivement et de matrices de variance-covariance $\sum_i$ (erreurs de mesures indépendantes) et B (variance-covariance des effets aléatoires). Les matrices X_{Yi} et Z_i sont des matrices de vecteurs ligne $X_{Yi}(t_{ij})^\intercal$ et $Z_i(t_{ij})^\intercal$. Le terme $\lambda_i(T_i|b_i;\theta_e)^{\delta_i} S_i(T_i|b_i,\theta_e)$ représente la densité d'un temps d'évènement pour données censurées à droite : $\lambda_i(T_i|b_i;\theta_e)$ est la fonction de risque instantané au temps T_i et $S_i(T_i|b_i,\theta_e)$ est la fonction de survie associée. Cette vraisemblance est difficile à optimiser puisque le calcul de l'intégrale sur les effets aléatoires n'a pas de solution analytique. Elle peut être calculée par quadrature gaussienne, par la méthode MCMC (Markov Chain Monte Carlo) ou par approximation de Laplace lorsque le modèle inclut de nombreux effets aléatoires (19). De plus, dans la majorité des cas, l'intégrale univariée sur le temps dans la fonction de survie n'a pas de solution analytique. Elle peut également être calculée par quadrature. La vraisemblance peut être maximisée par des algorithmes de Newton-Raphson, Quasi-Newton ou par l'algorithme EM (espérance-maximisation) en considérant les effets aléatoires comme des données manquantes. Les logiciels proposent de combiner ces différents algorithmes lorsqu'il y a des problèmes de convergence.

4.3 Adéquation du modèle

L'adéquation du modèle conjoint à effets aléatoires partagés s'effectue via l'adéquation des deux sous-modèles. Des prédictions marginales et spécifiques aux sujets peuvent être définies pour le modèle longitudinal. Il est préférable d'évaluer l'ajustement avec les prédictions conditionnelles aux effets aléatoires car la sortie d'étude consécutive à la survenue d'un évènement dépend des effets aléatoires du modèle linéaire mixte. Les résidus qui découlent de ces prédictions permettent notamment de vérifier les hypothèses de normalité et d'homoscédasticité des erreurs de mesures. L'adéquation du sous-modèle de survie peut être analysée en comparant les courbes de survies prédites aux estimations par la méthode de Kaplan-Meier. Les courbes de survie prédites sont estimées par :

$$1/N \sum_{i=1}^{N} S_i(t|\hat{b}_i, X_{Ti}; \hat{\theta}). \tag{1.36}$$

Les résidus de Martingales et de Schoenfeld peuvent être utilisés dans les mêmes buts que ceux décrits dans le Chapitre 1, Section 3.4. Les résidus de Martingales sont également utilisés pour évaluer la structure de dépendance entre Y et T à l'aide d'une représentation graphique des résidus versus la valeur prédite de la fonction $g_i(b_i, t)$ pour chaque individu. Elle permet de vérifier l'hypothèse de log-linéarité de $g_i(b_i, t)$.

5 Le modèle conjoint à classes latentes

Le modèle conjoint à effets aléatoires partagés décrit ci-dessus repose sur l'hypothèse d'une population homogène à la fois vis-à-vis de l'évolution du marqueur longitudinal et de la survenue de l'évènement. Or, dans de nombreux cas, les populations peuvent être hétérogènes c'est-à-dire qu'il est possible qu'il existe des groupes d'individus avec différentes évolutions du marqueur et différents risques de survenue de l'évènement. Basée sur un modèle de mélange (20), qui suppose que la population est composée de sous-populations, au sein desquelles les observations ont la même distribution, le modèle conjoint à classes latentes a été développé plus récemment ; il permet de créer des classes d'individus avec une évolution du marqueur et un risque d'évènement homogènes.

Ce modèle est une extension du modèle mixte à classes latentes défini dans le Chapitre 1, Section 2, qui permet de diviser la population en fonction de l'évolution d'un marqueur (4; 5). Tout comme le modèle à effets aléatoires partagés, le modèle conjoint à classes latentes combine un modèle linéaire mixte et un modèle de survie. La différence réside dans la structure latente qui décrit l'association entre la survenue d'un évènement et l'évolution du marqueur. Cette association est prise en compte par l'introduction des classes au sein desquelles ces 2 outcomes sont liés. L'appartenance à ces classes est définie par l'introduction d'un sous-modèle logistique multinomial.

5.1 Le modèle

Le modèle conjoint à classes latentes est composé de 3 sous-modèles : un modèle logistique multinomial qui définit la probabilité d'appartenir à une classe latente, un modèle linéaire mixte qui décrit l'évolution du marqueur dans chaque classe latente et un modèle de survie qui étudie le délai de survie dans chacune des classes.
Nous supposons que la population d'étude de N sujets est composée de G sous-populations homogènes, non identifiées. On note c_i la classe à laquelle appartient le sujet $i (i = 1, ..., N)$. La spécification de ces 3 sous-modèles est détaillée ci-après.

Le sous-modèle logistique multinomial

Comme dans le modèle linéaire mixte à classes latentes, défini dans le Chapitre 1, Section 2.1, le sous-modèle logistique multinomial pour un individu i défini la probabilité d'appartenir à la classe latente g sachant les variables explicatives X_{1i} :

$$\pi_{ig} = P(c_i = g|X_{1i}) = \frac{\exp(\xi_{0g} + X_{1i}^{\intercal}\xi_{1g})}{\sum_{l=1}^{G}\exp(\xi_{0l} + X_{1i}^{\intercal}\xi_{1l})}, \tag{1.37}$$

où l'on suppose que $\xi_{0G} = 0$ et $\xi_{1G} = 0$ afin que la classe G devienne la référence.

Le sous-modèle linéaire mixte

Le sous-modèle linéaire mixte est identique à celui développé dans le Chapitre 1, Section 2.1. Il est défini de la façon suivante pour chaque classe latente g :

$$Y_{ij} = X_{2i}(t_{ij})^{\intercal}\beta + X_{3i}(t_{ij})^{\intercal}\beta_g + Z_i(t_{ij})^{\intercal}b_i + \epsilon_{ij}, \tag{1.38}$$

avec $b_i \sim \mathcal{N}(\mu_g, B_g)$.

Le sous-modèle de survie

Tout comme le modèle linéaire mixte, chaque classe latente a un risque d'évènement spécifique qui est modélisé par la fonction de risque suivante :

$$\lambda_i(t|c_i = g) = \lambda_0(t;\zeta_g)\exp(X_{Ti}(t)^\intercal\gamma_g). \tag{1.39}$$

Les paramètres de la fonction de risque de base ζ_g, et de régression γ_g peuvent être communs ou spécifiques aux classes latentes. Il est également possible de faire une hypothèse de risque de base proportionnelle entre les classes. Dans ce cas le risque de base devient : $\lambda_0(t;\zeta)\exp^{\gamma_g}$. Le risque de base λ_0 peut également être paramétré pour différentes foncions (Weibull, Gamma) comme vu dans le Chapitre 1, Section 3.

Hypothèse d'indépendance conditionnelle

Le modèle conjoint à classes latentes admet l'hypothèse que la corrélation qui existe entre l'évolution du marqueur longitudinal et la survenue de l'évènement est entièrement prise en compte par les classes latentes. Jacqmin-Gadda *et al.* (21) ont développé un test du score permettant d'évaluer si, conditionnellement aux classes latentes, il n'existe pas de dépendance résiduelle entre le marqueur longitudinal et le risque d'évènement à travers des effets aléatoires partagés. Ce test consiste à définir un modèle conjoint à classes latentes en remplaçant le sous-modèle de survie par un modèle de survie avec des effets aléatoires partagés. L'Eq.(1.39) devient alors :

$$\lambda_i(t|c_i = g) = \lambda_{0g}(t;\zeta_g)\exp^{X_{Ti}^\intercal(t)\gamma_g + b_{ig}^{*\intercal}\eta}, \tag{1.40}$$

avec η le vecteur de paramètres associant les effets aléatoires du modèle longitudinal recentrés ($b_{ig}^* = b_i - \mu_g$) au temps d'évènement.

On teste alors si les effets aléatoires sont différents de 0 ou non c'est-à-dire si η=0. Si tel est le cas, l'hypothèse d'indépendance conditionnelle aux effets aléatoires est vérifiée. La statistique du test de score ne sera pas développée dans ce mémoire mais peut être trouvée dans la littérature (21; 3).

5.2 Estimations et inférence sur les paramètres

Tout comme pour le modèle mixte à classes latentes introduit dans le Chapitre 1, Section 2, l'estimation des paramètres du modèle conjoint à classes latentes s'effectue par le maximum de vraisemblance pour un nombre de classes fixé au préalable et dont la log-vraisemblance peut être maximisée par un algorithme de type Newton ou Marquardt (22). Les variances des paramètres peuvent être obtenues par l'inverse de la matrice Hessienne (23). Les paramètres à estimer sont ceux de l'Eq.(1.37), les paramètres du modèle linéaire mixte l'Eq.(1.14) et les paramètres du modèle de survie l'Eq.(1.39). Nous regroupons ces paramètres dans le vecteur θ_G. La log-vraisemblance peut être décomposée grâce à l'hypothèse d'indépendance des données répétées de Y et du temps d'évènement censuré à droite T, conditionnellement aux classes latentes définies par c. Ainsi, la log-vraisemblance est définie à partir de la vraisemblance pour un individu i L_i :

$$l(\theta_G) = \sum_{i=1}^{N} l_i(\theta_G) \tag{1.41}$$
$$= \sum_{i=1}^{N} \log \left(\sum_{g=1}^{G} \pi_{ig} f(Y_i|c_i = g; \theta_G) \lambda_i(T_i|c_i = g; \theta_G)^{\delta_i} S_i(T_i|c_i = g; \theta_G) \right),$$

où π_{ig} est la probabilité d'appartenir à la classe g pour i, $\lambda_i(T_i|c_i = g; \theta_G)$ est la fonction de risque instantanée définie dans l'Eq.(1.39), $S_i(T_i|c_i = g; \theta_G)$ est la fonction de survie spécifique à la classe g. $\lambda_i(T_i|c_i = g; \theta_G)^{\delta_i} S_i(T_i|c_i = g; \theta_G)$ représente la contribution d'un temps d'évènement pour des données censurées à droite. La densité des données répétées du marqueur dans la classe g est exprimée par $f(Y_i|c_i = g; \theta_G)$ où f est la densité normale. Cette densité a une expression analytique et ne nécessite donc pas d'intégration numérique comme celle d'un modèle à effets aléatoires partagés. L'intégrale sur les effets aléatoires est remplacée par une somme sur les classes latentes. Cela devient possible grâce à l'hypothèse d'indépendance entre Y et la survie, conditionnellement aux classes latentes.

D'autres méthodes d'estimation sont possibles comme l'utilisation d'un algorithme EM pour estimer les modèles conjoints (24; 25). Enfin Garre *et al.* (26) a proposé une approche bayésienne.

Le nombre de classes latentes peut être déterminé de la même manière que pour un modèle mixte à classes latente c'est-à-dire en faisant un compromis entre le BIC, le nombre d'individus par classe, le point de vue clinique et la classification *a posteriori* (voir Chapitre 1, Section 2.4).

5.3 Adéquation du modèle

L'adéquation du modèle conjoint à classes latentes se vérifie en analysant le pouvoir discriminant du modèle (via la classification *a posteriori*), le pouvoir prédictif du modèle (via les graphiques des prédictions *vs.* observations) et en vérifiant l'hypothèse d'indépendance conditionnelle définit dans le Chapitre 1, Section 5.1. Ces aspects sont détaillés ci-après.

Classification a posteriori

Comme dans le cadre des modèles linéaires mixtes à classes latente, la classification *a posteriori* permet de déterminer la probabilité d'un individu d'appartenir à une classe spécifique. Elle permet de vérifier la qualité de discrimination des classes. Une probabilité moyenne d'appartenir à la bonne classe supérieure à 80% est considérée comme suffisante. Dans les modèles conjoints à classes latentes à la différence des modèles mixtes à classes latentes, deux types de probabilités *a posteriori* peuvent être définis.

1. Probabilité conditionnelle aux données longitudinales et au temps d'évènement :

$$\hat{\pi}_{ig}^{Y,T} = P(c_i = g|Y_i, T_i, \delta_i; \hat{\theta}_G) \tag{1.42}$$
$$= \frac{\hat{\pi}_{ig} f_{Yi|c_i}(Y_i|c_i = g; \hat{\theta}_G) \lambda_i(T_i|c_i = g; \hat{\theta}_G)^{\delta_i} S_i(T_i|c_i = g; \hat{\theta}_G)}{\sum_{l=1}^{G} \hat{\pi}_{il} f_{Yi|c_i}(Y_i|c_i = l; \hat{\theta}_G) \lambda_i(T_i|c_i = l; \hat{\theta}_G)^{\delta_i} S_i(T_i|c_i = l; \hat{\theta}_G)},$$

où tous les paramètres sont définis comme dans l'Eq.(1.41). Ces probabilités permettent d'évaluer l'adéquation du modèle de plusieurs façons.

— Calculer la proportion de sujet ayant une probabilité *a posteriori* maximale au-dessus d'un seuil de 0.8 pour chaque classe obtenu *a posteriori*. Il sera donc possible de quantifier la proportion de sujets classés dans chaque classe selon le degré d'ambiguité de la classification.

— Mesurer l'entropie des données :
$1 - \frac{En}{N \log G}$ avec $En = -\sum_{i=1}^{N} \sum_{g=1}^{G} \hat{\pi}_{ig}^{Y} \log(\hat{\pi}_{ig}^{Y})$.
Plus elle sera proche de 1 plus la séparation des classes sera bonne.

— Établir une table de classification *a posteriori* qui consiste à fournir la matrice des moyennes des probabilités *a posteriori* $\hat{\pi}_{il}^{Y}$ d'appartenir à une classe l au sein de chaque classe *a posteriori* g. Une classification discriminante aura des valeurs proches de 1 pour $l = g$ et des valeurs proches de 0 pour tout $l \neq g$.

2. Probabilités conditionnelles aux données longitudinales seulement. Dans ce cas, l'équation est similaire à celle développée dans l'Eq.(1.17). Ces probabilités peuvent être utilisées pour faire la prédiction à partir des mesures répétées du marqueur longitudinal.

Le pouvoir prédictif du modèle

Tout comme dans le modèle mixte à classes latentes, deux types de prédictions spécifiques à la classe peuvent être calculés : les prédictions marginales et conditionnelles aux effets aléatoires. Pour le sujet i, la répétition au temps t_{ij} et la classe g, les prédictions marginales sont :

$$\hat{Y}_{ijg}^{(M)} = Z_{ij}^{\intercal} \hat{\mu}_g + X_{lij}^{\intercal} \hat{\beta}_g. \tag{1.43}$$

Les prédictions conditionnelles sont :

$$\hat{Y}_{ijg}^{(SS)} = Z_{ij}^{\intercal} (\hat{\mu}_g + \tilde{\hat{b}}_{ig}^{*}) + X_{lij}^{\intercal} \hat{\beta}_g, \tag{1.44}$$

où

$$\tilde{\hat{b}}_{ig}^{*} = \hat{B}_g Z_i^{\intercal} \hat{V}_{ig}^{-1} (Y_i - X_{li} \hat{\beta}_g - Z_i \hat{\mu}_g) \tag{1.45}$$

est l'estimateur empirique bayesien des effets aléatoires recentrés dans la classe g.
A partir de ces prédictions, il est possible d'obtenir deux types de résumés :

— une prédiction unique par sujet, en moyennant les prédictions sur les classes latentes

$$\hat{Y}_{ij}^{(.)} = \sum_{g=1}^{G} \hat{\pi}_{ig} \hat{Y}_{ijg}^{(.)}. \tag{1.46}$$

Ces moyennes sur les classes peuvent être comparées aux moyennes pondérées des observations $\bar{Y}_g(t) = \sum_{i=1}^{N(t)} \hat{\pi}_{ig} Y_i(t)$.

— une prédiction unique par classe, à un temps de mesure donnée t, en moyennant les prédictions sur les sujets

$$\hat{Y}_g(t)^{(.)} = \sum_{i=1}^{N(t)} \hat{\pi}_{ig} \hat{Y}_{ijg}^{(.)}, \tag{1.47}$$

avec $N(t)$ le nombre de sujets étant à risque au temps t. Ces prédictions moyennes par sujet peuvent être utilisées pour l'analyse des résidus qui permettent de vérifier les hypothèses de normalité et d'homoscédasticité des erreurs de mesures notamment comme dans le modèle à effets aléatoires partagés (Chapitre 1, Section 4.3).

A partir du sous-modèle de survie, il est possible d'obtenir des prédictions à l'aide de fonctions de survie conditionnelles aux classes $\hat{S}_{ig}$ qui peuvent être moyennées par sujet ou par classe comme les prédictions sur Y. Les courbes de survies prédites sont comparées aux courbes de survies estimées par la méthode de Kaplan-Meier pondérées par les probabilités d'appartenance aux classes. Les résidus de Martingales et de Schoenfeld peuvent être utilisés dans les mêmes buts que ceux décrits dans le Chapitre 1, Section 4.3

6 Résumé

Dans ce chapitre nous avons présenté les modèles utilisés dans la suite de ce mémoire ainsi que les modèles sur lesquels ils sont basés. Leur spécification, leur méthode d'estimation et l'étude de leur adéquation ont été mises en évidence. Les hypothèses sous-jacentes et les limites de ces modèles sont mentionnées. Les modèles linéaires mixtes et les modèles de durées font partie des sous-modèles du modèle conjoints à classes latentes que nous allons étudier plus en détails par la suite. Les modèles à effets aléatoires partagés sont une alternative aux modèles conjoints à classes latentes et seront appliqués dans le Chapitre 4.

La Figure 1.1 représente schématiquement les données analysées par les modèles conjoints à effets aléatoires partagés et à classes latentes. Ces modèles sont particulièrement intéressants pour l'étude de l'évolution d'un marqueur longitudinal censuré par un évènement.

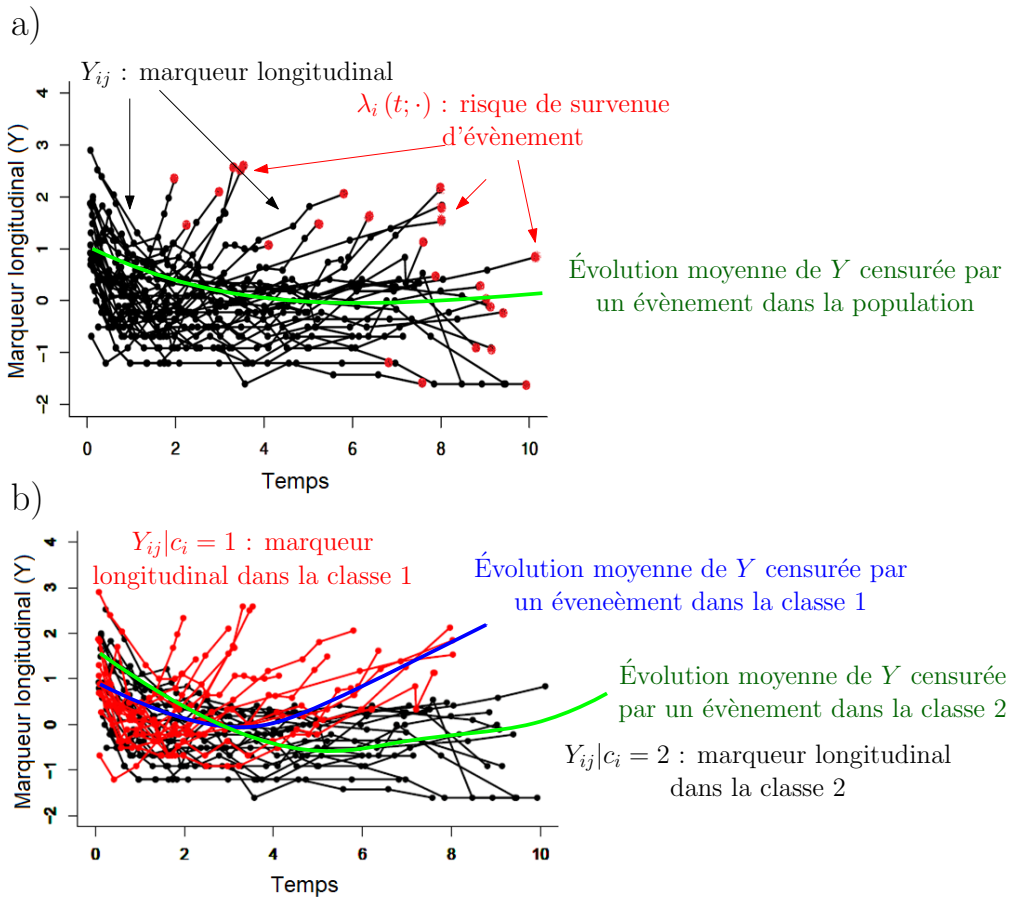


FIGURE 1.1 – Représentation schématique des évolutions des biomarqueurs dans les modèles conjoints à effets aléatoires partagés (a) et à classes latentes (b).

Chapitre 2

Étude du modèle conjoint à classes latentes

Ce chapitre consiste à répondre aux objectifs suivants :
— faire le bilan de l'utilisation des modèles conjoints à classes latentes en recherche clinique et des manquements mis en évidence ;
— mettre en évidence des règles quant à l'utilisation de covariables dans les sous-modèles selon les objectifs de l'étude ;
— mettre en évidence des propriétés quant à la taille d'échantillon (nombre de patients, nombre de mesures du facteur longitudinal, nombre d'évènements) via une étude de simulations ;
— faire un bilan des résultats trouvés afin de permettre aux cliniciens d'avoir plus de recul pour utiliser le modèle.

1 Utilisation en recherche clinique

Les modèles conjoints à classes latentes sont des modèles encore aujourd'hui très peu utilisés. Bien que leur utilité ne soit plus à démontrer, leur complexité et le manque d'implémentation au sein des logiciels ont limité leur utilisation. Depuis la publication de la revue de Proust et al. (27) en 2014, qui rend accessible leur utilisation grâce à l'explication intuitive des modèles et à leur implémentation sur **R** via le package *lcmm* (28), 17 articles utilisant ces modèles ont été publiés dans des revues cliniques. Ces articles sont décrits ci-dessous et un résumé est proposé dans le Tableau 2.1.

Á la suite de la publication de la revue de Proust *et al.* (27), une équipe française **Marioni *et al.*** (29) a publié le premier article clinique la même année, étudiant l'association entre le déclin cognitif et le risque de décès chez les personnes âgées. Cet article est un réel exemple d'application qui détaille avec précision la structure des sous-modèles, l'utilisation des covariables dans chacun d'entre eux et le choix du nombre de classes latentes. Les classes latentes trouvées sont interprétées en termes d'évolution du déclin cognitif et du risque d'évènement. Les covariables ajoutées dans le sous-modèles logistique multinomial (sexe, niveau d'éducation, emploi, engagement social) permettent d'expliquer cliniquement les différentes classes. Ces mêmes variables incluses dans le sous-modèle de durée permettent de tester leur

lien avec le risque de décès. Cette étude comportait 3653 patients avec 9 mesures de MMSE (Mini-mental state examination) par patient, pour étudier le déclin cognitif. Le nombre de décès au cours du temps s'élevait à 2921 ce qui représentait plus de 80% de la population.

Suite à ce premier article clinique, un second a ensuite été publié en 2016 par l'équipe de **Protegies *et al.*** (30), 6 en 2019 par **Stamenic *et al.*, Ogata *et al.*, Qin *et al.*, Jiang *et al.*, Brilleman *et al.*** et **Syrjala *et al.*** (31; 32; 33; 34; 35; 36), 5 ont été publiés en 2020 par **Carrier *et al.*** **Naygamon *et al.*, Raghavan *et al.*** **Khorashadizadeh *et al.*** et **Peter *et al.*** (37; 38; 39; 40; 41) et enfin 4 articles ont été publiés en 2021 par **Graillon *et al.*** **Tiruneh *et al.*, Yamanouchi *et al.*** et **Zheng *et al.*** (42; 43; 44; 45).

Bien qu'ils soient utilisés par de plus en plus de chercheurs français (4 articles sur les 17), les modèles conjoints à classes latentes commencent à être connus et utilisés mondialement. Par exemple les 13 autres articles ont été publiés par des chercheurs américains, iraniens, éthiopiens, allemands, japonais, chinois, australiens, hollandais ou encore finlandais. Leurs domaines d'applications étaient très variés tels que l'orthopédie, la neurologie, la néphrologie, la cardiologie ou encore le VIH ou la COVID-19.

Tous les articles avaient pour objectif d'établir un lien entre l'évolution d'un biomarqueur et la survenue d'un évènement. Cependant, nous avons remarqué un manque d'explication sur la construction des sous-modèles en termes d'inclusion des variables explicatives et de leur interprétation. En effet, il y a une importante hétérogénéité sur l'utilisation de ces covariables dans les 17 articles cliniques.
Par exemple, la sélection des covariables peut se faire de manière prédéfinie, comme dans l'article de Carrier *et al.* (37) ou de Marioni *et al.* (29), ou elle peut se faire à l'aide d'une sélection pas à pas, en amont ou pendant la création des classes (31; 42). Le nombre de covariables incluses dans les sous-modèles varie de 0 à 10 selon les articles.

A l'aide de ces articles cliniques et de la revue de Proust-Lima (27) nous avons pu extraire les règles suivantes quant à l'utilisation des variables explicatives dans les modèles conjoints à classes latentes.

- Lorsque l'on souhaite **connaître les facteurs associés aux classes latentes** il est possible 1) d'ajouter *a priori* les variables explicatives dans le sous-modèle logistique multinomial lorsqu'elles sont connues dans la littérature ou 2) de ne mettre aucune variable dans le sous-modèle logistique multinomial mais de comparer *a posteriori* (après la création des classes) les caractéristiques ; cela peut se faire par exemple à l'aide de régressions logistiques multinomiales multivariées.
- Lorsque l'on souhaite **connaître 1) l'impact de covariables sur le risque d'évènement ajusté sur l'évolution du marqueur ou 2) l'impact du marqueur sur le risque d'évènement ajusté sur des facteurs connus dans la littérature**, il suffit d'ajouter ces variables dans le sous-modèle de survie.

— Lorsque l'on souhaite **connaître 1) l'impact de covariables sur l'évolution du paramètre ajusté sur le risque d'évènement ou 2) l'impact du risque d'évènement sur l'évolution du marqueur ajusté sur des facteurs connus dans la littérature**, il suffit d'ajouter ces covariables dans le sous-modèle linéaire mixte.
— Les variables ajoutées dans les sous-modèles linéaire mixte et de survie peuvent être définies comme communes ou spécifiques aux classes.
— Aucune règle quant au nombre optimal de covariable à ajouter dans les sous-modèles n'a pu être mise en évidence.

De plus, au sein de ces 17 articles cliniques trouvés, les chercheurs utilisent des cohortes de tailles différentes (de 18 à plus de 1000 patients). Le nombre de mesures répétées par patient varie de 3 à 25 mesures suivant les études, le taux de censure était compris entre 0% et 95% de la population. Ces hétérogénéités nous ont amenés à constater un manque de recul concernant la taille d'échantillon (en termes de nombre d'individus, de mesures et d'évènements) et son impact sur le comportement du modèle et sur les estimations.

Malgré une revue de la littérature sur les articles statistiques basés sur les propriétés asymptotiques des modèles conjoints à classes latentes, seules quelques études de simulation concernant les modèles conjoints à classes latentes et ses extensions (risques concurrents, censure par intervalles, sous-modèle de survie multi-états) ont été trouvées (46; 47; 48; 49). Ces simulations se concentraient sur l'utilisabilité du modèle et visaient à valider la procédure d'estimation plutôt qu'à explorer les propriétés générales du modèle et ses propriétés en échantillon fini. Nous avons donc décidé de remédier à ce manque en proposant une étude de simulations plus générale, qui vise à étudier empiriquement les propriétés théoriques du modèle conjoint à classes latentes et de connaître son comportement vis-à-vis de la taille d'échantillon.

TABLEAU 2.1 – Descriptif des 17 articles cliniques portant sur l'utilisation des modèles conjoints à classes latentes.
X représente l'ensemble des covariables des sous-modèles ; X_{1i} représente la présence des covariables du sous-modèle logistique multinomial de l'Eq.(1.37) (Oui/Non) et leur nombre ; X_{2i} et X_{3i} représentent la présence des covariables du sous-modèle linéaire mixte communes et spécifiques aux classes de l'Eq.(1.14) autres que le temps (Oui/Non) et leur nombre ; X_{Ti} représente la présence des covariables du sous-modèle de survie de l'Eq.(1.39) (Oui/Non) et leur nombre. Abréviations : ANCOVA=analyse de covariance, JLCMM= Joint latent class mixed model, RLM=régression logistique multinomiale, AVC=accident vasculaire cérébral, eGFR=Epithelial Growth Factor Receptor, IMC=Indice de masse corporel, MMSE=Mini-Mental State Examination, NSP=Ne sait pas, PAS=Pression artérielle systolique, VIH=virus de l'immunodéficience humaine.

Auteurs	Pays	Année	Domaine	Objectif	Nb d'individus n	Marqueur longitudinal Y_{ij}	Evènement type et taux	Mode de sélection X	Méthode de caractérisation des classes *a posteriori*	X_{1i}	X_{2i} ou X_{3i}	X_{Ti}	Particularités
Marioni *et al.* (29)	France	2014	Neurologie	Étudier l'impact du MMSE sur le risque de décès	3656	MMSE : 9 mesures par patients	décès : 80%	*a priori*	JLCMM	Oui ; n=4	Non	Oui ; n=4	Variables dans le sous-modèle logistique pour tester leur impact sur la création des classes. Variables dans le sous-modèle de survie pour tester l'impact du MMSE sur la survie ajusté sur les facteurs pré-définis

Suite sur la page suivante

Tableau 2.1 Suite de la page précédente

Auteurs	Pays	Année	Domaine	Objectif	Nb d'individus n	Marqueur longitudinal Y_{ij}	Evènement : type et taux	Mode de sélection X	Méthode de caractérisation des classes *a posteriori*	X_{1i}	X_{2i} ou X_{3i}	X_{Ti}	Particularités
Protegis *et al.* (30)	Hollande	2016	Neurologie	Prédire le risque d'AVC à l'aide de l'évolution de la préssion artérielle systolique	6745	PAS : 5 mesures par patient	AVC : 15%	*a priori*	ANCOVA	Non	Oui : n=2	Oui : n=2	
Brilleman *et al.* (35)	Australie	2019	Nephrologie	Établir un lien entre l'évolution de l'IMC et le risque de décès et de transplantation chez les patients hémodialysés	16414	IMC : 5 mesures par patient	Décès : 34% ; Transplantation : 14%	*a priori*	JLCMM	Non	Non	Oui : n=10	
Jiang *et al.* (34)	Chine	2019	Nephrologie	Recherche de facteurs associés au profils d'évolution de l'eGFR corrélé au risque de décès chez chez les patients de diabete de type 2	6330	eGFR : 25 mesures par patient	Décès : 17%	Aucune	RLM	Non	Non	Non	Pas d'*a priori* sur les variables. Aucune dans les sous-modèles puis comparaisons des classes à l'aide d'une RLM
Ogata *et al.* (32)	Japon	2019	Cardiologie	Relation entre l'évolution de la glycémie à jeun et le risque cardiovasculaire	3120	Glucose : 3 mesures par patient	Évènement cardiovasculaire : 10%	*a priori*	RLM	Non	Oui : n=10	Oui : n=8	

Suite sur la page suivante

Tableau 2.1 Suite de la page précédente

Auteurs	Pays	Année	Domaine	Objectif	Nb d'individus n	Marqueur longitudinal Y_{ij}	Evènement : type et taux	Mode de sélection X	Méthode de caractérisation des classes *a posteriori*	X_{1i}	X_{2i} ou X_{3i}	X_{Ti}	Particularités
Qin *et al.* (33)	Chine	2019	Neurologie	Classer les patients en différents groupes de risque d'Alzheimer selon l'évolution cognitive	245	MMSE : inconnu	Alzeihmer : inconnu	*a priori*	JLCMM	Non	Oui : n=5	Oui : n=5	variables explicatives spécifiques aux classes latentes. Interprétation de leur association pour chaque profil d'évolution.
Stamenic *et al.* (31)	France	2019	Nephrologie	Étudier l'impact de la créatinine sur le risque de rejet de greffe	616	créatinine : 5 mesures par patient	Rejet de greffe : 11%	*a posteriori*	JLCMM	Oui : n=NSP	Oui : n=NSP	Oui : n=5	
Syrjälä *et al.* (36)	Finlande	2019	Néphrologie	Étudier l'impact de la consommation de viande, poisson et œuf sur le risque avancé d'auto-immunité des ilots	5545	consommation alimentaire : 5 mesures par patient	Diabète : 7%	*a priori*	NSP	Non	Non	Non	

Suite sur la page suivante

Tableau 2.1 Suite de la page précédente

Auteurs	Pays	Année	Domaine	Objectif	Nb d'individus n	Marqueur longitudinal Y_{ij}	Evènement : type et taux	Mode de sélection X	Méthode de caractérisation des classes *a posteriori*	X_{1i}	X_{2i} ou X_{3i}	X_{Ti}	Particularités
Carrier *et al.* (37)	France	2020	Orthopédie	Association entre les différents profils d'évolution de benzodiazépane et le risque de fracture de hanche et avant-bras	106437	benzodiazépane 30 mesures par patient	Diabète : 5%	*a priori*	Aucune	Non	Non	Oui : n=6	
Khorashad - izadeh *et al.* (40)	Iran	2020	Infectiologie : VIH	Estimer la survie des patients en fonction de l'évolution des CD4	213	CD4 : 5 mesures par patient	Décès : 24%	*a priori*	JLCMM	Oui : n=3	Oui : n=3	Oui : n=3	
Naymagon *et al.* (38)	USA	2020	Infectiologie : COVID-19	Étudier l'association entre l'évolution des Ddimère et 3 outcomes (décès, intubation et évènement thromboembolique (ETE))	2032	Ddimères : 3 mesures par patient au minimum	Décès et intubation : 30%, ETE : 3%	*a posteriori*	NSP	NSP	NSP	NSP	

Suite sur la page suivante

Tableau 2.1 Suite de la page précédente

Auteurs	Pays	Année	Domaine	Objectif	Nb d'individus n	Marqueur longitudinal Y_{ij}	Evènement type et taux	Mode de sélection X	Méthode de caractérisation des classes *a posteriori*	X_{1i}	X_{2i} ou X_{3i}	X_{Ti}	Particularités
Peter *et al.* (41)	Allemagne	2020	Cardiologie	Étudier l'impact de l'évolution de l'anxiété et la dépression sur le risque d'évènement cardiovasculaire	1206	Anxiété et dépréssion : 7 mesures par patient	Évènement cardio-vasculaire : 27%	*a priori*	Modèle log-binomial (50)	Non	Non	Oui : n=9	temps utilisé comme effet aléatoire partagé, Comparaison des caractéristiques entre les groupes à l'aide d'un modèle log-binomial (50)
Raghavan *et al.* (39)	USA	2020	Cardiologie	Étudier l'impact de l'évolution de la gylcémie sur le risque de décès	7780	Gycémie : 2 mesures par patient	Décès : 5%	*a priori*	Aucune	NSP	NSP	Oui : n=NSP	LCMM pour trouver les facteurs associés aux profils d'évolution de la glycémie avant JLCMM

Suite sur la page suivante

Tableau 2.1 Suite de la page précédente

Auteurs	Pays	Année	Domaine	Objectif	Nb d'individus n	Marqueur longitudinal Y_{ij}	Evènement : type et taux	Mode de sélection X	Méthode de caractérisation des classes *a posteriori*	X_{1i}	X_{2i} ou X_{3i}	X_{Ti}	Particularités
Graillon *et al.* (42)	France	2021	Neurologie	Étudier l'impact de l'évolution du volume du cerveau sur le risque de progression des patients atteints de méningiome	32	Volume cérébral : 5 mesures par patient	Décès : 100%	*a posteriori*	Aucune	Oui : n=4	Non	Non	Sélection backward pour sélectionner les facteurs à mettre dans le sous-modèle logistique multinomial
Tiruneh *et al.* (43)	Ethiopie	2021	Infectiologie : HIV	Estimer la survie des patients en fonction de l'évolution des CD4	358	CD4 cérébral : 5 mesures par patient	Décès : 27%	*a priori*	JLCMM	Oui : n=3	Oui : n=3	Oui : n=3	
Yamanouchi *et al.* (44)	Japon	2021	Nephrologie	Établir le lien entre l'évolution de l'albumine et le taux de maladie rénale au stade terminal le taux de mortalité toutes causes	329	Albumine : 9 mesures par patient	Maladie rénale : 22%, Décès : 7%	NSP	NSP	NSP	NSP	NSP	Modèle pour risques concurrents pour comparer les taux d'évènement entre les classes trouvées

Suite sur la page suivante

Tableau 2.1 Suite de la page précédente

Auteurs	Pays	Année	Domaine	Objectif	Nb d'individus n	Marqueur longitudinal Y_{ij}	Evènement : type et taux	Mode de sélection X	Méthode de caractérisation des classes *a posteriori*	X_{1i}	X_{2i} ou X_{3i}	X_{Ti}	Particularités
Zheng *et al.* (45)	Chine	2021	Cardiologie	Établir le lien entre l'évolution du peptide N-terminal pro-B-type et les évènements indésirables à court terme chez les enfants avec une maladie cardiaque congénitale	873	peptides : 4 mesures par patient	évènements indésirables : 30%	*a priori*	RLM	Non	Oui : n=7	Oui : n=7	temps utilisé comme effet aléatoire partagé

2 Étude empirique des propriétés et du comportement du modèle

Dans cette partie les résultats des simulations de Monte-Carlo pour étudier les propriétés du modèle conjoint à classes latentes sont présentés. Nous avons étudié précisément la robustesse des estimateurs en fonction du nombre d'individus, du nombre d'évènements et du degré de séparation des classes afin de répondre aux questions que se posent en pratique les cliniciens.

2.1 Design des simulations

Cadre général

Plusieurs cas de simulations sont effectués selon le nombre d'individus n, $n = \{100, 500, 1000, 5000\}$, et selon le taux de censure τ, $\tau = \{5\%, 10\%, 15\%, 25\%, 50\%\}$, ce qui permet d'explorer les deux directions asymptotiques possibles : le nombre d'individus et le nombre d'évènements observé (51). La capacité du modèle à distinguer les classes latentes est étudiée en considérant deux cas différents en termes de séparation des classes : *Grande separation* GS (les classes sont très différentes en termes d'évolution longitudinale du marqueur) et *Faible separation* FS (les classes sont assez similaires).
Étant donné la complexité de la fonction de vraisemblance, l'algorithme d'optimisation rencontre parfois des problèmes de convergence. C'est pourquoi, pour chaque cas (en termes de n, τ et de séparation des classes), 120 ensembles de données ont été générés afin de garantir l'obtention d'au moins 100 résultats par cas.

La distribution de chacun des paramètres estimés a ensuite été analysée en termes de normalité, de biais relatif et de taux de couverture.

- **La normalité** a été évaluée graphiquement par des diagrammes quantile-quantile plutôt que par les tests de normalité qui auraient souvent rejetés l'hypothèse nulle en raison de valeurs aberrantes dans les estimations des paramètres. Cette situation est probable en raison de la complexité de la vraisemblance ; elle se traduit par des maximum locaux et/ou d'une puissance de test élevée.
- **Le biais relatif** en valeur absolue pour un paramètre θ estimé par $\hat{\theta}$ à partir d'un échantillon de taille n sur K simulations est calculé comme suit :

$$RB(\theta, n) = \left| \frac{\frac{1}{K}\sum_{h=1}^{K} \hat{\theta}_{n,h} - \theta}{\theta} \right|,$$

  avec $\frac{1}{K}\sum_{h=1}^{K} \hat{\theta}_{n,h}$ la moyenne empirique d'estimation du paramètre calculée à partir de n individus sur K simulations de Monte-Carlo, et θ la vraie valeur du paramètre.
- **Le taux de couverture** était calculé pour chaque cas de simulation et représentait le nombre de fois où la vraie valeur était comprise dans l'intervalle de confiance de chaque estimation.
- La capacité du modèle à **distinguer les classes latentes** est évaluée par le pourcentage de patients bien classés.

Le cadre des simulations est schématiquement représenté sur le Tableau 2.2.

TABLEAU 2.2 – Plan de simulation ; 120 jeux de données ont été simulés pour chaque cas en termes de nombre de patients n, taux de censure τ, niveau de séparation. Abréviations : FS= Faible séparation ; GS= Grande séparation.

Taux de censure τ	Degré de séparation	Nombre de patients			
		100	500	1000	5000
50%	GS				
	FS	Paramètres calculés :			
25%	GS				
	FS	Normalité			
15%	GS	Biais relatif			
	FS	Taux de couverture			
10%	GS	Taux de bon classement			
	FS				
5%	GS				
	FS				

Simulation de données

Les vraies valeurs des paramètres ont été choisies pour imiter les données réelles de l'article de Stamenic *et al.* (31), portant sur un outil de pronostic pour la prédiction individualisée du risque d'échec du greffon dans les dix ans suivant la transplantation rénale, en utilisant la progression de la créatinine sérique comme marqueur longitudinal. Nous avons choisi cet article car il nous semblait intuitif et suffisamment détaillé pour comprendre la construction du modèle et l'interprétation des résultats. Les données ont été simulées à partir des sous-modèles présentés dans le Chapitre 1, Section 5.1 (Eq.(1.37) - Eq.(1.39)) :

$$
\begin{cases}
\pi_{i1} = \text{Constant} \\
\quad \text{pour un modèle à 2 classes : } \xi_{01} = \ln\left(\frac{\pi_{i1}}{1-\pi_{i1}}\right), \text{ voir Eq.(1.37)} \\
Y_{ij}|(c_i = g) = \beta_{0g} + \beta_{1g} t_{ij} + b_i + \epsilon_i, \text{voir Eq.(1.14)} \\
\quad b_{ig} \sim \mathcal{N}\left(0, \sigma^2_{b,g}\right), \epsilon_{ig} \sim \mathcal{N}\left(0, \sigma^2_{\epsilon,g}\right) \\
S(t)|(c_i = g) = \exp\left(-\left(\frac{t}{\zeta_{1g}}\right)^{\zeta_{2g}}\right) \\
\quad T^\star \sim \mathcal{W}eibull\left(\zeta_{1g}, \zeta_{2g}\right) \\
\quad \lambda_i(t|c_i = g) = \zeta_{1g}^{\zeta_{2g}} \zeta_{2g} t^{\zeta_{2g}-1}, \text{voir Eq.(1.21) et Eq.(1.39)} \\
M(\tilde{t})|(c_i = g) = \exp\left(-\left(\frac{\tilde{t}}{\tilde{\zeta}_{1g}}\right)^{\tilde{\zeta}_{2g}}\right) \\
\quad \tilde{T} \sim \mathcal{W}eibull\left(\tilde{\zeta}_{1g}, \tilde{\zeta}_{2g}\right) \\
\quad \lambda_i(\tilde{t}|c_i = g) = \tilde{\zeta}_{1g}^{\tilde{\zeta}_{2g}} \tilde{\zeta}_{2g} \tilde{t}^{\tilde{\zeta}_{2g}-1}, \text{voir Eq.(1.21) et Eq.(1.39).}
\end{cases}
$$

Le vecteur de paramètres pour le modèle à deux classes latentes avec un effet aléatoire commun aux classes et la variance de l'erreur du sous-modèle mixte est

donc le suivant :

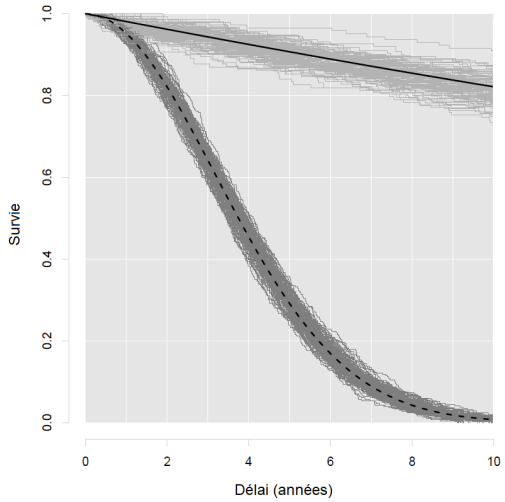
$$\boldsymbol{\theta} = \Big(\xi_{01}, \beta_{01}, \beta_{11}, \beta_{02}, \beta_{12}, \sigma_b^2, \sigma_\epsilon^2, \zeta_{11}, \zeta_{21}, \zeta_{12}, \zeta_{22}\Big). \tag{2.1}$$

Les paramètres ont été choisis de la manière suivante.

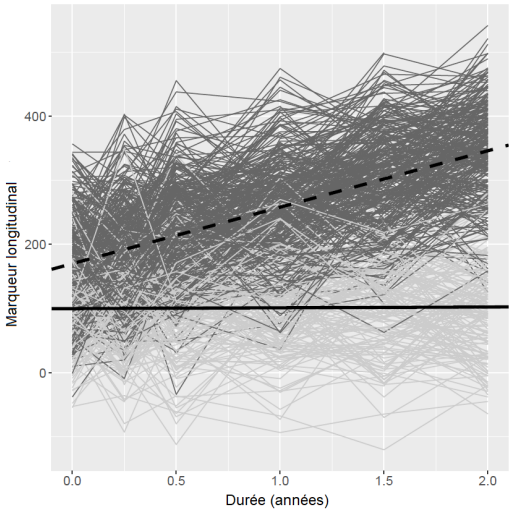
1. **Modèle logistique multinomial** : notons que le fait qu'il n'y ait pas de covariable dans le **modèle logistique** pour l'appartenance à une classe implique une probabilité constante pour l'appartenance à chaque classe.
   La probabilité d'appartenance à la classe 1 a été fixée à 0,3 dans les deux cas (*Grande séparation* et *Faible séparation*), ce qui a donné le paramètre du modèle logistique de l'Eq.(1.37) $\xi_{01} = -0,84$.
2. Le **modèle longitudinal** considéré est un modèle linéaire mixte à intercept aléatoire et il implique que dans l'Eq.(1.14), X_{2ij}^T est une matrice nulle (pas de covariables communes aux classes) et $X_{3ij}^T = (1 \quad t_{ij})$, seul le temps est ajouté au modèle pour créer des classes avec des évolutions temporelles différentes. Le modèle s'écrit alors : $Y_{ij}|(c_i = g) = \beta_{0g} + \beta_{1g}t_{ij} + b_{0i} + \epsilon_i$. Les temps de mesures répétées pour le marqueur longitudinal sont fixés à 1, 3, 6, 12, 18 et 24 mois selon l'article de Stamenic *et al.*(31). Les vrais valeurs des paramètres de ce sous-modèle étaient choisis comme suit :
   (a) cas d'une *GS*. Les valeurs sont directement tirées de l'article de Stamenic *et al.* (31) : $\beta_{01} = 170$, $\beta_{02} = 100$, $\beta_{11} = 88$ par an, $\beta_{12} = 1,2$ par an, $\sigma_{b,1}^2 = \sigma_{b,2}^2 = 50$;
   (b) cas d'une *FS*. Pour ce cas, les valeurs du modèle linéaire mixte du cas de *Grande séparation* ont été divisées par deux afin d'obtenir des classes relativement similaires en termes d'évolution longitudinale des marqueurs et les paramètres aléatoires $\sigma_{b,1}^2$et$\sigma_{b,2}^2$ n'ont pas été modifiés, ce qui donne : $\beta_{01} = 135$, $\beta_{02} = 100$, $\beta_{11} = 44$ par an, $\beta_{12} = 1,2$ par an.

   Les exemples de trajectoires simulées pour les cas de *GS* et de *FS* sont illustrés sur les Figures 2.1b et 2.1c où nous observons que les trajectoires longitudinales observées sont plutôt confuses dans le cas *FS* par rapport au cas *GS*.
3. **Modèle de survie** : les **distributions de survie et de censure** considérées impliquent que les temps de survie et de censure sont des variables aléatoires de Weibull. Les paramètres de la distribution des temps d'évènement ont été choisis pour ressembler aux données de l'article. $M(t)$ étant la fonction de survie de la distribution de censure et $\tilde{T}$ le temps de censure.

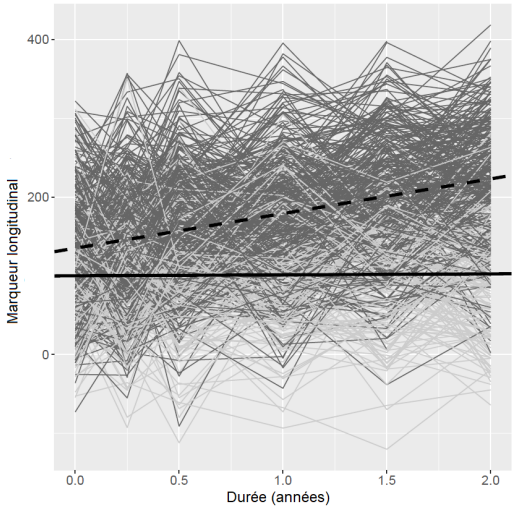
   Les vraies valeurs des paramètres étaient choisis comme suit pour le cas d'une *GS* et *FS* : $\zeta_{11} = 4,5$, $\zeta_{21} = 2$, $\zeta_{12} = 50$, $\zeta_{22} = 1,01$.
   Dans les deux cas (en termes de séparation des classes), le paramètre de forme de la distribution de Weibull pour la censure a été fixé à 1,5. Ce choix est inspiré de la vie réelle, où la probabilité de censure augmente avec le temps. Le paramètre d'échelle de cette distribution a été dérivé empiriquement en fonction du paramètre de forme pour répondre au taux de censure requis.



(a) courbes de survie pour 2 classes (même résultats pour les cas de *GS* et *FS*).



(b) courbes d'évolution du marqueur longitudinal pour le cas d'une *GS*.



(c) courbes d'évolution du marqueur longitudinal pour le cas d'une *FS*.

FIGURE 2.1 – Courbes de survie et trajectoires du marqueur longitudinal simulées : un exemple avec n=500, $\tau = 5\%$. Classe 1 : trajectoires individuelles en gris foncé, pointillés pour la trajectoire moyenne ; Classe 2 : trajectoires individuelles en gris clair, ligne continue pour la trajectoire moyenne.

2.2 Interprétation des résultats

Dans cette partie nous présentons les résultats des simulations qui permettent d'évaluer les propriétés du modèle conjoint à classes latentes selon les critères définis dans le Chapitre 2, Section 2.1 : normalité, biais relatif, taux de couverture, taux de bon classement. Les tendances générales seront d'abord énoncées, suivies par les recommandations pratiques basées sur ces tendances (Chapitre 2, Section 2.3).

Normalité des paramètres du modèle

La normalité des paramètres estimés est évaluée en traçant des diagrammes quantile-quantile pour chaque paramètre de chaque classe, chaque taille d'échantillon n et chaque taux de censure τ.

En ce qui concerne le cas de *GS*, plus le taux de censure est petit, plus la normalité des paramètres est respectée ; une forte censure (50%) implique des déviations de la normalité, notamment pour les paramètres du sous-modèle de survie (Figure 2.2 et dans l'Annexe Figures A.1 à A.3). Cette déviation est d'autant plus importante pour n petit, par exemple pour 100 individus, taux de censure 5% et 50% dans le cadre *Grande séparation* (Annexe Figure A.4). Cependant, comme le montre les Figures A.2 et A.3 dans l'Annexe, la déviation diminue lorsque le nombre d'individus augmente.

Il convient de noter que la normalité des paramètres du sous-modèle longitudinal n'est pas fortement influencée par une petite taille d'échantillon et/ou une forte censure. De même, la normalité de l'estimateur du maximum de vraisemblance n'est pas considérablement influencée par le degré de séparation des classes. Toutefois, cette conclusion doit être considérée avec prudence, car elle peut être différente pour différents degrés de séparation des classes.

Indépendamment du taux de censure, les écarts par rapport à la normalité diminuent avec l'augmentation du nombre d'individus (Figure 2.3 pour le paramètre d'échelle et les paramètres de forme de Weibull).

Dans le cas d'une *FS* les résultats de la normalité des paramètres sont similaires à ceux trouvés pour le cas d'une *GS* (Annexe, Figures A.5 à A.8). Les écarts à la normalité surviennent surtout pour les forts taux de censures pour les paramètres du sous-modèle de survie (Annexe, Figure A.9).

Il est à noter que la plupart des articles traitant des propriétés asymptotiques des modèles de survie se concentrent sur les coefficients de régression et très peu sur les paramètres de la distribution de Weibull. Sirvanci et Yang (52) dérivent la normalité asymptotique des paramètres du modèle de Weibull pour les données de censure de type I (durée de suivi fixe). Cependant, dans notre étude, empiriquement, les écarts à la normalité sont signalés pour des échantillons de petites tailles en termes de nombre d'évènements et/ou de nombre d'individus ; en ce sens, le problème de normalité n'est pas spécifique au modèle conjoint à classes latentes, mais est plutôt hérité de l'analyse de survie. Les courbes de survies dérivées des différents paramètres estimés sont présentés sur la Figure 2.4. On note que pour un faible taux de censure, un petit nombre d'individu montre une plus grande variation des estimations (Figures 2.4a) et 2.4c).

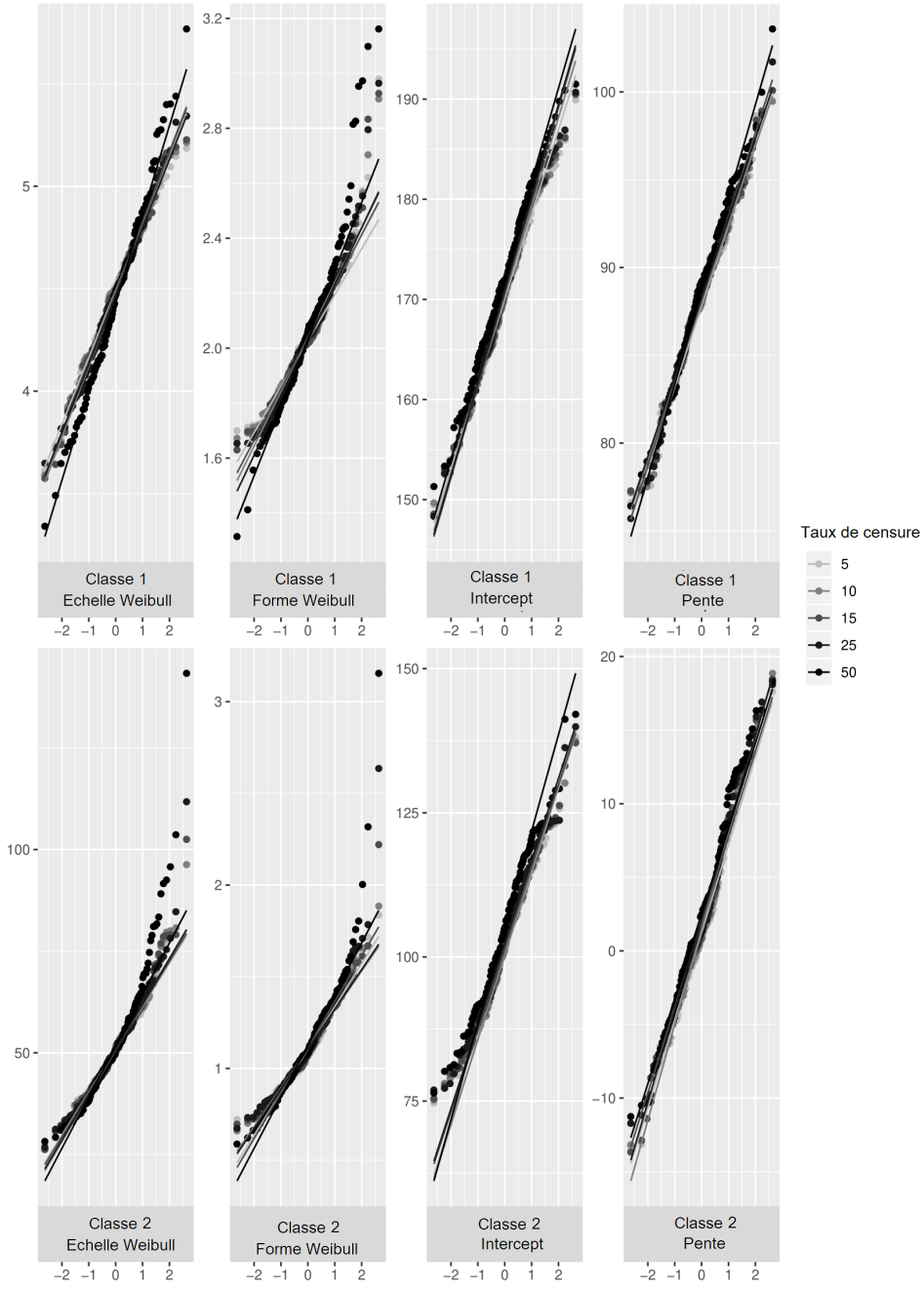


FIGURE 2.2 – Diagramme quantile-quantile selon le taux de censure en %. n=100, cas de *GS*. Les résultats pour les paramètres ζ_{11} (échelle de Weibull pour la classe 1), ζ_{21} (forme de Weibull pour la classe 1), ζ_{12},(échelle de Weibull pour la classe 2), ζ_{22} (forme de Weibull pour la classe 2), β_{01} (intercept de la classe 1), β_{02} (intercept de la classe 2), β_{11} (pente de la classe 1) et β_{22} (intercept de la classe 2).

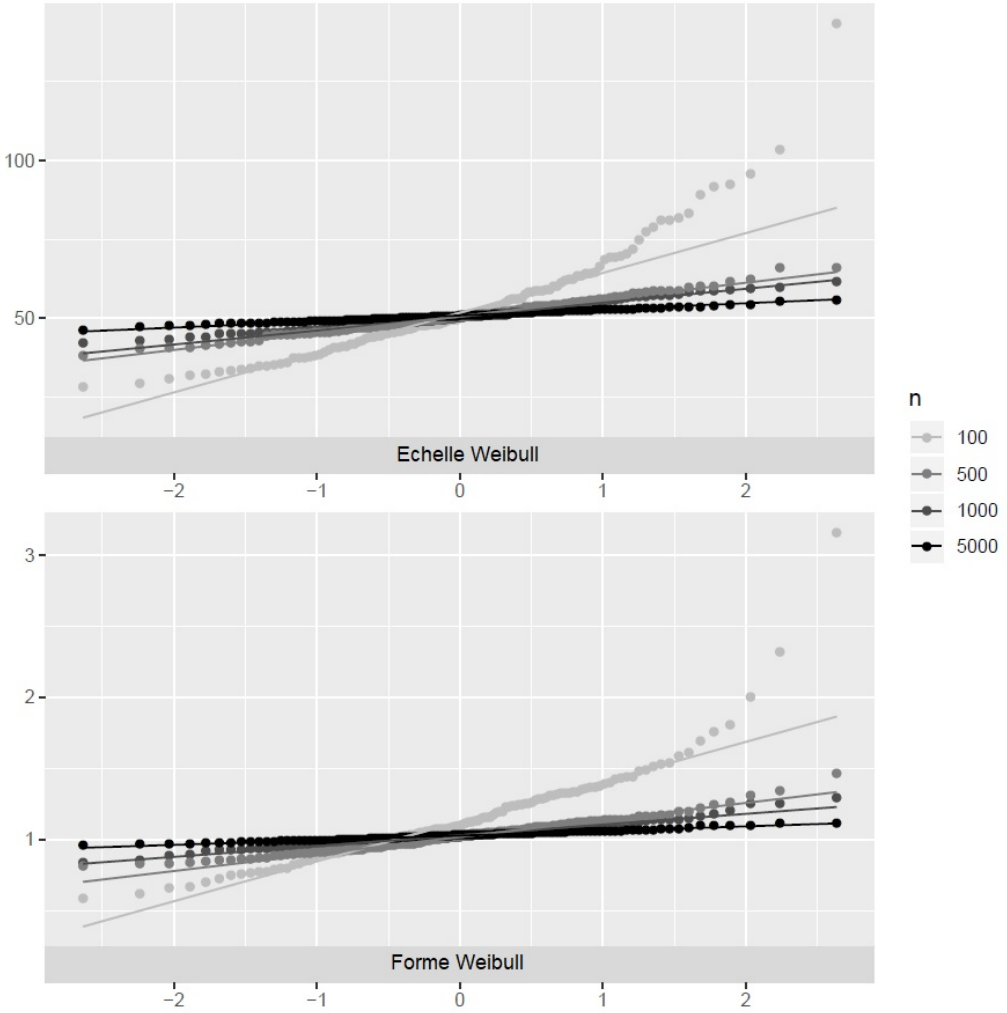
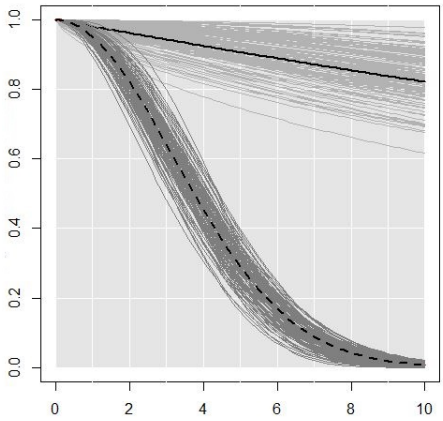
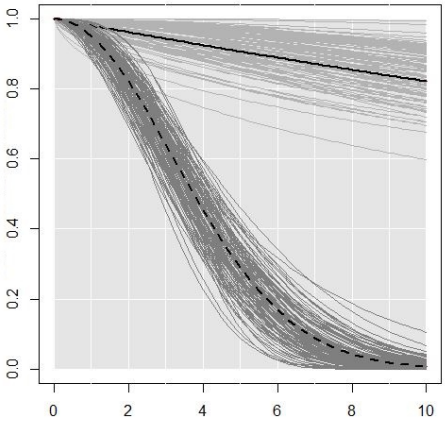


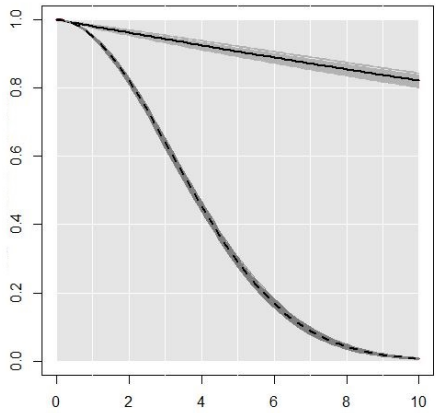
FIGURE 2.3 – Diagramme quantile-quantile pour les estimations des paramètres dans la classe 2, cas *Grande séparation*, $\tau = 5\%$. Résultats pour les paramètres d'échelle (ζ_{12}) et de forme(ζ_{22}) de Weibull selon le nombre de patients.



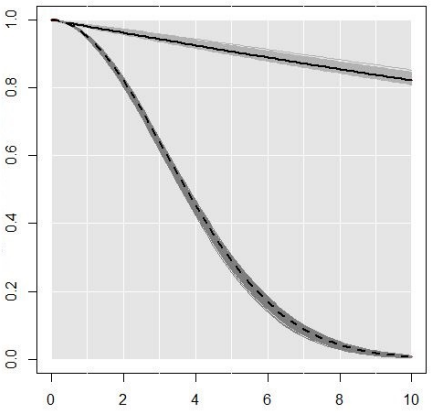
(a) courbes de survie pour 2 classes n=100, τ=5%



(b) courbes de survie pour 2 classes n=100, τ=50%



(c) courbes de survie pour 2 classes n=5000, τ=5%



(d) courbes de survie pour 2 classes n=5000, τ=50%

FIGURE 2.4 – Courbes de survies estimées à partir des jeux de données simulés. Classe 1 : trajectoires individuelles en gris foncé, pointillés pour la trajectoire moyenne ; classe 2 : trajectoires individuelles en gris clair, ligne continue pour la trajectoire moyenne.

Évaluation du biais relatif

Le biais relatif (RB) des estimations des paramètres spécifiques à la classe du modèle linéaire mixte et du modèle de survie est illustré dans la Figure 2.5 et la Figure 2.6 pour le cas d'une *GS* et dans la Figure 2.7 et la Figure 2.8 pour le cas d'une *FS*. Les résultats numériques détaillés sont fournis dans l'Annexe dans les Tableaux B.1 et B.2 pour les cas de *GS* et *FS* respectivement.

Les tendances générales du biais relatif et de son évolution par rapport à la taille de l'échantillon et du taux de censure dépendent du paramètre du modèle et du degré de séparation des classes. En ce qui concerne les paramètres de variance (la variance de l'erreur et de l'effet aléatoire dans le sous-modèle linéaire mixte), il n'y a pas de tendance claire du biais relatif ; pour les paramètres suivants en revanche, des tendances ont pu être mis en évidence.

— Concernant les **valeurs absolues** du biais relatif, dans le cas d'une *GS* (Figure 2.6 et Tableau B.1 dans l'Annexe), le RB est plus important pour deux paramètres de la classe 2 :

1. **le paramètre de forme Weibull** du sous-modèle de survie : RB supérieur à 10% pour un petit nombre d'individus
2. **le paramètre de pente** du sous-modèle linéaire mixte : RB varie de 10% à 120% selon le nombre d'individus et le taux de censure. Le nombre moyen de mesures du marqueur dans le pire des cas (100 patients et un taux de censure de 50%) est de 5,1.

Pour les autres paramètres, le RB ne dépasse pas 10%. La tendance est assez similaire pour le cas d'une *FS* (Figure 2.8 et Tableau B.2 dans l'Annexe), mais dans une plus grande mesure : le RB varie de plus de 30% à 530% dans le pire des cas (petit n et haut τ) pour le paramètre de la pente.

— Concernant l'impact du **taux de censure sur le biais**, le RB augmente linéairement avec le taux de censure pour un nombre d'individus donné. Cette tendance est la même pour les deux degré de séparation des classes (*GS* et *FS*), bien qu'elle soit d'une plus grande mesure pour le cas d'une *FS* comme pour la valeur absolue.
Plus précisément, dans le cas d'une *GS*, en passant d'un taux de censure de 5 à 50%, le RB augmente d'environ 1% pour les paramètres de la classe 1 (Figure 2.5) (*vs.* 2-8% pour une *FS*, Figure 2.7) et d'environ 3-5% pour les paramètres de la classe 2 (Figure 2.6)(*vs.* 2-15% pour une *FS*, Figure 2.8). Concernant le paramètre de pente du modèle linéaire mixte de la classe 2 on observe une augmentation de 100% du RB dans le cas d'une *GS* (Figure 2.6) (*vs.* 400% dans le cas de *FS*, Figure 2.8) pour τ allant de 5% à 50%.
Notons que la tendance linéaire de l'évolution du RB en fonction de τ n'est pas toujours respectée pour les petits n (exemple du paramètre de forme pour la classe 1 et *GS*, Figure 2.5).

— Bien que l'augmentation du **nombre d'individus** n ne semble pas avoir un impact fort sur le RB, le paramètre de forme semble être plus influencé que le paramètre d'échelle de Weibull. De même, le cas d'une *faible séparation* est plus influencé que le cas d'une *Grande séparation*.

En résumé, le biais relatif est plus sensible au taux de censure qu'au nombre de patients. Les paramètres les plus impactés sont la pente du modèle mixte et le

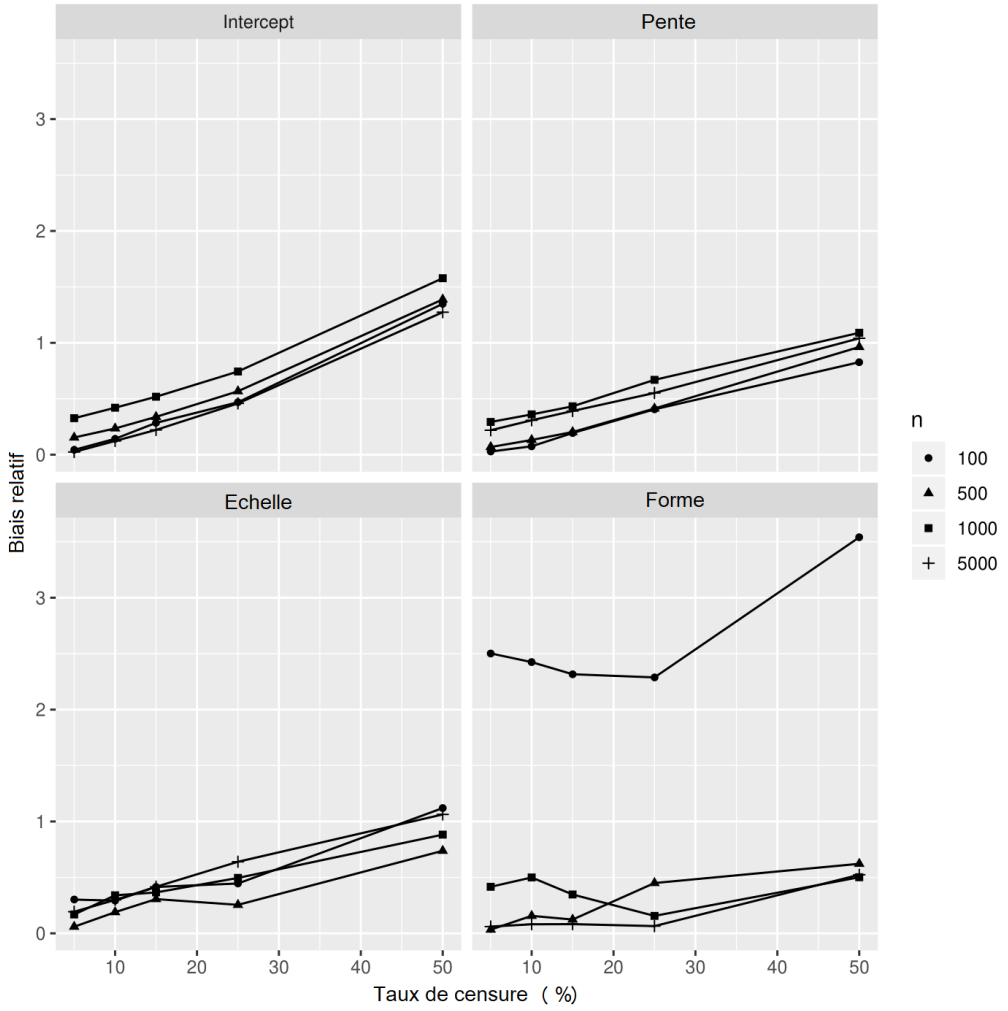


FIGURE 2.5 – Biais relatif (en %) des estimations des paramètres ζ_{11} (échelle de Weibull), ζ_{21} (forme de Weibull), β_{01} (intercept), β_{11} (pente) de la classe 1 en fonction du taux de censure τ et du nombre d'individus n, cas d'une *Grande separation GS*. La même échelle des abscisses est utilisée pour les quatre figures.

paramètre de forme du modèle de Weibull. Notez néanmoins que la classe 2 a le plus petit nombre de patients avec un risque de décès plus faible ; par conséquent, les paramètres de cette classe sont plus affectés par le taux de censure. De plus, le biais relatif élevé pour le paramètre de pente de la classe 2 s'explique par la faible valeur théorique de ce paramètre ($\beta_{12} = 1,2$).

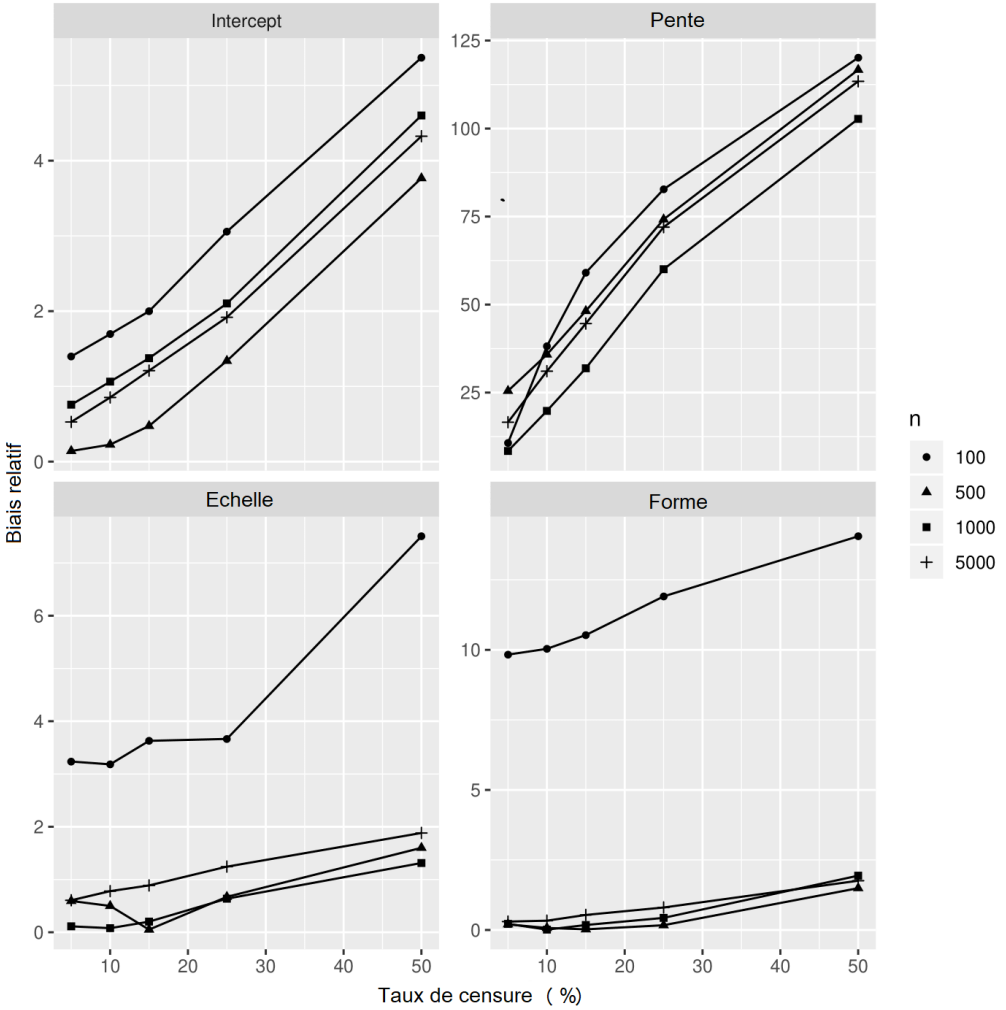


FIGURE 2.6 – Biais relatif des estimations des paramètres ζ_{12} (échelle de Weibull), ζ_{22} (forme de Weibull), β_{02} (intercept), β_{12} (pente) de la classe 2 en fonction du taux de censure τ et du nombre d'individus n, cas d'une *Grande separation GS*. La même échelle des abscisses est utilisée pour les quatre Figures.

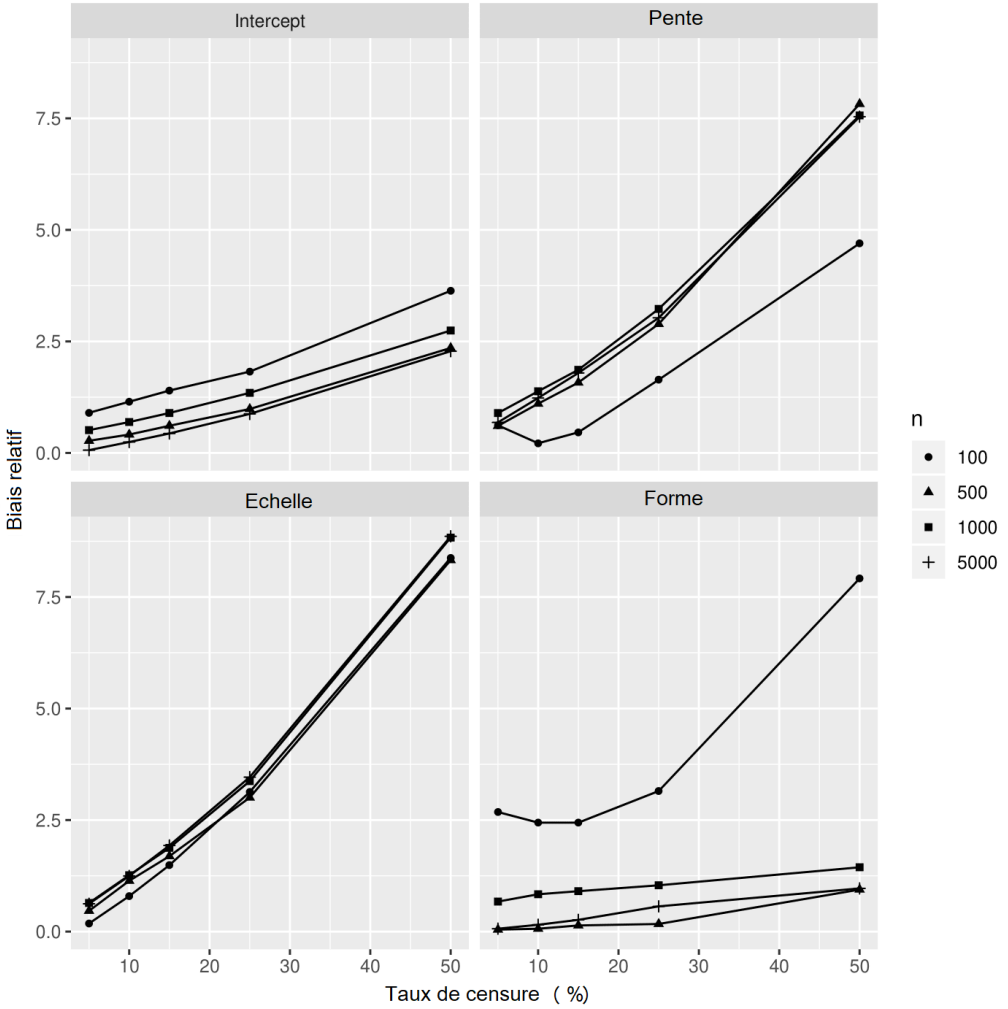


FIGURE 2.7 – Biais relatif des estimations des paramètres ζ_{11} (échelle de Weibull), ζ_{21} (forme de Weibull), β_{01} (intercept), β_{11} (pente) de la classe 1 en fonction du taux de censure τ et du nombre d'individus n, cas d'une *Faible separation FS*. La même échelle des abscisses est utilisée pour les quatre Figures.

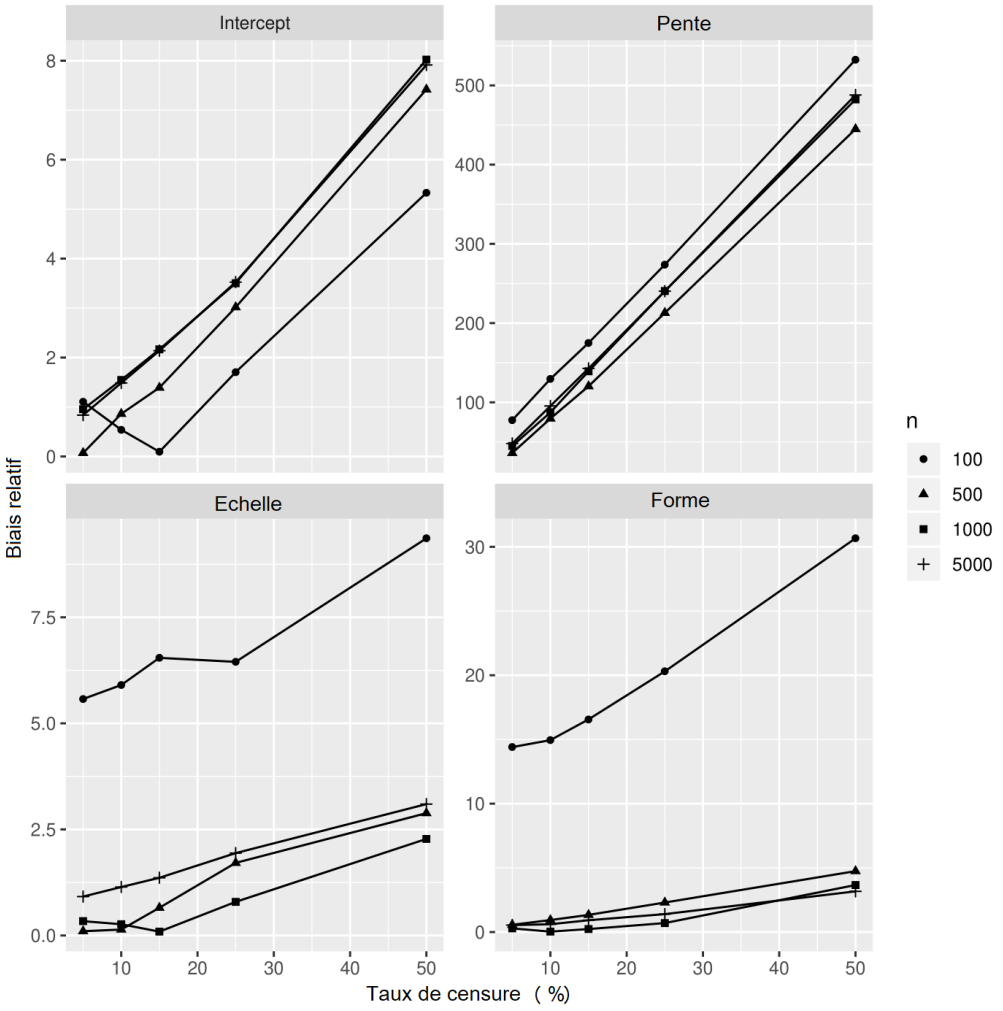


FIGURE 2.8 – Biais relatif des estimations des paramètres ζ_{12} (échelle de Weibull), ζ_{22} (forme de Weibull), β_{02} (intercept), β_{12} (pente) de la classe 2 en fonction du taux de censure τ et du nombre d'individus n, cas d'une *Faible separation FS*. La même échelle des abscisses est utilisée pour les quatre Figures.

Évaluation du taux de couverture

Le taux de couverture est globalement satisfaisant avec un taux autour de 95% pour tous les paramètres étudiées dans les cas de *GS* (Tableau 2.3). Cependant, le taux de recouvrement semble légèrement diminuer lorsque le taux de censure augmente. Cette diminution est d'autant plus importante lorsque la taille de l'échantillon est grande en termes de nombre d'individus, ce qui se traduit par des intervalles de confiance plus petits, entraînant un taux de couverture empirique plus faible. Les déviations de la normalité déjà mentionnées pour ces paramètres peuvent également être une cause de ce phénomène.

TABLEAU 2.3 – Résultats des simulations : taux de couverture en fonction du nombre d'individus, n, et du taux de censure, τ, pour le cas d'une *GS*. Les résultats pour l'intercept et la pente du sous-modèle longitudinal ($\hat{\beta}_{0g}$, $\hat{\beta}_{1g}$ respectivement) et pour l'échelle et la forme de Weibull du sous-modèle de survie ($\hat{\zeta}_{1g}$ et $\hat{\zeta}_{2g}$ respectivement) sont présentés. g : identification des classes.

n	τ	$\hat{\beta}_{0g}$		$\hat{\beta}_{1g}$		$\hat{\zeta}_{1g}$		$\hat{\zeta}_{2g}$	
		$g=1$	$g=2$	$g=1$	$g=2$	$g=1$	$g=2$	$g=1$	$g=2$
100	5	0,9664	0,9496	0,9328	0,9496	0,8824	0,8655	0,9580	0,9160
	10	0,9664	0,9496	0,9328	0,9412	0,8571	0,8655	0,9496	0,9076
	15	0,9664	0,9496	0,9412	0,9496	0,8824	0,8908	0,9412	0,9328
	25	0,9580	0,9496	0,9412	0,9580	0,8487	0,7899	0,9496	0,9076
	50	0,9328	0,9076	0,9748	0,9328	0,8403	0,7899	0,8824	0,8571
500	5	0,9667	0,9667	0,9833	0,9500	0,9250	0,9167	0,9333	0,9500
	10	0,9667	0,9750	0,9833	0,9417	0,9583	0,9250	0,9250	0,9500
	15	0,9667	0,9667	0,9750	0,9750	0,9250	0,9083	0,9417	0,9333
	25	0,9583	0,9500	0,9750	0,9417	0,8667	0,7750	0,9083	0,9167
	50	0,8750	0,9167	0,9333	0,9083	0,9000	0,8333	0,9333	0,9250
1000	5	0,9833	0,9333	0,9500	0,9250	0,8667	0,8750	0,9500	0,9417
	10	0,9750	0,9167	0,9500	0,9333	0,8833	0,9417	0,9583	0,9417
	15	0,9750	0,9167	0,9333	0,9333	0,8833	0,8917	0,9833	0,9583
	25	0,9500	0,9000	0,9167	0,9083	0,8917	0,8917	0,9417	0,9000
	50	0,8667	0,8000	0,8917	0,9083	0,9000	0,7000	0,9083	0,9167
5000	5	0,9750	0,9083	0,9333	0,8750	0,8917	0,9000	0,9667	0,9500
	10	0,9833	0,9083	0,9417	0,8583	0,9000	0,9250	0,9500	0,9417
	15	0,9667	0,8667	0,9083	0,8333	0,8250	0,8750	0,9417	0,9333
	25	0,9333	0,7917	0,9000	0,8250	0,8167	0,8667	0,8917	0,9333
	50	0,5000	0,2833	0,8000	0,7167	0,7750	0,7750	0,8500	0,9000

Concernant le cas de *FS* (Tableau 2.4), le taux de recouvrement est globalement plus faible que pour la cas de *GS* avec des taux autours de 85%. Le taux de recouvrement est très impacté par l'augmentation des taux de censure quelque soit le nombre de patient (taux de recouvrement pouvant atteindre les 0% pour n=5000 et taux de censure=50%).

TABLEAU 2.4 – Résultats des simulations : taux de couverture en fonction du nombre d'individus, n, et du taux de censure, τ, pour le cas d'une *FS*. Les résultats pour l'intercept et la pente du sous-modèle longitudinal ($\hat{\beta}_{0g}$, $\hat{\beta}_{1g}$ respectivement) et pour l'échelle et la forme de Weibull du sous-modèle de survie ($\hat{\zeta}_{1g}$ et $\hat{\zeta}_{2g}$ respectivement) sont présentés. g : identification des classes.

n	τ	$\hat{\beta}_{0g}$		$\hat{\beta}_{1g}$		$\hat{\zeta}_{1g}$		$\hat{\zeta}_{2g}$	
		$g=1$	$g=2$	$g=1$	$g=2$	$g=1$	$g=2$	$g=1$	$g=2$
100	5	0,9580	0,9496	0,9244	0,9328	0,8151	0,7899	0,9076	0,8571
	10	0,9412	0,9412	0,9244	0,8908	0,8235	0,7899	0,9076	0,8487
	15	0,9496	0,9328	0,9580	0,9160	0,8403	0,8403	0,9160	0,8487
	25	0,9580	0,8908	0,8486	0,8823	0,8739	0,8571	0,9160	0,8655
	50	0,9496	0,8151	0,8655	0,7479	0,6722	0,7815	0,8403	0,7731
500	5	0,9583	0,9500	0,9667	0,9583	0,9083	0,8833	0,9167	0,9333
	10	0,9583	0,9500	0,9667	0,9250	0,8833	0,8917	0,8417	0,9500
	15	0,9417	0,9417	0,9667	0,8917	0,8750	0,8917	0,9333	0,9167
	25	0,9250	0,9083	0,9250	0,8000	0,8333	0,8500	0,9417	0,8750
	50	0,8750	0,7250	0,8000	0,5417	0,5333	0,8500	0,9167	0,9083
1000	5	0,9667	0,9167	0,9583	0,9417	0,9167	0,9250	0,9667	0,9583
	10	0,9667	0,9167	0,9417	0,9083	0,8883	0,9167	0,9667	0,9667
	15	0,9667	0,9000	0,9417	0,8833	0,8667	0,9000	0,9417	0,9333
	25	0,9333	0,8500	0,8667	0,7333	0,7333	0,9167	0,9667	0,9667
	50	0,8167	0,4417	0,6583	0,2000	0,2167	0,8083	0,9000	0,9333
5000	5	0,9833	0,9083	0,9583	0,8667	0,8417	0,9000	0,9667	0,9667
	10	0,9917	0,8250	0,8667	0,7833	0,7167	0,8917	0,9083	0,9250
	15	0,9583	0,7250	0,8250	0,6250	0,5583	0,9000	0,9417	0,9417
	25	0,8417	0,5083	0,5750	0,1667	0,1167	0,8667	0,8333	0,9417
	50	0,4000	0	0,0083	0	0	0,7667	0,8750	0,8750

Évaluation du taux de bon classement

La qualité de la prédiction de l'appartenance à une classe est globalement satisfaisante (Tableau 2.5) : elle est supérieure à 90% pour la majorité des paramètres en termes de n et de τ. Cependant, cette qualité est globalement plus faible pour le cas d'une *FS* (moins de 95% par rapport à un taux supérieur à 95% pour le cas d'une *GS*) et pour un fort taux de censure (83-85% pour le cas d'une *FS*, taux de censure 50%). Un taux de censure décroissant entraîne une amélioration de 1 à 3% de l'identification des classes pour tous les n, à l'exception du plus fort taux de censure (τ=50%). La taille de l'échantillon n n'influence pas considérablement la qualité des prédictions. La qualité de l'identification de l'appartenance à une classe dépend donc du nombre d'évènements observé plutôt que du nombre d'individus observé.

TABLEAU 2.5 – Résultats des simulations : taux de bon classement, calculé comme le taux d'appartenance aux classes correctement prédites, en fonction du nombre d'individus, n, et du taux de censure, τ. La différence entre les taux pour les cas de *GS* et *FS* est représentée.

n	τ	*GS*	*FS*	Différence
100	5	0,9760	0,9418	-0,0342
	10	0,9767	0,9347	-0,0420
	15	0,9748	0,9248	-0,0500
	25	0,9689	0,9039	-0,0650
	50	0,9556	0,8335	-0,1221
500	5	0,9790	0,9440	-0,0350
	10	0,9778	0,9376	-0,0402
	15	0,9764	0,9321	-0,0443
	25	0,9720	0,9148	-0,0572
	50	0,9586	0,8458	-0,1128
1000	5	0,9814	0,9477	-0,0337
	10	0,9798	0,9419	-0,0379
	15	0,9782	0,9354	-0,0428
	25	0,9745	0,9186	-0,0559
	50	0,9605	0,8488	-0,1017
5000	5	0,9817	0,9480	-0,0337
	10	0,9801	0,9417	-0,0384
	15	0,9784	0,9348	-0,0436
	25	0,9748	0,9189	-0,0559
	50	0,9618	0,8504	-0,1114

Convergence numérique

L'évaluation des propriétés du modèle a été effectuée après avoir éliminé les simulations présentant des problèmes de convergence de l'estimation. Les problèmes de convergence sont principalement dus aux valeurs initiales des paramètres utilisés dans la procédure d'estimation numérique. De telles situations sont assez rares : 1/120 (0,8%) pour un n=100 dans le cas d'une *GS* et 9/120 fois (7,5%) pour un n=100 dans le cas d'une *FS*. Les autres paramètres n'ont pas été affectés.

2.3 Conclusions sur les simulations

Afin de donner plus de recul aux cliniciens quant à l'utilisation des modèles conjoints à classes latentes, les résultats des simulations nous ont permis de mettre en évidence les propriétés suivantes.

— En général, les propriétés MLE des paramètres du modèle sont influencées par le nombre d'individus ainsi que par le nombre d'évènements observés et par le nombre d'observations longitudinales. Les deux derniers étant régis par le taux de censure.

— La fréquence des observations des marqueurs longitudinaux détermine également le nombre de mesures observé, bien que ce paramètre soit laissé fixe dans la présente étude.

— Les écarts par rapport à la **normalité** sont particulièrement présents pour les paramètres du sous-modèle de survie, et ces écarts disparaissent pour un nombre assez grand d'événements observés (petit taux de censure) et/ou une taille d'échantillon assez grande (à partir de 500 individus la normalité est généralement respectée même pour une forte censure).
— Pour le **biais relatif**, les tendances sont plus complexes. Les paramètres du sous-modèle de survie sont également plus impactés, surtout pour une petite taille d'échantillon n. Le grand nombre d'individus ne permet pas de compenser une forte censure, comme c'était le cas pour la normalité. Il n'y a pas de tendance particulière en fonction de n, sauf pour les paramètres du sous-modèle de survie, dont le biais est considérablement augmenté pour un petit n (n=100). Le biais diminue de façon quasi-linéaire pour presque tous les paramètres avec l'augmentation du nombre d'événements observés (taux de censure décroissant). Les estimations dans le cas de *FS* sont moins robustes à la taille de l'échantillon et au taux de censure que dans le cas de *GS*.
— Le **taux de recouvrement** est bon pour le cas de *GS*. Il n'est pas influencé par le nombre d'individus mais il diminue légèrement lorsque le taux de censure augmente. Pour le cas de *FS*, le taux de recouvrement diminue dans tous les cas et l'impact du taux de censure est encore plus important.
— L' **exactitude de l'identification des classes** est légèrement supérieure dans le cas de *GS* et d'une censure plus faible, mais n'est pas considérablement influencée par le nombre d'individus, sauf dans le cas d'une forte censure ; dans le cas du *FS*, l'exactitude de l'identification des classes est assez faible.

D'après ces propriétés, les recommandations suivantes ont pu être mises en évidence pour l'utilisation de ces modèles par les cliniciens :

— en ce qui concerne la mise en œuvre, lorsque nous sommes face à une **petite différence entre les classes dans les pentes du sous-modèle longitudinal** constatée *a posteriori* (*FS*), la procédure d'optimisation de la vraisemblance est plus susceptible de converger vers un maximum local. Ainsi, **plusieurs estimations avec différentes valeurs initiales des paramètres doivent être effectuées** pour s'assurer que l'estimation obtenue est le maximum global,
— une vigilance particulière doit être donnée lorsque la taille de l'échantillon en termes de nombre d'individus est **petite**. Cela entraîne des **déviations de la normalité**, en particulier pour les paramètres du sous-modèle de survie. Les intervalles de confiance fournis peuvent ne pas être valides. Comme représenté sur la Figure 2.4, le nombre d'individus et le taux de censure impactent les estimations des paramètres du modèle de survie ;
— un **fort taux de censure** implique :
  — un **biais dans l'estimation des paramètres** même lorsque le nombre d'individus est important ;
  — une **mauvaise précision d'identification des classes latentes** ;
— en cas de **mauvaise séparation** (*FS*) entre les classes latentes, le biais augmente et la précision des prédictions de classe diminue, les résultats doivent être interprétés avec prudence ;

— les paramètres du modèle sont généralement plus sensibles au taux de censure qu'au nombre d'individus en termes de biais, ainsi, **augmenter le temps d'observation est plus bénéfique pour la précision des estimations que d'augmenter la taille de l'échantillon en termes de nombre d'individus** ;
— les petits groupes latents avec peu d'événements (fort taux de censure) doivent être caractérisés avec prudence, car les estimations des paramètres peuvent être considérablement biaisées.

3 Résumé

Ce chapitre a permis de mettre en évidence des lacunes concernant l'utilisation du modèle conjoint à classes latentes, telles qu'un manque d'information sur l'utilisation des covariables dans les sous-modèles et un manque d'information sur la taille d'échantillon (nombre de patients, nombre de mesures répétées ou nombre d'évènements optimal). A l'aide d'une recherche bibliographique sur les articles cliniques et une étude de simulations de Monte-Carlo, nous avons pu conclure à différentes recommandations pour l'utilisation de ces modèles.

— **La place des covariables dans les sous-modèles dépend des objectifs de l'étude** (Chapitre 2, Section 1).
— Dans la majorité des cas, il est préférable d'utiliser un modèle sans covariable si l'on n'a aucun a priori sur les classes. Il est tout à fait possible par la suite de comparer les classes selon les caractéristiques des individus à l'aide d'une régression logistique multinomiale.
— **Le nombre maximum de covariable** à utiliser n'était pas dans les objectifs de ce travail. Cependant nous conseillons d'utiliser **les règles de Pedduzzi et Concato** (53) pour les sous-modèles logistique multinomial et de survie :
  — pour le sous-modèle logistique multinomial, nous conseillons de vérifier que le nombre de covariable inclus dans ce modèle ne dépasse pas 1 covariable pour 10 patients dans la plus petite classe. Par exemple, si la plus petite classe comprends 50 patients il faut s'assurer qu'il n'y a pas plus de 5 covariables incluses dans ce sous-modèle (50/10) ;
  — pour le sous-modèle de survie, deux cas de figures sont possibles :
    1. pour les covariables communes aux classes latentes, nous conseillons de vérifier que le nombre de covariables inclus dans ce modèle ne dépasse pas 1 covariable pour 10 évènements au total. Les paramètres du modèle de survie doivent également compter comme des covariables, c'est-à-dire que si l'on choisit une distribution de Weibull, les deux paramètres de forme et d'échelle devront être comptabilisés dans le nombre de covariable (54). Par exemple si le jeu de données comporte 60 évènements, il faut veiller à ne pas inclure plus de 6 covariables dans le sous modèle ;
    2. pour les covariables spécifiques aux classes latentes, nous conseillons de vérifier que le nombre de covariables inclues dans ce modèle ne dépasse pas 1 covariable pour 10 évènement dans la plus petite classe. Par exemple, si la plus petite classe comprend 20 évènements il faut

s'assurer qu'il n'y a pas plus de 2 covariables incluses dans ce sous-modèle (20/10).

- Pour le sous-modèle linéaire mixte, aucune règle n'a pu être trouvée mais les recommandations peuvent être moins strictes. En effet, en pratique les données disponibles sont souvent de grandes ampleurs puisque il y a plusieurs mesures par patients.

— L'impact de la taille de l'échantillon sur la normalité des paramètres, le biais relatif, le taux de couverture et le taux de bon classement a été résumé dans le Chapitre 2, Section 2.3. Ces recherches ont permis de mettre en évidence des solutions pour limiter les biais dues au manque de sujets, de mesures ou d'évènements et d'indiquer les cas de figures où les interprétations doivent être faites avec vigilance.

Chapitre 3

Modèle conjoint à classes latentes pour la stratification de patients : application dans l'étude TROPHOS

Dans ce chapitre, nous abordons l'application du modèle conjoint à classes latentes à l'analyse de l'évolution de la *sclérose latérale amyotrophique* (SLA) dans le cadre d'une étude TROPHOS. Il sera décomposé comme suit.

- Définition de la pathologie *sclérose latérale amyotrophique* : son incidence, délai de survie, causes, traitements.
- Description de l'étude TROPHOS : type d'étude, objectifs, résultats.
- Description des patients inclus dans l'étude TROPHOS : caractéristiques initiales, délai de survie et évolution du handicap.
- Application des modèles conjoints à classes latentes pour répondre à l'objectif principal du clinicien à l'origine de cette thèse : identifier des caractéristiques initiales des patients associées aux différents profils d'évolution de la SLA en prenant en compte le risque d'évènement composite (décès ou mise sous ventilation non invasive ou trachéotomie) corrélé à l'évolution du handicap (via l'échelle d'évaluation fonctionnelle SLA révisée).

1 Contexte clinique

1.1 *Sclérose latérale amyotrophique*

La *Sclérose Latérale Amyotrophique* (SLA) ou maladie de Charcot, est la plus fréquente des maladies rares du neurone moteur. En France, sa prévalence est de 4 à 6 malades pour 100 000 personnes et son incidence est de 2,5 nouveaux cas pour 100 000 habitants par an (55). La SLA constitue une dégénérescence progressive des motoneurones présents dans le cerveau, dans la moelle épinière et dans le bulbe rachidien. Les motoneurones sont responsables du conduit de l'information du cerveau jusqu'aux muscles, et leur mort se traduit par une atrophie musculaire des bras, des jambes, du visage ainsi que des muscles respiratoires. Après le diagnostic, le décès survient 3-4 ans après le début des symptômes, le plus souvent par insuffisance respiratoire. Dix pourcent des patients peuvent néanmoins vivre plus de 10 ans.

Les causes de cette dégénérescence des motoneurones ne sont pas connues. Plu-

sieurs hypothèses ont été faites mais aucune d'entre elles n'a été vérifiée puisque la SLA regroupe plusieurs types de maladies. En effet, la SLA peut se présenter sous deux formes différentes : la forme bulbaire et la forme spinale. La forme bulbaire se traduit par le début de la dégénérescence des motoneurones au niveau du bulbe rachidien, ce qui conduit dans un premier temps à une paralysie des muscles de la bouche. Elle touche principalement les femmes après 60 ans. La forme spinale elle, se traduit par la dégénérescence des motoneurones situés dans la moelle épinière ce qui conduit à une paralysie des membres inférieurs et supérieurs. Cette forme est la plus répandue, elle représente les deux tiers des cas de SLA et affecte d'avantage les hommes après 55 ans. En général, cette maladie intervient de manière isolée mais dans 5 à 10% des cas elle se transmet génétiquement. Cette transmission est caractérisée par une apparition de la maladie à un âge plus jeune (avant 50 ans) et l'évolution est souvent plus rapide que les SLA isolées. Un gène responsable de cette transmission de la maladie dans 20% des cas a été mis en évidence, il s'agit du gène SOD1 qui présente chez ces patients des anomalies mutationnelles. Pour les autres formes génétiques, aucun gène responsable n'a encore été trouvé.

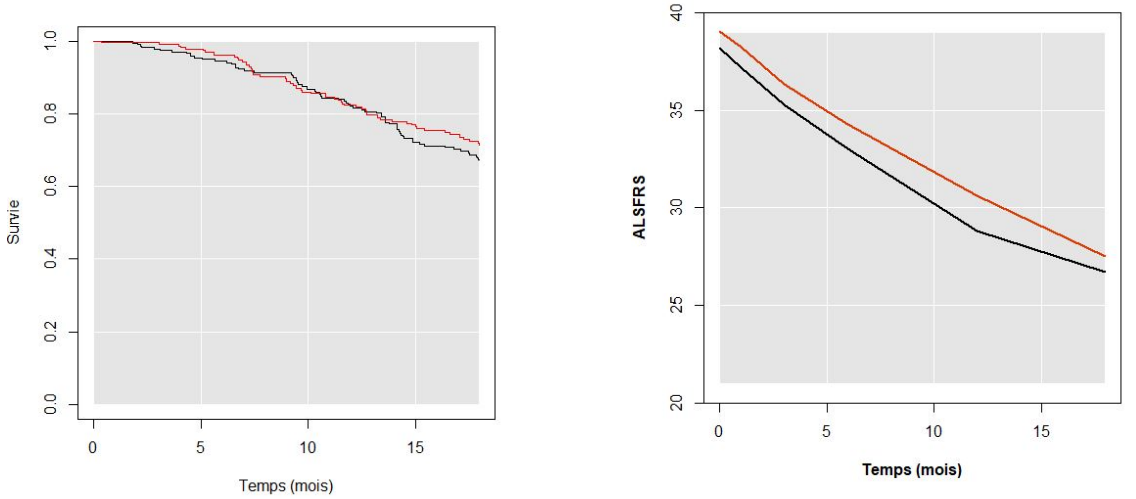
Face à cette maladie fatale et rapide, les recherches s'activent afin de trouver un traitement capable de ralentir ou au mieux d'empêcher la dégénérescence des motoneurones chez les personnes atteintes ou potentiellement atteintes de SLA. L'objectif actuel est de repousser au maximum l'évolution du handicap et ainsi d'augmenter l'espérance de vie en « bonne » santé. Un seul médicament, le *riluzole*, est à ce jour approuvé pour les patients atteints de SLA. Il permet de diminuer de manière très faible l'évolution du handicap mais augmente l'espérance de vie chez certaines personnes atteintes de SLA. Il consiste à diminuer le taux de glutamate qui est un messager nerveux et qui serait, suivant une hypothèse, trop important chez les personnes atteintes de SLA et responsable de la dégénérescence des motoneurones.

1.2 Étude TROPHOS

L'étude TROPHOS est un essai clinique ayant pour but de tester l'efficacité et la tolérance de l'*olesoxime*, une molécule aux propriétés neuroprotectrices, chez les patients atteints de SLA et traités avec le traitement de référence : le *riluzole*. C'est une étude en double aveugle, randomisée, et multicentrique. Entre mai 2009 et mars 2010, 512 patients avec une SLA définie ou probable et une capacité vitale lente d'au moins 70% ont été inclus dans l'étude et suivis pendant 18 mois.

L'évaluation des patients se réalisait via différents critères.

1. **La survie globale**. La durée de survie était déterminée comme la durée entre l'inclusion dans l'étude et le décès lié à la SLA ou la durée jusqu'à la date où le patient a été vu vivant la dernière fois.
2. **La survie sans évènement composite** : mise sous ventilation non invasive (VNI) ou trachéotomie ou décès lié à la SLA, qui sont des évènements considérés comme des marqueurs d'aggravation importants. La durée de survie était déterminée comme la durée entre l'inclusion dans l'étude et l'évènement composite ou la durée jusqu'à la date où le patient a été vu vivant la dernière fois.
3. **L'évolution du handicap** via le score d'ALSFRS (56). C'est une échelle d'évaluation reconnue sur le plan international de l'activité quotidienne des



(a) Courbes de la survie globale

(b) Courbes d'évolution du score ALSFRS

FIGURE 3.1 – Étude TROPHOS : Courbes de survie globales et courbes d'évolution du score ALSFRS moyen au cours du temps selon les groupes de traitement *Olesoxime* (en rouge) ou Placebo (en noir).

patients atteints de SLA et qui retranscrit de manière fiable l'état moteur et respiratoire des patients. Cette échelle est composée de 4 items évaluant l'état moteur bulbaire (salivation, parole, déglutition, respiration), 2 items pour les membres supérieurs (écriture, couper la nourriture) et 2 items pour les membres inférieurs (marche, escaliers), et enfin 2 items pour évaluer l'état de dépendance du patient (habillage et hygiène, se tourner dans le lit). Il est noté de 0 à 48 où 48 représente une autonomie totale du patient.

4. **L'évolution de la capacité vitale lente (CVL)** est le volume maximal d'air qui peut être lentement exhalé des poumons après une inspiration maximale. Elle est caractéristique d'une aggravation de la capacité respiratoire du patient et donc du handicap.
5. **L'évolution de la capacité musculaire** via le MMT qui est un examen clinique standardisé communément utilisé pour mesurer la force des groupes de périphériques du muscle squelettique.

Les patients étaient suivis de la manière suivante. Les visites étaient réalisées aux mois suivants : 0, 1, 2, 3, 6, 9, 12, 15 et 18. Ainsi la date exacte de survenue de l'évènement n'est pas connue. L'ALSFRS, le CVL et le MMT ont été récoltés aux visites suivantes : 0, 1, 3, 6, 12 et 18 mois.

Les résultats de cette étude ont montré que les patients traités par *olesoxime* en plus du *riluzole* n'avaient pas une meilleure survie globale (Figure 3.1a) ou sans mise sous ventilation non invasive ou trachéotomie.

Le risque de décès était augmenté avec les caractéristiques initiales des patients suivants : la forme bulbaire de la maladie *vs* la forme spinale, une augmentation de l'âge, une augmentation de la durée de la maladie et une diminution de la capacité musculaire et la capacité respiratoire.

L'*olesoxime* n'avait pas d'impact sur l'évolution de l'ALSFRS, qui diminuait d'environ 12 points sur les 18 mois de suivis dans les deux groupes de traitement

(Figure 3.1b).

Le traitement n'avait également pas d'impact sur l'évolution de la MMT ou la CVL. Aucune différence concernant les effets indésirables et la tolérance des médicaments n'a été retrouvée dans cette étude.
La conclusion de cet essai clinique était que l'*olesoxime* était bien tolérée mais ne montrait aucun bénéfice sur les patients atteints de SLA et traités par *riluzole*.

1.3 Descriptif des patients TROPHOS

Le Tableau 3.1 représente les caractéristiques d'inclusion des 512 patients tous les groupes confondus. Parmi les 512 patients diagnostiqués comme atteints de *Sclérose Latérale Amyotrophique*, 107 (20,9%) étaient diagnostiqués de manière certaine et 404 (79,1%) de manière probable. L'étude TROPHOS comportait 331(64,6%) hommes et 181(35,4%) femmes, 101(19,8%) personnes étaient atteintes de la forme bulbaire de la maladie et 410(80,2%) étaient atteintes par la forme spinale. Les patients avaient en moyenne 56 (±11,2) ans à l'inclusion et ils avaient en moyenne 55 (±11,2) ans aux premiers symptômes.

TABLEAU 3.1 – Étude TROPHOS : descriptif des patients SLA, n=512. Les valeurs sont exprimées en effectifs(%) ou moyenne±SD selon le type de variable.

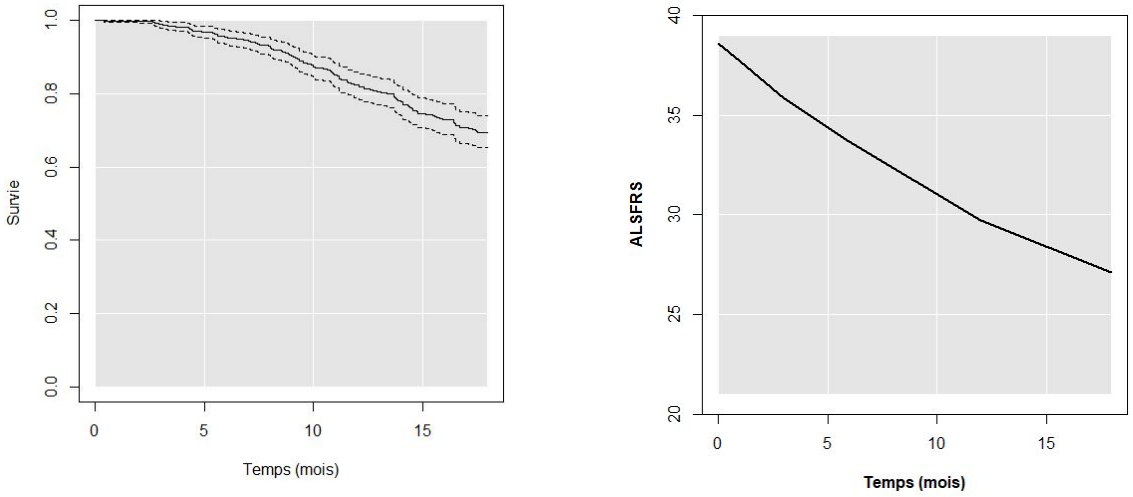
Variable	n	Valeurs
Traitement		
Placebo	512	252 (49,3)
Olesoxime	512	259 (50,7)
Age	511	56,6±11, 2
Hommes	511	330 (64,6)
Age au début des signes	511	55,2±11, 2
Délai depuis le diagnostic, mois	510	7,2±6, 0
Délai depuis les 1ers signes, mois	510	16,41±8, 0
Forme de la maladie		
Bulbaire	511	101 (19,8)
Spinale	511	410 (80,2)
Diagnostic de sclérose		
Probable	511	286 (56,0)
Probable avec signes	511	118 (23,1)
Définie	511	107 (20,9)
ALSFRS	510	38,6±5, 0
IMC	508	24,7±3, 6
Capacité musculaire	512	127±18, 4
Capacité vitale lente	512	3,6±1, 0
Pression artérielle systolique	510	131±16, 4
Pression artérielle diastolique	510	81,3±11, 6
Cholestérol total	510	5,9±1, 1

Le Tableau 3.2 reprends le nombre de patients à chaque visite où l'ALSFRS a été mesuré. Le nombre moyen de visite par patient était de 5.1 ± 1.2.

L'ALSFRS diminue pratiquement de manière constante au cours du temps avec une pente de -0,96 par mois chez les 512 patients TROPHOS (Figure 3.2b).

TABLEAU 3.2 – Étude TROPHOS : tableau descriptif du nombre de patients par visite. Les valeurs sont exprimées en effectifs (%).

Délai depuis l'inclusion (mois)	$n(\%)$
0	512 (100)
1	510 (99,6)
3	493 (96,3)
6	458 (89,5)
12	358 (69,9)
18	283 (55,3)



(a) Courbes de survie sans évènement composite

(b) Courbes d'évolution moyenne de l'ALSFRS

FIGURE 3.2 – Étude TROPHOS : Courbe de survie sns évènement composite et courbe d'évolution du score ALSFRS moyen au cours du temps quel que soit le groupe de traitement (N=512).

Au cours du suivi 132 patients ont eu l'évènement composite. La médiane de survie n'a pas été atteinte mais à la fin de l'étude 74,2% des patients ont survécu sans avoir eu l'évènement spécifique (Figure 3.2a).

2 Application du modèle conjoint à classes latentes

Un modèle conjoint à classes latentes a été utilisé pour analyser les données des patients inclus dans l'étude TROPHOS afin de répondre à l'objectif initié par le clinicien à l'origine de cette thèse : **rechercher les caractéristique initiales des patients associées aux profils d'évolution de la SLA**. L'évolution de la SLA est mesurée par l'évolution du score de handicap ALSFRS et à la survenue de l'évènement composite. La problématique clinique réside dans le fait que l'évolution de la maladie est très hétérogène d'un patient à l'autre même si le décès survient généralement 3-4 ans après le début des symptômes. Cette hétérogénéité pourrait expliquer que les essais cliniques réalisés jusqu'à présent, dont TROPHOS, soient tous non concluant en termes d'efficacité du traitement. Il serait alors intéressant de pouvoir trouver des facteurs prédictifs de l'évolution de la maladie et de les utiliser comme critères d'inclusion dans les essais cliniques afin de sélectionner un groupe de patients atteints de SLA plus homogène et d'ainsi augmenter les chances de trouver un traitement efficace. Afin de répondre à cet objectif, nous allons dans cette section :

1. identifier des classes latentes de patients avec des évolutions différentes de la maladie en utilisant un modèle conjoint à classes latentes sans covariable ;
2. comparer les caractéristiques initiales des patients entre les classes trouvées en utilisant des analyses univariées ;
3. trouver les facteurs indépendamment associés aux classes latentes à l'aide d'un modèle de régression logistique multinomial multivarié avec une sélection pas à pas descendante ;
4. identifier des seuils pour les facteurs du modèle final afin de fournir des recommandations aux cliniciens quant aux critères d'inclusion des patients à prendre en compte lors des prochains essais cliniques.

2.1 Estimation des paramètres du modèle

Dans le modèle, le paramètre longitudinal sera l'évolution du score ALSFRS et l'évènement sera un évènement composite composé du décès, de la mise sous VNI et/ou de la trachéotomie. Nous avons choisi de n'inclure aucune variable explicative dans les sous-modèles pour la création des classes et de chercher ensuite les facteurs associés aux profils identifiés par le modèle à l'aide d'un modèle de régression logistique multinomial multivarié, avec une sélection pas à pas descendante en suivant les travaux antérieurs (30).

En suivant la spécification générale du modèle (Eq.(1.14) et Eq.(1.39)), les composantes suivantes y sont incluses :

— Un **modèle logistique multinomial** décrivant la probabilité d'appartenance aux classes. Aucune variable explicative n'est inclue dans ce sous-modèle.
— Un **modèle linéaire mixte** décrivant l'évolution de l'ALSFRS, à intercept aléatoire avec une fonction quadratique du temps spécifique aux classes latentes. Les variances de l'effet aléatoire (σ_b^2) et de l'erreur de mesure (σ_ϵ^2) étaient considérées communes aux classes latentes.

— Un **modèle de survie** pour la survenue d'évènement composite est considéré spécifique aux classes latentes avec une distribution de Weibull du temps jusqu'à la survenue d'un évènement.

Dans les données originales les temps de survie étaient censurés par intervalles. En effet, comme décrit dans le Chapitre 3, Section 1.2, les évènements sont répertoriés uniquement aux dates de visites et les dates exactes d'évènement ne sont pas connues. L'estimation des paramètres du modèle conjoint à classes latentes en présence de censures par intervalle est implémentée dans le package *lcmm*. Néanmoins, afin de nous rapprocher de la configuration adoptée par les simulations effectuées (Chapitre 2), une imputation des temps exacts a été effectuée de la manière suivante (voir Figure 3.3 pour la représentation schématique) :

1. estimation des paramètres de la distribution de Weibull en présence de données censurées par intervalles sur l'ensemble des temps d'évènements. Pour les patients avec évènement le dernier temps de visite sans évènement a été considéré comme borne inférieure de l'intervalle et le temps de visite où l'évènement a été repéré comme borne supérieure. Pour les patients sans évènement, la borne inférieure de l'intervalle correspond au dernier temps de visite du patient et la borne supérieure à un temps infini ;
2. imputation des temps d'évènements pour les patients avec évènement. Pour cela, un temps d'évènement issu de la distribution de Weibull estimée dans l'étape 1), tronquée par sa borne inférieure et sa borne supérieure a été imputé. Si le patient n'avait pas subi d'évènement le temps considéré correspondait à son dernier temps de visite.

Le modèle conjoint à classes latentes appliqué aux données TROPHOS est le suivant :

$$\begin{cases} \pi_{ig} = \frac{e^{\xi_{0g}}}{\sum_{l=1}^{G} e^{\xi_{0l}}} \text{ voir Eq.(1.37).} \\ \\ Y_{ij}|(c_i = g) = \beta_{0g} + \beta_{1g} t_{ij} + \beta_{2g} t_{ij}^2 + b_{0i} + b_{1i} t_{ij} + b_{2i} t_{ij}^2 + \epsilon_i, \\ \qquad b_i \sim \mathcal{N}(0, B), \epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2) \text{ voir Eq.(1.14).} \\ \\ S(t_i)|(c_i = g) = \exp\left(- \left(\frac{t_i}{\zeta_{1g}} \right)^{\zeta_{2g}} \right), \\ \qquad T^\star \sim \mathcal{W}eibull(\zeta_{1g}, \zeta_{2g}), \\ \quad \lambda_i(t|c_i = g) = \zeta_{1g}^{\zeta_{2g}} \zeta_{2g} t^{\zeta_{2g}-1}, \text{ voir Eq.(1.39),} \end{cases} \tag{3.1}$$

avec B la matrice de covariance des effets aléatoires.
Comme décrit dans le Chapitre 1, Section 2.4, le choix du nombre de classes s'effectue à l'aide du BIC, du nombre de patients par classe et de la cohérence clinique. Nous avons donc fait varier le nombre de classes latentes de 1 à 4, regardé les différents BIC, le nombre de patients par classe et la cohérence clinique, et un modèle à 4 classes latentes a été retenu (Tableau 3.3).

2.2 Interprétation des résultats

Les courbes de survie estimées et les profils d'évolution d'ALSFRS prédits sont illustrés dans la Figure 3.4. Le Tableau 3.4 représente les résultats des estimations.

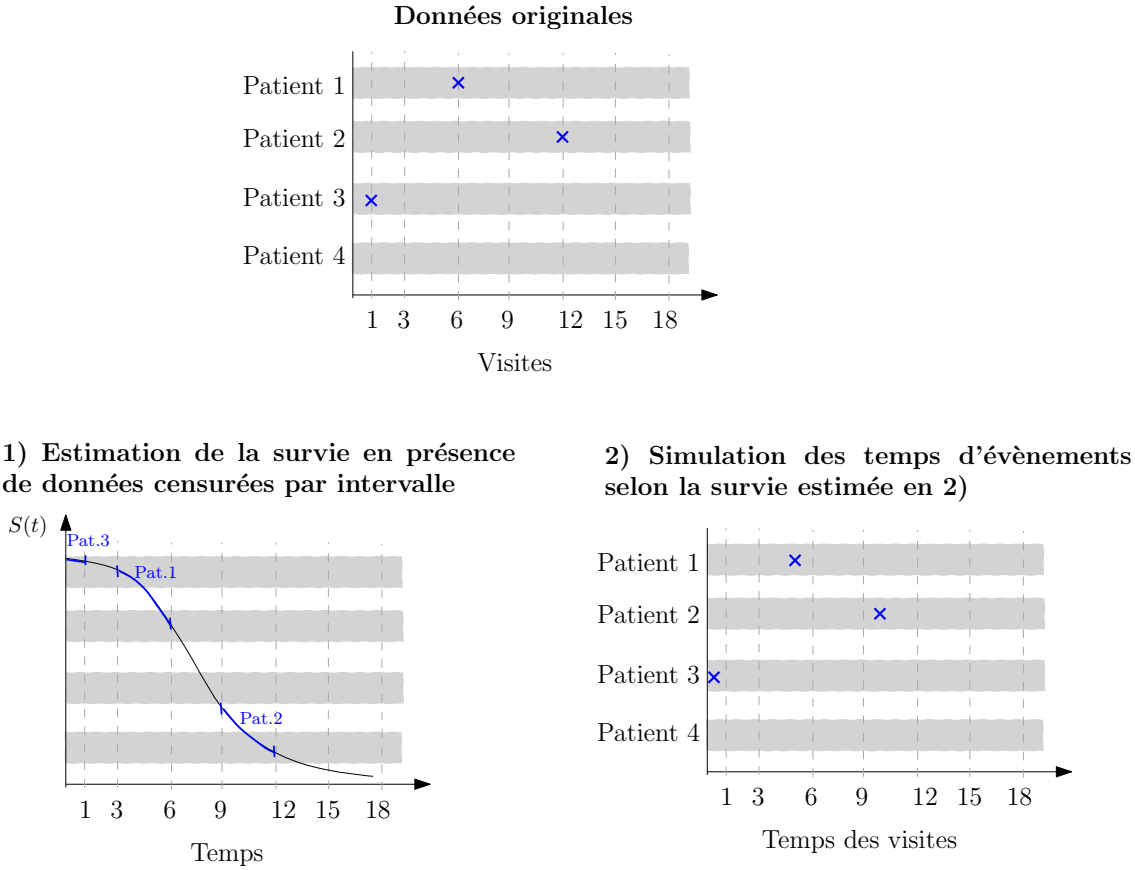
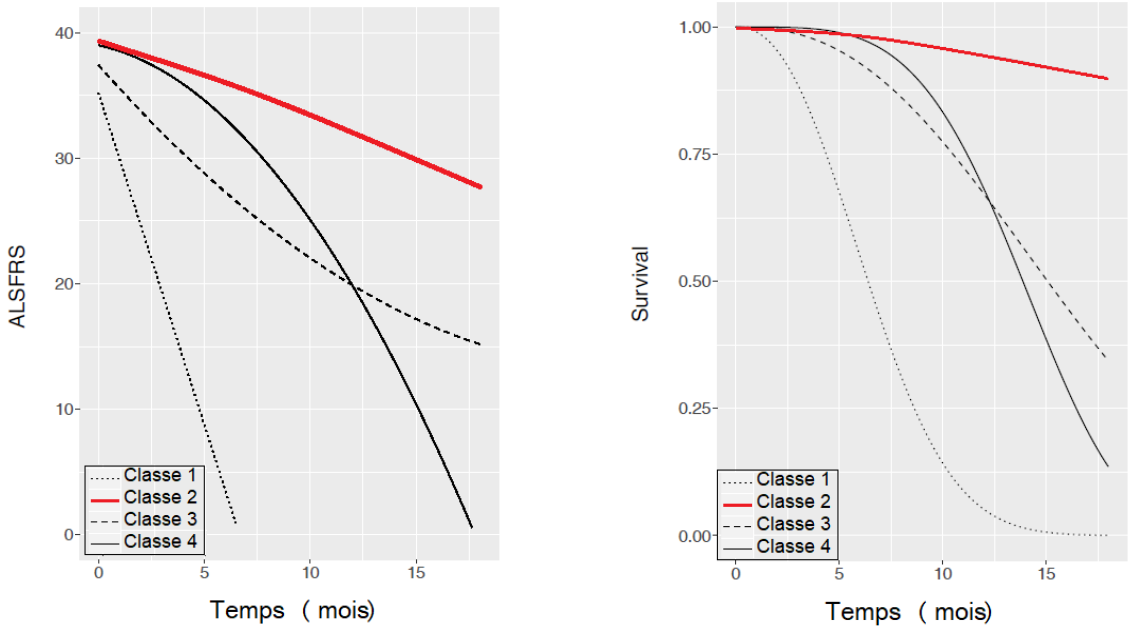


FIGURE 3.3 – Étude TROPHOS : schéma de l'imputation des temps initialement censurés par intervalles.

TABLEAU 3.3 – Étude TROPHOS : nombre de patients et BIC selon le nombre de classes latentes testé.

	Nombre de patients				BIC
	Classe 1	Classe 2	Classe 3	Classe 4	
Modèle à 1 classe	511	0	0	0	15110
Modèle à 2 classes	462	49	0	0	14941
Modèle à 3 classes	400	80	31	0	14911
Modèle à 4 classes	26	350	109	26	14901



(a) Évolution du score d'ALSFRS au cours du temps selon les 4 classes latentes

(b) Courbes de survie sans évènement composite selon les 4 classes latentes

FIGURE 3.4 – Étude TROPHOS : évolution de l'ALSFRS et courbes de survie sans évènement composite selon les 4 classes latentes retrouvées avec le modèle conjoint à classes latentes sans covariable.

L'interprétation des classes latentes est résumée dans le Tableau 3.5. Les classes 1 et 4 comportent chacune 5,1% de la population, soit 26 patients chacune. Elles représentent les patients avec le plus important déclin d'ALSFRS et un plus fort risque d'évènement composite. Les médianes de survie des classes 1 et 4 étaient de 7 et 14 mois respectivement. La classe 2 est la classe qui regroupe la majorité des patients (68,5%, n=350) et est caractérisée par la plus faible évolution de l'ALSFRS et le plus faible risque d'évènement composite. Enfin, la classe 3 est composée de 21,3% (n=109) de la population. Elle correspond aux patients qui ont une évolution du handicap quasi-similaire aux patients de la classe 2 mais avec un niveau d'ALSFRS de base plus faible (37 dans la classe 3 *vs.* 39 dans la classe 1 (Tableau 3.4)). La survie de la classe 3, quant à elle est bien plus faible que la classe 2 avec une médiane de survie autour de 15 mois.

En résumé, les profils identifiés sont :

— Des patients critiques (Classe 1) avec une **forte progression du handicap**

TABLEAU 3.4 – Étude TROPHOS : estimations des paramètres du modèle à classes latentes avec les erreurs standard et les p-values.

Nombre de mesures d'ALSFRS		2591	
Nombre de patients		511	
Nombre moyen de mesures d'ALSFRS par patient		5	
Nombre d'évènements composites		132	
Taux de censure		0,74	
Sous-modèles	**Paramètres**	**Estimateurs (se)**	**p-values**
Modèle logistique multinomial	ξ_{01}	-0,29 (0,49)	$0,55$
	ξ_{02}	2,26 (0,44)	$< 0,001$
	ξ_{03}	1,21 (0,53)	0,022
Modèle de survie (Weibull)	ζ_{11}	0,37 (0,02)	$<,001$
	ζ_{21}	1,52 (0,13)	$< 0,001$
	ζ_{12}	0,12 (0,02)	$< 0,001$
	ζ_{22}	1,25 (0,14)	$< 0,001$
	ζ_{13}	0,24 (0,01)	$< 0,001$
	ζ_{23}	1,56 (0,12)	$< 0,001$
	ζ_{14}	0,26 (0,01)	$< 0,001$
	ζ_{24}	2,02 (0,31)	$< 0,001$
Modèle linéaire mixte : effets fixes	β_{01}	35,22 (1,11)	$< 0,001$
	β_{11}	-5,29 (0,32)	$< 0,001$
	β_{21}	0,31 (0,04)	$< 0,001$
	β_{02}	39,37 (0,34)	$< 0,001$
	β_{12}	-0,52 (0,06)	$< 0,001$
	β_{22}	-0,01 (0,00)	0,007
	β_{03}	37,44 (0,66)	$< 0,001$
	β_{13}	-1,92 (0,16)	$< 0,001$
	β_{23}	0,04 (0,01)	$< 0,001$
	β_{04}	39,00 (1,30)	$< 0,001$
	β_{14}	-0,36 (0,20)	$< 0,001$
	β_{24}	-0,10 (0,02)	$< 0,001$
Modèle linéaire mixte : effets alatoires	$\sigma^2_{b_0}$	22,93	
	$\sigma^2_{b_1}$	0,20	
	$\sigma^2_{b_2}$	0,00	
	$\sigma^2_{\epsilon,1}$	1,67	

TABLEAU 3.5 – Étude TROPHOS : interprétation des 4 classes latentes trouvées à partir du modèle conjoint.

	Classe 1	Classe 2	Classe 3	Classe 4
Nombre de patients	26 (5,1%)	350 (68,5%)	109 (21,3%)	26 (5,1%)
ALSFRS à l'inclusion	Faible	Fort	Moyen	Fort
Déclin ALSFRS	Très fort	Très faible	Moyen	Fort
Risque d'évènement composite	Très fort	Très faible	Moyen	Moyen
Médiane de survie	7 mois	non atteinte	15 mois	14 mois

et une **faible survie**.

— Des patients dont **le handicap évolue rapidement** mais dont la **survie est moyenne** (Classe 4)

— Des patients avec une **faible évolution du handicap** et une **survie moyenne** (Classe 3).
— Des patients avec une évolution normale c'est à dire une **faible évolution du handicap** et une **forte survie** (Classe 2).

2.3 Recherche des facteurs associés aux classes

L'identification des profils et leur caractérisation *a posteriori* ne permet pas d'identifier les caractéristiques d'inclusion des patients associés à ces profils. Ainsi, nous réalisons une analyse permettant de rechercher des critères d'inclusion permettant de sélectionner la population regroupant la majorité des patients SLA et la plus homogène vis à vis de d'évolution de la maladie. Nous avons choisi de regrouper les classes 1, 3 et 4 afin de déterminer les facteurs associés à la classe 2 (classe avec la majorité des patients et une évolution "normale" du handicap et du risque de d'évènement composite).

Le Tableau 3.6 compare les caractéristiques d'inclusion entre la classe 2 et les 3 autres classes regroupées. Les analyses univariées ont été réalisées à l'aide d'un test du Chi-deux pour les paramètres qualitatifs, d'un test T de Student pour les paramètres continus gaussiens ou à l'aide d'un test U de Mann-Whitney pour les paramètres continus non-gaussiens. Les résultats de ces analyses montrent les tendances suivantes :

— **les patients de la classe 2 sont plus jeunes** au moment du début des signes que les patients des autres classes (54,2 (11,7) *vs.* 57,1 (9,9), p=0,004)) ;
— **les patients de la classe 2 ont une capacité musculaire et respiratoire** (capacité vitale lente) à l'inclusion **plus importantes** que les patients des autres classes (129 (18,1) *vs.* 122,9 (18,5) et 3,7 (1,0) *vs.* 3,3 (0,9) respectivement) ;
— la durée depuis de début les signes est également un facteur associé aux classes latentes. **Les patients de la classe 2 ont une durée depuis les signes plus importante** que les patients des autres classes (17,5 mois (8,3) *vs.* 14,0 (7,0), p<0,001).

TABLEAU 3.6 – Étude TROPHOS : comparaisons des caractéristiques initiales selon les classes latentes (2 *vs.* 1+3+4) : analyses univariées.

Variables	Classes 1+3+4 (n=161)	Classe 2 (n=350)	p-value
Age au début des signes	57,1±9, 9	54,2±11, 7	0,004
Sexe (Hommes)	99/161 (61,5)	231/349 (66,2)	0,30
Délai depuis les 1ers signes, mois	12,0 (8,0 à 18,0)	16,0 (11,0 à 22,0)	<0,001
Forme de la maladie			
Bulbaire	39/161 (24,2)	61/349 (17,5)	0,075
Spinale	122/161 (75,8)	288/349 (82,5)	
IMC	24,4±3, 3	24,9±3, 7	0,22
Capacité musculaire (MMT)	122,9 ±18, 5	129,0 ±18, 1	<0,001
Capacité vitale lente (CVL) (%)	88,3 ±12, 8	95,3 ±15, 4	<0,001
Pression artérielle systolique	132,3 ±17, 8	131,1 ±15, 8	0,47
Pression artérielle diastolique	81,7 ±11, 9	81,1 ±11, 6	0,55
Cholesterol total	6,0 ±1, 0	5,8 ±1, 2	0,23

A l'aide d'un modèle de régression logistique multivariée avec une sélection pas à pas descendante incluant tous les facteurs associés au seuil de 0,20 en univarié, nous avons obtenu le modèle final (Tableau 3.7). Une importante capacité musculaire et respiratoire ainsi qu'une plus grande durée depuis le début des signes étaient associées à une plus grande probabilité d'appartenir à la classe 2. Ce modèle final a un c-statistique (57) de 0,720 ce qui veut dire que la probabilité d'appartenir à la classe 2 est expliquée à 72,0% par ces variables.

TABLEAU 3.7 – Étude TROPHOS : caractéristique initiales des patients de TROPHOS associés à la classe 2. Le modèle final est obtenu avec un modèle logistique binomial avec une sélection pas à pas descendante. OR=odds ratios calculé avec son intervalle de confiance avec les classes 1+3+4 comme référence.

Variables	OR (95%CI)	p-value
Délai depuis les 1ers signes, mois	1,10 (1,06 to 1,12)	<0,001
Capacité musculaire	1,02 (1,01 to 1,04)	<0,001
Capacité vitale lente (%)	1,57 (1,27 to 1,92)	<0,001

2.4 Définition des seuils pour les variables associées aux classes

Après avoir trouvés les facteurs d'inclusion associés à l'appartenance de la classe 2 (classe avec une évolution "normale" de SLA) il est important de définir des valeurs seuils de la capacité musculaire et respiratoire ainsi que pour la durée depuis le début des signes permettant ainsi aux cliniciens d'ajouter ces critères d'inclusion à leurs études. Les valeurs seuils optimales selon les classes 2 *vs.* 1+3+4 ont été trouvées en utilisant la courbe ROC et la méthode de maximisation de l'index de Youden (58). Les valeurs seuils optimales retrouvées étaient de 129 pour la capacité musculaire (MMT), 96% pour la capacité vitale lente (CVL), et de 17 mois pour la durée depuis le début des signes.

D'après cette analyse, nous conseillons donc de sélectionner des patients SLA avec une capacité musculaire d'au moins 129, une capacité vitale lente d'au moins 96% (et non pas 70% comme dans l'étude TROPHOS) et un délai depuis le début des signes d'au moins 17 mois en plus des critères de sélection déjà utilisés : patients sous *riluzole* avec une SLA définie ou probable afin d'obtenir une population plus homogène vis-à-vis de l'évolution de la maladie.

3 Résumé

La *sclérose latérale amyotrophique* est la maladie rare la plus fréquente qui se caractérise par une dégénérescence des motoneurones conduisant à un handicap moteur progressif et in fine au décès. A ce jour aucun traitement efficace ne permet de guérir de cette maladie.
L'évolution du handicap et le risque de survenue du décès chez les patients sont très hétérogènes. Cette hétérogénéité a été supposée comme l'une des principales raisons des décennies d'échecs thérapeutiques.

Un modèle conjoint à classes latentes a été utilisé dans le cadre de l'étude TROPHOS afin d'identifier 4 différents profils de patients avec des évolutions homogènes

de la maladie à la fois vis-à-vis de l'évolution du handicap (score ALSFRS) et de la survenue d'un évènement composite (décès ou mise sous ventilation non-invasive ou trachéotomie).
L'identification de ces classes nous a permis de déterminer les caractéristiques initiales des patients associées à ces différents profils. Ces facteurs, indépendamment associés aux classes (délai depuis les premiers signes, capacité musculaire et capacité vitale lente), ont été seuillés et caractérisés comme de nouveaux critères d'inclusion à prendre en compte dans les prochains essais cliniques, permettant de recruter une population plus homogène. Ainsi, il sera possible d'établir une recherche de traitement en s'affranchissant des patients avec des évolutions extrêmes.

Ces résultats permettront d'optimiser *in fine* la prise en charge des patients.

Chapitre 4

Recherche de facteurs associés aux *outcomes* dans l'étude TROPHOS : comparaisons des modèles marginaux et des modèles conjoints

Afin de tester l'impact de covariables sur l'évolution du handicap, mesurée par le score ALSFRS, et sur le risque de survenue de l'évènement composite, plusieurs modèles statistiques peuvent être utilisés tels que :

1. les modèles marginaux :
   - un **modèle linéaire mixte** pour étudier les facteurs associés à l'évolution du handicap ;
   - un **modèle de survie** pour étudier les facteurs associés au risque de survenue de l'évènement composite ;
2. les modèles conjoints :
   - un **modèle conjoint à classes latentes** pour étudier les facteurs associés aux deux critères lorsque l'on suppose qu'il existe des classes latentes d'évolution de la maladie ;
   - un **modèle à effets aléatoires partagés** pour étudier les facteurs associés aux deux critères lorsque l'on suppose que la population est homogène vis à vis de l'évolution de la maladie.

Les avantages et inconvénients de chaque méthode seront détaillés afin de tester l'apport des modèles conjoints à classes latentes comparé aux autres modèles existants.

1 Les modèles marginaux

1.1 Modèle linéaire mixte pour l'évolution du handicap

Les modèles linéaires mixtes sont les modèles les plus couramment utilisés lorsque l'on souhaite étudier les facteurs associés à l'évolution d'un paramètre longitudinal comme un biomarqueur. Comme détaillé dans le Chapitre 1, Section 1.1, les modèles linéaires mixtes sont une extension du modèle linéaire dans lequel des effets aléatoires

sont ajoutés afin de tenir compte des potentielles corrélations au sein des mesures répétées d'un même individu. En d'autres termes, ce modèle tient compte du fait que les mesures d'un même patient sont plus corrélées entre elles que deux mesures de deux patients différents.
Nous allons appliquer ce modèle aux données TROPHOS afin de déterminer les facteurs associés à l'évolution du handicap (ALSFRS). Les covariables à tester sont des facteurs sélectionnés par le clinicien comme la durée depuis le début des signes, l'IMC, la capacité musculaire (MMT), la capacité vitale lente (CVL) et le volume corpusculaire moyen (MCV). Le modèle linéaire mixte estimé a la forme suivante :

$$\begin{aligned} ALSFRS_{ij} = \; & \beta_0 + \beta_1 t_{ij} + \beta_2 SO_i + \beta_3 IMC_i + \\ & \beta_4 MMT_i + \beta_5 CVL_i + \beta_6 MCV_i + \\ & t_{ij} \times (\beta_7 SO_i + \beta_8 MMT_i + \beta_9 CVL_i) + \\ & b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}, \\ & b_i \sim \mathcal{N}(0, \sigma_b^2), \\ & \epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2), \text{ voir Eq.(1.14).} \end{aligned} \tag{4.1}$$

Les matrices de l'Eq.(4.1) associées à ce modèle sont explicitées dans l'annexe C. Les résultats des estimations de ce modèles sont représentés dans le Tableau 4.1.

TABLEAU 4.1 – Étude TROPHOS : résultats d'estimations du modèle linéaire mixte des paramètres avec les erreurs standard et les p-values.

nombre d'observations	2525	
nombre de patients	497	
nombre moyen de mesures d'ALSFRS	5	
Paramètres	Estimateur (se)	p-values
β_0	8,94 (4,03)	0,027
β_1	-2,99 (0,26)	$< 0,001$
β_2 (SO)	-0,04 (0,02)	0,069
β_3 (IMC)	-0,13 (0,05)	0,012
β_4 (MMT)	0,16 (0,01)	$< 0,001$
β_5 (CVL)	1,09 (0,18)	$< 0,001$
β_6 (MCV)	0,09 (0,04)	0,012
β_7 (SO$\times t_j$)	0,03 (0,00)	$< 0,001$
β_8 (MMT$\times t_j$)	0,01 (0,00)	$< 0,001$
β_9 (CVL$\times t_j$)	0,08 (0,03)	0,004
$\sigma^2_{b_0}$	3,83	
σ^2_{b1}	0,57	
$\sigma^2_{\epsilon,1}$	2,00	

Note : Les covariables suivantes et leurs interactions avec le temps pour certaines d'entre elles sont présentées : SO (délai depuis le début des symptômes), IMC (Indice de masse corporelle), MMT (capacité musculaire), CVL (capacité vitale lente), MCV (volume corpusculaire moyen)).

Les résultats de ce modèle montrent que le score ALSFRS diminue de 2,99 points par mois lorsque les autres paramètres sont à 0. Plus la capacité musculaire (MMT),

la capacité vitale lente (CVL) et le volume corpusculaire moyen (MCV) sont importants à l'inclusion, plus la valeur de l'ALSFRS à l'inclusion sera grande. Au contraire, plus l'IMC sera important à l'inclusion, plus le score ALSFRS sera faible à l'inclusion. La durée depuis les premiers symptômes n'est pas significativement associée au score ALSFRS à l'inclusion (p=0,069). Cependant, une importante durée depuis les premiers symptômes, une capacité musculaire et une capacité vitale lente plus élevées à l'inclusion sont des facteurs protecteurs et sont associées à un ralentissement du déclin du score ALSFRS et donc du handicap.

Ce modèle est simple d'utilisation mais comporte un biais important. En effet, les données manquantes de l'ALSFRS peuvent être dues à la survenue de l'évènement composite : la mort, la mise sous ventilation non invasive ou la trachéotomie peuvent mettre un terme aux mesures du score du handicap. Or, l'estimation des paramètres via le maximum de vraisemblance est basée sur l'hypothèse que les données manquantes sont aléatoires. La présence de données manquantes informatives peut provoquer des biais dans l'estimation des paramètres. Le modèle linéaire mixte n'est donc pas optimal pour ce type de données (voir Chapitre 1, Section 1.1 pour plus de détails).

1.2 Modèle de survie pour la survenue de l'évènement composite

Le modèle de survie est un modèle permettant d'étudier les facteurs associés au risque de survenue de l'évènement. Dans notre cas, nous souhaitons analyser l'impact de covariables sélectionnées par le clinicien : l'âge au début des signes (AO), la durée depuis le début des signes (SO), l'IMC, la capacité musculaire (MMT) et la capacité respiratoire (capacité vitale lente CVL) sur le risque d'évènement composite (décès, mise sous ventilation non invasive ou trachéotomie). Le modèle de survie que nous avons choisi est spécifié par une distribution de Weibull du temps jusqu'à la survenue d'un évènement. La fonction de hazard pour ce modèle est :

$$\begin{aligned}\lambda_i(t) = \; & \underbrace{\zeta_1^{\zeta_2}\zeta_2 t^{\zeta_2-1}}_{\lambda_0(t)} \exp(\gamma_1 SO_i + \gamma_2 IMC_i + \\ & \gamma_3 MMT_i + \gamma_4 CVL_i + \gamma_5 AO_i), \text{ voir Eq.(1.24).}\end{aligned} \tag{4.2}$$

Les résultats d'estimations de ce modèle de survie sont détaillés dans le Tableau 4.2.

Plus le délai depuis les premiers symptômes, l'IMC, la capacité musculaire et la capacité vitale lente sont importants à l'inclusion, plus le risque de survenue de l'évènement composite est faible. Á l'inverse, plus l'âge au début des signes est important, plus le risque de survenue de l'évènement composite est important. Il est potentiellement possible d'inclure la variable décrivant le score ALSFRS dans ce modèle en tant que variable dépendante du temps. Néanmoins, cela nécessiterait de connaître les valeurs du score pour chaque temps d'évènement (voir Chapitre 1, Section 4). Or, les mesures du score du handicap sont effectuées à des temps précis,

TABLEAU 4.2 – Étude TROPHOS : Résultats des estimations du modèle de survie des paramètres avec les erreurs standard et les p-values.

nombre d'observations	2525	
nombre de patients	497	
nombre d'évènements	129	
taux de censure	0.74	
Paramètres	Estimateur (se)	p-values
ζ_1	0,67	0,004
ζ_2	1,96	$< 0,001$
γ_1 (SO)	-0,07 (0,01)	$< 0,001$
γ_2 (IMC)	-0,07 (0,03)	0,005
γ_3 (MMT)	-0,03 (0,01)	$< 0,001$
γ_4 (CVL)	-0,35 (0,12)	0,002
γ_5 (AO)	0,03 (0,01)	$< 0,001$

Note : Les covariables suivantes et leurs interactions avec le temps pour certaines d'entre elles sont présentées : SO (délai depuis le début des symptômes), IMC (Indice de masse corporelle), MMT (capacité musculaire), CVL (capacité vitale lente), MCV (volume corpusculaire moyen), AO (âge au début des signes).

0, 1, 3, 6, 12 et 18 mois. Les évènements décès, mise sous trachéotomie ou mise sous ventilation non invasives peuvent également être répertoriés à 2, 9 et 15 mois. De plus, l'ALSFRS est une variable endogène (sa valeur est influencée par la survenue de l'évènement composite, voir Chapitre 1 Section 4).
L'utilisation d'un modèle de conjoint pour cet objectif est donc plus adapté.

2 Les modèles conjoints

2.1 Modèle conjoint à classes latentes pour l'évolution du handicap et pour le risque de survenue de l'évènement composite

Les modèles conjoints à classes latentes permettent d'étudier l'impact de covariables sur l'évolution du handicap (via le score ALSFRS) et sur le risque de survenue de l'évènement composite conjointement dans une population hétérogène. Les objectifs de l'analyse dans ce contexte sont les suivants.

- Trouver un nombre de classes latentes optimal expliquant la relation entre l'évolution de l'ALSFRS et le risque d'évènement composite.
- Interpréter les associations entre les covariables : l'âge au début des signes (AO), la durée depuis le début des signes (SO), l'IMC, la capacité musculaire (MMT) et la capacité respiratoire (capacité vitale lente CVL) et la survenue de l'évènement composite.
- Interpréter les associations entre les covariables (SO, IMC, MMT, CVL et MCV) et l'évolution de l'ALSFRS.

Le modèle général (Eq.(1.14) et Eq.(1.39)) s'écrit dans ce contexte de la manière suivante :

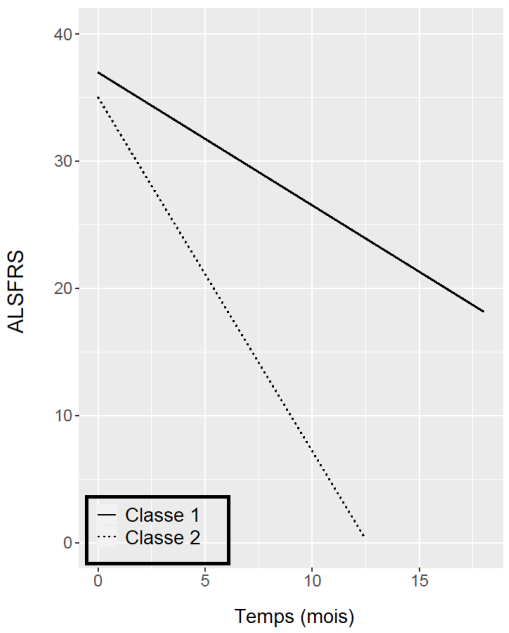
$$\begin{cases} \pi_{ig} & = \frac{e^{\xi_{0g}}}{\sum_{l=1}^{G} e^{\xi_{0l}}}, \text{ voir Eq.(1.37).} \\ Y_{ij}|(c_i = g) & = \beta_{0g} + \beta_{1g} t_{ij} + \beta_2 SO_i + \beta_3 IMC_i + \\ & \beta_4 MMT_i + \beta_5 CVL_i + \beta_6 MCV_i + \\ & t_{ij} \times (\beta_7 SO_i + \beta_8 MMT_i + \beta_9 CVL_i) + \\ & b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}, \\ & b_i \sim \mathcal{N}(0, \sigma_b^2), \\ & \epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2), \text{ voir Eq.(1.14).} \\ \lambda_i(t)|(c_i = g) & = \underbrace{\zeta_{1g}^{\zeta_{2g}} \zeta_{2g} t^{\zeta_{2g}-1}}_{\lambda_0(t)} \exp(\gamma_1 SO_i + \gamma_2 IMC_i + \\ & \gamma_3 MMT_i + \gamma_4 CVL_i + \\ & \gamma_5 AO_i), \text{ voir Eq.(1.39).} \end{cases} \tag{4.3}$$

Dans ce modèle les effets des covariables sont supposées communes aux classes latentes. C'est-à-dire que d'une classe à une autre, l'impact de la covariable sur l'évolution du handicap et sur le risque de survenue de l'évènement composite est supposé identique. Comme vu précédemment, il est également possible de tester un effet spécifique des covariables selon les classes latentes mais ce cas ne sera pas testé dans cette thèse puisqu'il n'y avait pas d'intérêt clinique.
Nous avons fait varier le nombre de classes latentes de 1 à 4 afin de déterminer le nombre de classes optimal. A l'aide d'un consensus entre le BIC, le nombre de patients par classe et la cohérence clinique, un modèle à 2 classes latentes a été retenu (Tableau 4.3).

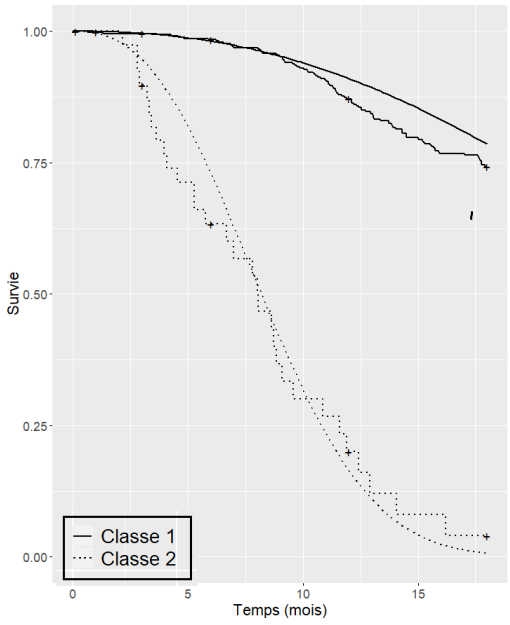
TABLEAU 4.3 – Étude TROPHOS : Nombre de patients et BIC selon le nombre de classes latentes testé.

	Nombre de patients				BIC
	Classe 1	Classe 2	Classe 3	Classe 4	
Modèle à 1 classe	497	0	0	0	14517
Modèle à 2 classes	452	45	0	0	14408
Modèle à 3 classes	370	107	20	0	14410
Modèle à 4 classes	333	124	26	14	14423

Les courbes de survie estimées et les profils d'évolution d'ALSFRS prédits sont illustrés dans la Figure 4.1. Le risque d'évènement est légèrement sous-estimé dans la classe 1 comparé à la courbe observée (Kaplan-Meier). Ce biais s'explique par le fait que la classe 1 est caractérisée par un mélange de distributions de Weibull des temps d'évènements (distribution bi-modale, voir Figure en AnnexeD.1) alors que nous n'avions spécifié qu'une seule distribution de Weibull par classe.
Les paramètres estimés du modèle sont fournis dans le Tableau 4.4. La classe 1 comprend la majorité des patients (90.9%) ; elle est caractérisée par une évolution plus modérée de l'ALSFRS que la classe 2 et par une meilleure survie que la classe 2 (médiane de survie à 20 mois comparé à 8 mois pour la classe 2 pour des valeurs



(a) Évolution du score d'ALSFRS au cours du temps selon les 2 classes latentes



(b) Courbes de survie observées (Kaplan-Meier) et estimée sans évènement composite selon les 2 classes latentes

FIGURE 4.1 – Étude TROPHOS : évolution de l'ALSFRS et courbes de survie sans évènement composites selon les 2 classes latentes retrouvées avec le modèle conjoint à classes latentes avec des valeurs moyennes des covariables.

moyennes de covariables). Ces analyses ont donc montré qu'un fort risque d'évènement est associé à un plus fort déclin de l'ALSFRS.

Une fois les 2 classes latentes trouvées, l'impact des covariables sur les deux outcomes (ALSFRS et évènement composite) a été étudié. Les paramètres estimés correspondant à l'Eq.(4.3) sont résumés dans le Tableau 4.4. Les résultats des **facteurs associés à l'ALFSRS** sont les suivants. En tenant compte du risque de décès et quel que soit la classe latente :

- plus la capacité musculaire (**MMT**), la capacité vitale lente (**CVL**) et le volume corpusculaire moyen (**MCV**) sont importants à l'inclusion, plus **la valeur de l'ALSFRS à l'inclusion sera grande**.
- Plus l'**IMC** et la durée depuis les premiers signes (**SO**) seront importants à l'inclusion, plus **la valeur de l'ALSFRS sera faible à l'inclusion**.
- Une importante durée depuis les premiers signes (**SO**), capacité musculaire (**MMT**) et capacité vitale lente (**CVL**) est associées à un **ralentissement du déclin du score ALSFRS et donc du handicap**.

Les résultats des **facteurs associés au risque de survenue de l'évènement composite** sont les suivants. En tenant compte de l'évolution du handicap (ALSFRS) et quel que soit la classe latente :

- plus le délai depuis les premiers symptômes (SO), la capacité musculaire (MMT) et la capacité vitale lente (CVL) sont importants à l'inclusion, plus **le risque de survenue de l'évènement composite** est faible.
- Plus l'âge au début des signes (AO) est important, plus **le risque de survenue de l'évènement composite** est important.

TABLEAU 4.4 – Étude TROPHOS : résultats des estimations des paramètres (avec les erreurs standard et les p-values) des 2 classes latentes obtenues avec le modèle conjoint à classes latentes.

	nombre d' observations	2525	
	nombre de patients	497	
	nombre moyen de mesures d'ALSFRS	5	
	nombre d'évènement	129	
	taux de censure	0,74	
Sous-modèles	Paramètres	Estimate (se)	p-values
Modèle logistique multinomial	ξ_{01}	2,22 (0,31)	$< 0,001$
Modèle de Weibull	ζ_{11}	0,68 (0,21)	0,001
	ζ_{21}	1,64 (0,12)	$< 0,001$
	ζ_{12}	0,48 (0,17)	0,004
	ζ_{22}	1,48 (0,08)	$< 0,001$
	γ_1 (SO)	-0,05 (0,01)	0,008
	γ_2 (IMC)	-0,05 (0,03)	0,079
	γ_3 (MMT)	-0,03 (0,01)	$< 0,001$
	γ_4 (CVL)	-0,41 (0,12)	$< 0,001$
	γ_5 (AO)	0,04 (0,01)	$< 0,001$
Linear mixed model : fixed effects	β_{01}	9,79 (4,02)	0,015
	β_{11}	-2,32 (0,27)	$< 0,001$
	β_{02}	7,83 (4,12)	0,057
	β_{12}	-4,06 (0,39)	$< 0,001$
	β_2 (SO)	-0,06 (0,02)	$< 0,001$
	β_3 (IMC)	-0,13 (0,05)	0,009
	β_4 (MMT)	0,16 (0,00)	$< 0,001$
	β_5 (CVL)	1,04 (0,18)	$< 0,001$
	β_6 (MCV)	0,10 (0,04)	0,007
	β_7 (SO$\times t_j$)	0,02 (0,00)	$< 0,001$
	β_8 (MMT$\times t_j$)	0,01 (0,00)	$< 0,001$
	β_9 (CVL$\times t_j$)	0,06 (0,02)	0,018
Linear mixed model : random effects	$\sigma^2_{b_0}$	14,10 (0,00)	
	σ^2_{b1}	0,18	
	$\sigma^2_{\epsilon,1}$	1,97	

Note : Les covariables suivantes et leurs interactions avec le temps (si significatives) sont présentées : SO (délai depuis le début des symptômes), IMC (Indice de masse corporelle), MMT (capacité musculaire), CVL (capacité vitale lente), MCV (volume corpusculaire moyen), AO (âge au début des signes).

— L'IMC n'est pas significativement associé au **risque de survenue de l'évènement composite**.

2.2 Modèle conjoint à effets aléatoires partagés pour l'évolution du handicap et pour le risque de survenue de l'évènement

Les modèles conjoints à effets aléatoires partagés peuvent être utilisés pour analyser l'impact des covariables sur le risque de survenue de l'évènement composite et

sur l'évolution du handicap. Tout comme les modèles conjoints à classes latentes, ils permettent de tenir compte de données manquantes informatives issues de la survenue de l'évènement composite dans l'analyse de l'évolution de l'ASLFRS mais également de tenir compte de l'évolution de l'ALSFRS dans l'analyse de la survenue de l'évènement composite. La différence qui réside entre les modèles à classes latentes et les modèles à effets aléatoires partagés est que ces derniers ne font pas d'hypothèse sur l'hétérogénéité de la population. Le modèle conjoint à effets aléatoires partagés appliqué aux données TROPHOS est le suivant.

$$\begin{cases} ALSFRS_{ij} = & \beta_0 + \beta_1 t_{ij} + \beta_2 SO_i + \beta_3 IMC_i + \\ & \beta_4 MMT_i + \beta_5 CVL_i + \beta_6 MCV_i + \\ & t_{ij} \times (\beta_7 SO_i + \beta_8 MMT_i + \beta_9 CVL_i) + \\ & b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}, \\ & b_i \sim \mathcal{N}(0, \sigma_b^2), \\ & \epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2). \\ \lambda_i(t) = & \underbrace{\zeta_1^{\zeta_2} \zeta_2 t^{\zeta_2 - 1}}_{\lambda_0(t)} \exp(\gamma_1 SO_i + \gamma_2 IMC_i + \\ & \gamma_3 MMT_i + \gamma_4 CVL_i + \gamma_5 AO_i + g_i(b_i, t)^\intercal \eta). \end{cases} \tag{4.4}$$

L'association entre le modèle linéaire et le modèle de survie est représentée par la fonction $g_i(b_i, t)$ qui peut être uni- ou multidimensionnelle et par le vecteur de paramètres η. Etant donné que l'utilisation de ce modèle dans notre cas consistait à étudier l'association entre les covariables et l'ALSFRS ou l'évènement composite, une fonction simple des effets aléatoires $g_i(b_i, t)$ a été choisie. Elle représentait la valeur de l'ALSFRS au temps courant c'est-à-dire que le risque instanné de l'évènement au temps t dépendait de la valeur de l'ALSFRS en t débarrassée de l'erreur de mesure. D'après les Eq.(1.31) et Eq.(1.32) le modèle de survie devient :

$$\begin{aligned} \lambda_i(t) = & \ \zeta_1^{\zeta_2} \zeta_2 t^{\zeta_2 - 1} \exp(\gamma_1 SO_i + \gamma_2 IMC_i + \\ & \gamma_3 MMT_i + \gamma_4 CVL_i + \gamma_5 AO_i + \\ & \eta \tilde{Y}_i(t)). \end{aligned} \tag{4.5}$$

L'implémentation de ce modèle dans le logiciel R a rencontré des problèmes de convergence. Il a donc été décidé de retirer l'interaction entre le temps et la variable capacité musculaire à l'inclusion ($\beta_8 MMT_i$) du modèle.
Les résultats d'estimations de ce modèle sont représentés dans le Tableau 4.5.

La valeur de l'estimation de l'association entre l'évolution de l'ALSFRS et la survenue de l'évènement composite était de -0,11 (0,01) et la p-value<0,0001, ce qui signifie que plus la valeur du score ALSFRS est élevée, plus le risque de survenue de l'événement est faible.
Dans ce modèle, l'augmentation du risque de survenue de l'évènement composite était également associée à un court délai depuis les premiers symptômes (SO), à un faible IMC ainsi qu'à un âge au début des symptômes (AO) important.
Plus l'IMC à l'inclusion était grand plus le score ALSFRS à l'inclusion était faible. A l'inverse, une forte capacité musculaire (MMT), capacité vitale lente (CVL) et un fort volume corpusculaire moyen (MCV) étaient associés à un important score d'ALSFRS à l'inclusion. De plus, plus la durée depuis le début des signes (SO) et la

TABLEAU 4.5 – Étude TROPHOS : résultats des estimations des paramètres (avec les erreurs standard et les p-values) du modèle conjoint à effets aléatoires partagés.

nombre d'observations	2525	
nombre de patients	497	
nombre moyen de mesures d'ALSFRS par patient	5	
nombre d'évènements	129	
taux de censure	0.74	
Paramètres	Estimate (se)	p-values
η	-0,11 (0,01)	$< 0,001$
ζ_1	0,78 (0,25)	$< 0,001$
ζ_2	1,54 (0,10)	$< 0,001$
γ_1 (SO)	-0,03 (0,01)	0,026
γ_2 (IMC)	-0,08 (0,03)	0,002
γ_3 (MMT)	-0,006 (0,01)	0,27
γ_4 (CVL)	-0,15 (0,11)	0,17
γ_5 (AO)	0,04 (0,01)	$< 0,001$
β_0	9,16 (4,06)	0,024
$\beta_1(t_j)$	-1,87 (0,13)	$< 0,001$
β_2 (SO)	-0,05 (0,02)	0,055
β_3 (IMC)	-0,13 (0,05)	0,012
β_4 (MMT)	0,16 (0,01)	$< 0,001$
β_5 (CVL)	1,09 (0,18)	$< 0,001$
β_6 (MCV)	0,10 (0,04)	0,010
β_7 (SO$\times t_j$)	0,03 (0,00)	$< 0,001$
β_9 (CVL$\times t_j$)	0,11 (0,03)	$< 0,001$

Note : Les covariables suivantes et leurs interactions avec le temps pour certaines d'entre elles sont présentées : SO (délai depuis le début des symptômes), IMC (Indice de masse corporelle), MMT (capacité musculaire), CVL (capacité vitale lente), MCV (volume corpusculaire moyen), AO (âge au début des signes).

capacité vitale lente (CVL) était important à l'inclusion moins le déclin d'ALSFRS au cours du temps était important.

3 Comparaisons des résultats des différents modèles

Quatre modèles ont été utilisés dans le but d'identifier les facteurs associés à l'évolution de l'ALSFRS et au risque de survenue de l'évènement composite dans le cadre de l'étude TROPHOS : modèles classiques (modèle linéaire mixte, modèle de survie) et modèles conjoints (modèle conjoint à classes latentes et modèle à effets aléatoires partagés). La comparaison des résultats fournis par ces quatre modèles sera effectuée par l'analyse des paramètres estimés (et donc les facteurs de risque identifiés) et par l'étude de leur adéquation.

3.1 Comparaison des paramètres estimés

Le Tableau 4.6 fourni les résultats des estimations des paramètres.

TABLEAU 4.6 – Étude TROPHOS : résultats des estimations du modèle linéaire mixte, du modèle de survie, du modèle conjoint à classes latentes et du modèle conjoint à effets aléatoires partagés. Estimations des paramètres avec les erreurs standard et les p-values.

Variables expliquées	Covariables	MLM*		MDS*		MCCL*		MCEAP*	
		Estimate (se)	p-values	Estimate (se)	p-values	Estimate (se)	p-values	Estimate (se)	p-values
Evènement composite	γ_1 (SO)			-0,07 (0,01)	$< 0,001$	-0,05(0,01)	0,008	-0,03 (0,01)	0,026
	γ_2(IMC)			-0,07 (0,03)	$< 0,001$	-0,05 (0,03)	0,079	-0,08 (0,03)	0,002
	γ_3(MMT)			-0,03 (0,01)	$< 0,001$	-0,03 (0,01)	$< 0,001$	-0,01 (0,01)	0,27
	γ_4 (CVL)			-0,35 (0,12)	0,002	-0,41 (0,12)	$< 0,001$	-0,15 (0,11)	0,17
	γ_5 (AO)			0,03(0,01)	$< 0,001$	0,04 (0,01)	$< 0,001$	0,04 (0,01)	$< 0,001$
ALSFRS	β_2 (SO)	-0,04 (0,02)	0,069			-0,06 (0,02)	$< 0,001$	-0,05 (0,02)	0,055
	β_3 (IMC)	-0,13 (0,05)	0,012			-0,13 (0,05)	0,009	-0,13 (0,05)	0,012
	β_4 (MMT)	0,16 (0,01)	$< 0,001$			0,16 (0,00)	$< 0,001$	0,16 (0,01)	$< 0,001$
	β_5 (CVL)	1,09 (0,18)	$< 0,001$			1,04 (0,18)	$< 0,001$	1,09 (0,18)	$< 0,001$
	β_6 (MCV)	0,09 (0,04)	0,012			0,10 (0,04)	0,007	0,10 (0,04)	0,010
	β_7 (SO$\times t_j$)	0,03 (0,00)	$< 0,001$			0,02 (0,00)	$< 0,001$	0,03 (0,00)	$< 0,001$
	β_8 (MMT$\times t_j$)	0,01 (0,00)	$< 0,001$			0,01 (0,00)	$< 0,001$		
	β_9 (CVL$\times t_j$)	0,08 (0,03)	0,004			0,06 (0,02)	0,018	0,11 (0,03)	$< 0,001$

Les cases grisées correspondent aux résultats discordants de ceux obtenus avec le modèle conjoint à classes latentes (MCCL).
Abréviations : MLM : modèle linéaire mixte, MDS : modèle de survie, MCCL : modèle conjoint à classes latentes, MCEAP : modèle conjoint à effets aléatoires partagés, se : erreur standard. Note : Les covariables suivantes et leurs interactions avec le temps (si significatives) sont présentées : SO (délai depuis le début des symptômes), IMC (Indice de masse corporelle), MMT (capacité musculaire), CVL (capacité vitale lente), MCV (volume corpusculaire moyen), AO (âge au début des signes).

Différentes tendances ont pu être mises en évidence :

— Les **courbes de survie**, **estimées** à l'aide du modèle de survie classique et des deux modèles conjoints sont différentes (nous nous référons à la Figure 4.2 pour la représentation graphique des fonctions de survie estimées pour toutes les covariables à zéro). La courbe de survie du modèle conjoint pour la classe 2 s'éloigne le plus des 3 autres puisqu'elle comprend 45 patients qui ont une très forte évolution de la maladie.

— Bien que les estimations des paramètres soient proches, **les covariables significativement associées au risque de survenue de l'évènement** diffèrent d'un modèle à l'autre. Notamment, le modèle de survie classique a tendance à identifier plus de facteurs associés au risque d'évènement. L'impact de certains de ces facteurs n'est pas significatif dans les modèles conjoints, mais le sont dans le modèle classique : l'IMC dans le modèle conjoint à classes latentes, la capacité musculaire et la capacité vitale lente dans le modèle conjoint à effets aléatoires partagés ne sont plus associées au risque de survenue de l'évènement. Ces résultats laissent à penser que l'utilisation des modèles marginaux, sans tenir compte du lien entre les deux processus, implique une surestimation de certains effets, notamment ceux associés à l'évènement d'intérêt dans notre cas.

— Bien que les estimations des paramètres soient proches, **les covariables significativement associées au score ALSFRS** et à son évolution diffèrent d'un modèle à l'autre. Notamment, dans le modèle linéaire mixte et dans le modèle conjoint à effets aléatoires partagés, la durée depuis les premiers symptômes n'est pas associée au score ALSFRS à l'inclusion, alors qu'elle l'est dans le modèle conjoint à classes latentes. Ce résultat amène à penser qu'un modèle conjoint à classes latentes est à privilégier lorsque l'on suppose qu'il existe des profils différents de patients. En effet, dans ce cas certaines covariables peuvent être significativement associées à l'évolution du marqueur dans les sous-populations, alors qu'elles ne le sont pas dans la population globale.

3.2 Étude de l'adéquation des modèles

Pour le modèle linéaire mixte (classique et considéré comme sous-modèle dans les modèles conjoints), l'adéquation sera évaluée graphiquement, via les résidus conditionnels aux effets aléatoires, définies dans l'Eq.(1.10) et par l'erreur quadratique moyenne (REQM), définie de la manière suivante :

$$\text{REQM} = \sqrt{\frac{1}{n \times \sum_i^n n_i} \sum_i \sum_j \hat{\epsilon}_{ij}}$$

où n_i est le nombre de mesures répétées de l'individu i et $\hat{\epsilon}_{ij}$ le résidu estimé. Le critère d'information d'Akaike (AIC) et le critère d'information bayésien (BIC) peuvent également être utilisés comme un critère de comparaison des modèles.

Pour le modèle de survie (classique et considéré comme sous-modèle dans les modèles conjoints), l'adéquation sera évaluée graphiquement via les résidus de martingales cumulés, définis dans l'Eq.(1.28).

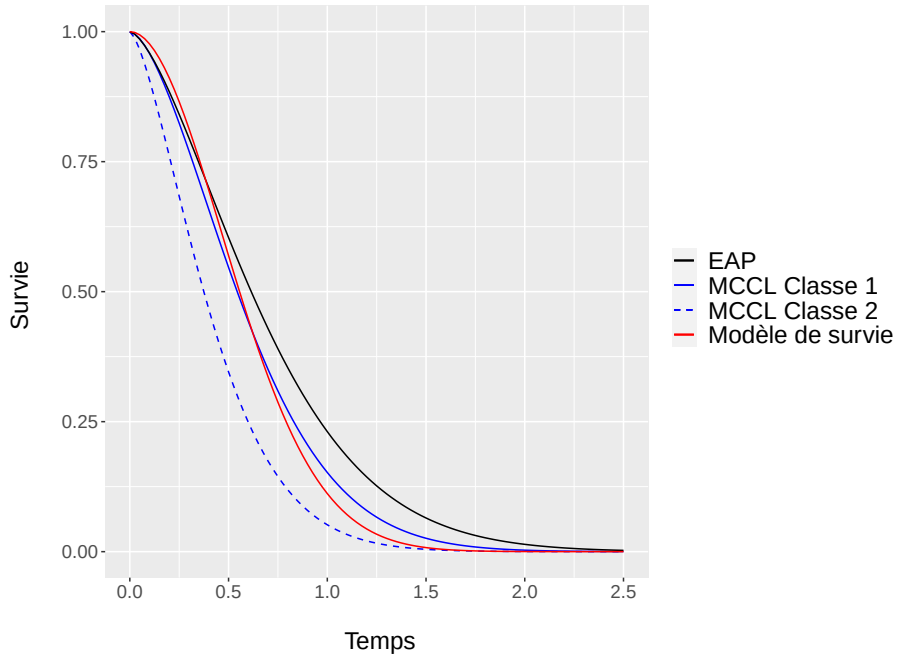


FIGURE 4.2 – Étude TROPHOS : courbes de survie estimées par différents modèles (fonctions de survie de base) : EAP (effets aléatoires partagés), MCCL (modèle conjoint à classes latentes), modèle de survie de Cox (avec toutes les variables explicatives à 0).

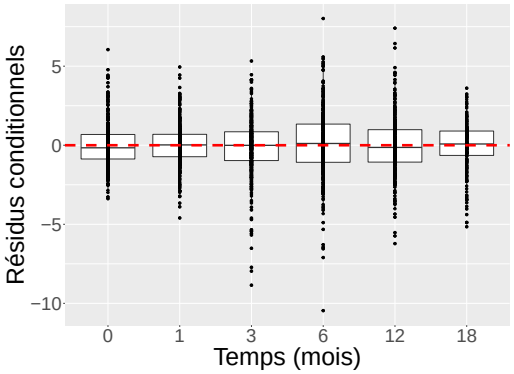
L'adéquation du modèle linéaire mixte est globalement satisfaisante. Les résidus sont centrés autours de zéro pour chaque modèle étudié (nous nous référons à la Figure 4.3).

Une plus grande variabilité des résidus de la classe 2 du modèle conjoint à classes latentes est observée pour les premières visites. Ce phénomène peut s'expliquer par un plus faible effectif dans cette classe au début de l'étude impliquant une diminution de la précision des prédictions. Pour les visites suivantes, le nombre de patients diminue dans la classe 1 à cause de la censure et les résidus deviennent comparables entre les deux classes.

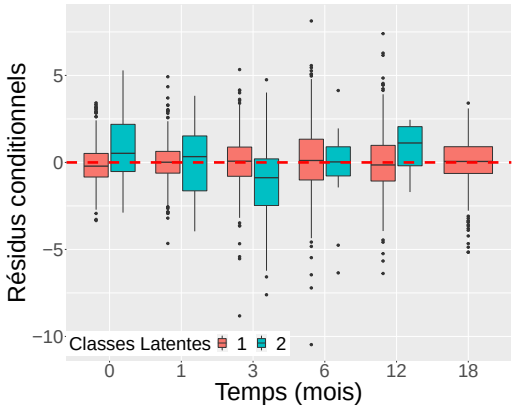
Le REQM des différents modèles est fourni dans le Tableau 4.7. Le modèle conjoint à classes latentes a une meilleure capacité prédictive (REQM=1,60 comparée à 1,63 pour le modèle marginal et le modèle à effets aléatoires partagés). Dans notre étude le modèle conjoint est ainsi le plus adapté pour l'étude de l'évolution de l'ALSFRS.

TABLEAU 4.7 – Étude TROPHOS : adéquation du modèle linéaire mixte, du modèle conjoint à classes latentes et du modèle conjoints à effets aléatoires partagés : erreur quadratique moyenne (REQM).

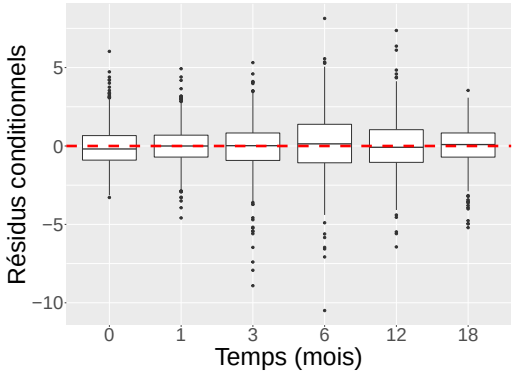
Modèles	REQM
Linéaire mixte	1.63
Conjoint à classes latentes	1.60
Conjoint à effets aléatoires partagés	1.63



(a) Modèle linéaire mixte.



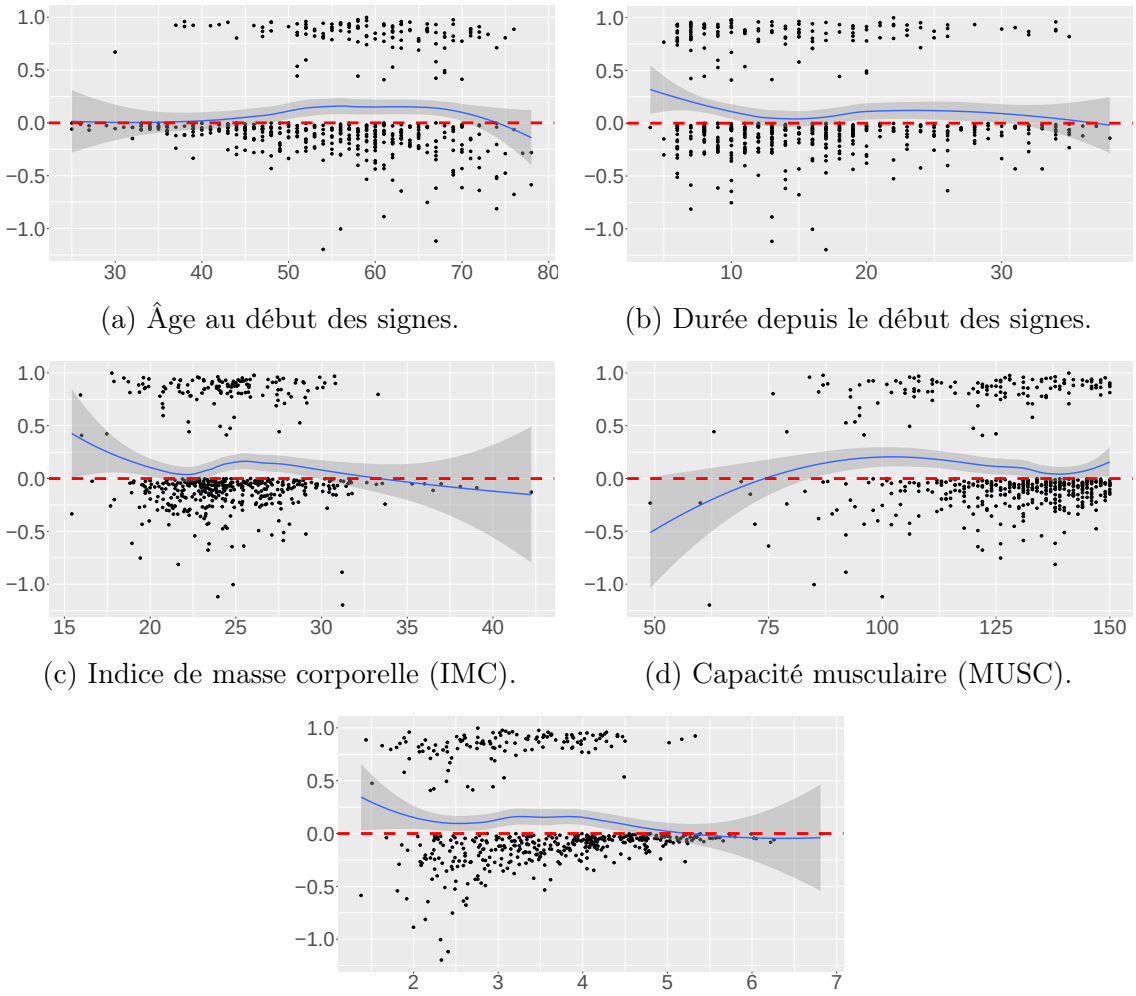
(b) Modèle conjoint à classes latentes.



(c) Modèle conjoint à effets aléatoires partagés.

FIGURE 4.3 – Étude TROPHOS : résidus conditionnels aux effets aléatoires en fonction du temps (mois), obtenus avec le modèle linéaire mixte et les deux modèles conjoints (à classe latentes et à effets aléatoires partagés).

En ce qui concerne **l'adéquation du modèle de survie** (classique ou comme sous-modèle des modèles conjoints), la spécification globale du modèle est correcte. Les résidus de Martingales en fonction des variables explicatives pour le modèle conjoint à classes latentes sont représentés sur la Figure 4.4 (les résultats pour le modèle classique ne sont pas présentés, les tendances sont semblables).



(a) Âge au début des signes.

(b) Durée depuis le début des signes.

(c) Indice de masse corporelle (IMC).

(d) Capacité musculaire (MUSC).

(e) Capacité respiratoire (CVL).

FIGURE 4.4 – Étude TROPHOS : résidus martingales issus du modèle conjoint à classes latentes en fonction des variables explicatives.

Les résidus martingales cumulés au cours du temps permettent de comparer l'adéquation globale des différents modèles ; nous nous référons à la Figure 4.5 pour la représentation graphique. Le modèle de survie classique et le modèle à effets aléatoires partagés montrent une meilleure adéquation que le modèle conjoint à classes latentes. Cela peut être expliqué par une importante différence entre les deux classes latentes en termes de survie (voir Figure 4.2), impliquant les erreurs de prédiction plus importantes. Ainsi, lorsque l'objectif principal est de prédire la survenue de l'évènement, le modèle conjoint à effets aléatoires partagés est mieux adapté. Lorsque l'objectif est d'identifier et caractériser les sous-populations homogènes, le modèle conjoint à classes latentes reste le plus adapté.

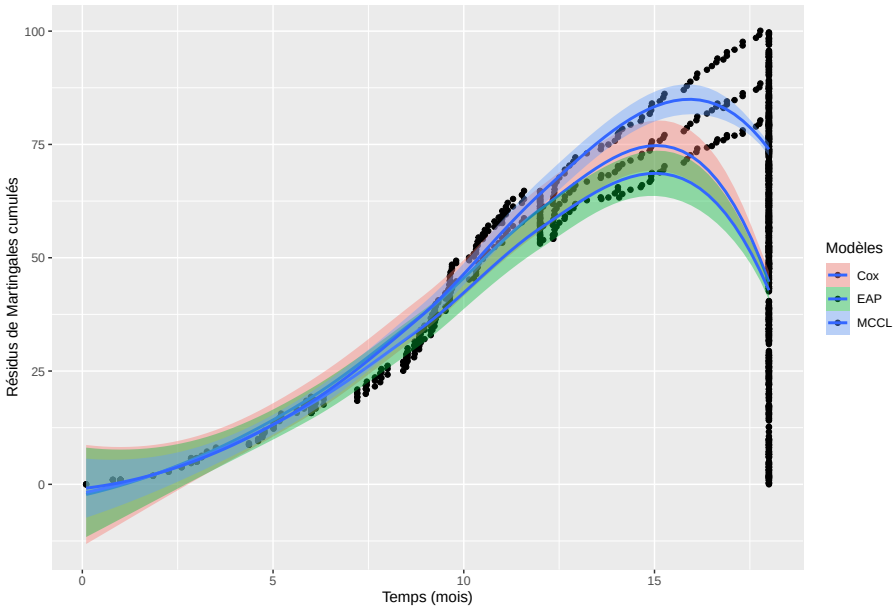


FIGURE 4.5 – Étude TROPHOS : résidus de Martingales cumulés obtenus avec les différents modèles.

4 Résumé

La recherche de facteurs associés à l'évolution de l'ALSFRS et au risque de survenue de l'évènement composite a été réalisée à l'aide de quatre modèles statistiques différents : le modèle linéaire mixte, le modèle de survie, le modèle conjoint à classes latentes et le modèle conjoint à effets aléatoires partagés. Cette étude permet d'avoir un recul nécessaire sur leur utilisation en pratique clinique et notamment, de formuler les éléments suivants :

— Les modèles classiques et les modèles conjoints n'identifient pas les mêmes facteurs de risque pour les outcomes considérés. Les paramètres associés à la survie sont particulièrement concernés dans notre cas d'application. Plus précisément, les modèles classiques ont tendance à trouver plus de facteurs de risque de survie et moins de facteurs influençant l'évolution du biomarqueur. Ainsi, lorsque les mesures du biomarqueur sont censurées par un évènement, nous suggérons l'utilisation des modèles conjoints, qui tiennent compte du lien entre les deux outcomes.

— Les deux modèles conjoints considérés ne fournissent pas les mêmes résultats ni en termes d'identification des facteurs de risque, ni en termes d'adéquation aux données. L'hypothèse sur la structure de la population et l'objectif de l'étude doivent guider le choix du modèle à utiliser. En présence des classes de patients homogènes vis-à-vis de l'évolution de la maladie, un modèle conjoint à classes latentes est mieux adapté. Il présente également une meilleure adéquation pour l'évolution du biomarqueur. En revanche, ce modèle est moins performant en termes d'adéquation pour la survie. Ainsi, lorsque l'objectif principal est l'étude de survie sans identification de sous-population, le modèle à effets aléatoires partagés est mieux adapté.

Conclusions générales et perspectives

Le modèle conjoint à classes latentes permet d'analyser simultanément l'évolution d'un biomarqueur et le délai de survenue d'un évènement en tenant compte de leur dépendance. Il est plus flexible que le modèle conjoint à effets aléatoires partagés car il tient compte d'une hétérogénéité pouvant exister dans la population étudiée sur ces 2 critères ; cette hétérogénéité est modélisée par des classes latentes. Ces classes latentes distinguent des patients présentant des évolutions différentes du biomarqueur et des risques différents de survenue de l'évènement.
Les propriétés mathématiques du modèle conjoint à classes latentes sont bien établies, l'hypothèse sous-jacente étant l'indépendance, conditionnellement aux classes latentes, entre l'évolution longitudinale du biomarqueur et le risque d'évènement. L'algorithme d'estimation des paramètres et les propriétés asymptotiques des estimateurs sont bien identifiés et un package **R** spécifique (*lcmm*) a été développé et publié (46; 59).
L'usage de ce modèle en recherche clinique est potentiellement important et permet : d'étudier l'évolution d'un biomarqueur en tenant compte de la censure par un évènement ; d'étudier le lien entre un biomarqueur longitudinal et le délai d'apparition d'un évènement ; d'étudier des facteurs associés à l'évolution d'un biomarqueur ou à la survenue d'un évènement en tenant compte des classes latentes ; de stratifier les patients en groupes différents selon le délai d'apparition d'un évènement et l'évolution d'un biomarqueur.
Malgré cela, le modèle conjoint à classes latentes demeure sous-utilisé en recherche clinique. L'une des raisons de ce sous-usage réside probablement dans la complexité du modèle (basé sur 3 sous-modèles) et dans la possibilité d'inclure des variables explicatives dans chacun de ces sous modèles avec potentiellement un grand nombre de paramètres à estimer. Aucune étude publiée ne s'est intéressée aux règles d'usage et aux propriétés de ce modèle sur des échantillons finis.

Dans ce travail, nous avons relevé seulement 17 applications médicales utilisant le modèle conjoint à classes latentes, publiées dans des domaines divers, avec des objectifs également différents. Nous avons analysé ces applications médicales afin de fournir aux chercheurs cliniciens des clefs pratiques quant à son utilisation et plus précisément sur l'utilisation de covariables : sans surprise, l'introduction de ces covariables dans tel ou tel sous-modele dépend de l'objectif de l'étude mais les règles ont été explicitées.
Nous avons ensuite réalisé une étude empirique des propriétés et du comportement du modèle. Cette étude se base sur des simulations de Monte-Carlo. Le plan de simulation fait varier le nombre de sujets, le taux de censure (ce qui permet de faire varier le nombre d'évènements mais aussi le nombre de mesures du biomarqueur),

ainsi que le niveau de séparation des classes en termes d'évolution du biomarqueur. Les paramètres du modèle théorique sont choisis pour refléter une situation clinique réelle (étude de Stamenic : « outil pronostique pour la prédiction individualisée du risque d'échec du greffon dans les dix ans suivant la transplantation rénale, en utilisant la progression de la créatinine sérique comme marqueur longitudinal »).

Nous avons montré que l'écart à la normalité des estimations des paramètres peut se produire pour le sous-modèle de survie quand le nombre d'évènements est faible. Le biais relatif (dépendant de la différence entre estimation et valeur réelle) est plus sensible au taux de censure qu'au nombre de patients. Les paramètres les plus impactés sont la pente du sous-modèle mixte et le paramètre de forme de la distribution de Weibull faisant partie du sous-modèle de survie. Ces tendances sont plus prononcées pour les classes faiblement séparées en terme d'évolution du biomarqueur. Le taux de couverture (estimation de la probabilité que la vraie valeur du paramètre soit dans l'intervalle de confiance à 95%) est satisfaisant pour les classes bien séparées. Il est moins bon (environ 85%) quand le taux de censure est élevé pour les classes faiblement séparées. La précision de l'identification des classes est très bonne pour toutes les tailles d'échantillon considérées pour les classes bien séparées. Elle est plus faible, mais satisfaisante quelle que soit la taille de l'échantillon pour les classes faiblement séparées. Les petits groupes latents avec peu d'évènements (fort taux de censure) doivent être interprétés avec prudence, car les estimations des paramètres peuvent être considérablement biaisées.
Ces simulations n'ont pas été réalisées dans l'objectif de proposer des règles pour l'estimation de la taille de l'échantillon. Ce travail reste à faire. En l'absence d'étude spécifique, nous proposons de suivre les recommandations de Peduzzi (53) (reprises par Rizopoulos (9) pour le modèle à effets aléatoires partagés) en les appliquant au sous-modèle logistique (1 covariable pour 10 patients dans la plus petite classe latente) et au sous-modèle de survie (1 variable pour 10 évènements au total). Si des variables spécifiques aux classes sont incluses dans le modèle de survie, cette règle doit s'appliquer en anticipant le plus faible effectif des classes (1 covariable pour 10 évènements dans la plus petite classe latente).

Le modèle conjoint à classes latentes a ensuite été appliqué aux données de l'étude TROPHOS (essai randomisé pour évaluer un nouveau traitement (Olesoxime) chez des patients atteints de *sclérose latérale amyotrophique* (SLA)). L'objectif du clinicien était d'identifier des critères d'inclusion permettant de recruter une population de patients SLA homogène vis-à-vis de l'évolution du handicap (mesurée par le biomarqueur ALSFRS) et de la survenue d'un événement majeur (décès, mise sous ventilation non invasive ou trachéotomie), ce qui permettrait d'augmenter les chances de trouver un traitement efficace pour cette pathologie. La réalisation de cet objectif a été à l'origine de ce travail de thèse.
Quatre classes ont été identifiées à l'aide du modèle conjoint à classes latentes. Ces 4 classes comportaient des patients ayant des évolutions du handicap et des risques de survenue d'évènements très différents. Seule la classe regroupant les patients avec une évolution lente selon les 2 critères de jugement intéressait les neurologues. Elle correspondait à la majorité des patients. Les facteurs indépendants recueillis au diagnostic qui expliquait cette classe par rapport aux autres patients étaient : le

délai depuis les premiers signes, la capacité musculaire et la capacité vitale lente à l'inclusion. Nous avons donc pu recommander aux cliniciens d'inclure des patients avec une capacité musculaire d'au moins 129 points, une capacité vitale de minimum 96% et un délai depuis les premiers signes d'au moins 17 mois.

Trois autres modèles sont fréquemment utilisés pour répondre à certains des objectifs indiqués pour le modèle conjoint à classes latentes : le modèle linéaire mixte classique (sans tenir compte de la censure informative induite par la survenue de l'évènement), le modèle de survie classique (modèle paramétrique) et le modèle conjoint à effets aléatoires partagés. En ce qui concerne le modèle linéaire mixte, bien qu'il soit encore utilisé dans certaines études publiées, il est bien établi que ces résultats sont biaisés en raison de la censure informative induite par l'évènement (60). Si l'objectif est d'étudier le lien entre le biomarqueur et le délai d'apparition de l'évènement, il n'est pas non plus correct d'introduire le biomarqueur comme une variable temps-dépendante dans le modèle de survie classique (61). Ceci est dû au fait que le biomarqueur est une variable endogène et que ses valeurs ne sont généralement pas connues à tous les temps d'évènement. Le modèle conjoint à effets aléatoires partagés est adapté aux objectifs indiqués, mais il ne permet pas de tenir compte d'une hétérogénéité pouvant exister dans la population et pour cette raison, il apparaît moins adapté que le modèle conjoint à classes latentes pour stratifier les patients. Nous avons évalué l'impact des covariables pré-sélectionnées par les cliniciens sur l'évolution du biomarqueur ALSFRS et sur le délai d'apparition de l'évènement composite. Nous avons comparé les résultats obtenus par les 3 modèles mentionnés ci-dessus qui pourraient formellement être employés pour répondre à cet objectif (en ne perdant pas de vue que le modèle linaire mixte est biaisé) à ceux obtenus avec le modèle conjoint à classes latentes.
Les résultats des modèles classiques (linéaire mixte et de survie) et des modèles conjoints diffèrent en particulier sur 2 éléments importants : (1) le délai depuis le début des symptômes est identifié comme un facteur significativement associé au score ALSFRS à la baseline uniquement par le modèle conjoint à classes latentes ; (2) le modèle de survie identifie plus de facteurs associés au risque d'évènement que les modèles conjoints (IMC pour le modèle conjoint à classes latentes ; CVL et MMT pour le modèle à effets aléatoires partagés).
La modélisation conjointe des 2 critères, plutôt que l'analyse séparée du biomarqueur et du délai de l'évènement, doit être la méthode de choix puisqu'elle tient compte de la dépendance entre les 2 critères. Le modèle conjoint à effets aléatoires partagés conduit à des résultats différents de ceux du modèle conjoint à classes latentes lorsqu'il existe une hétérogénéité dans la population sur les critères de jugement comme dans l'exemple de TROPHOS. Il est donc important de s'interroger au préalable sur la plausibilité d'une telle hétérogénéité ou, si aucune connaissance a priori existe, d'effectuer un modèle à classes latentes pour évaluer l'existence d'une hétérogénéité.

Nous espérons que le travail effectué et présenté dans cette thèse contribuera à augmenter l'utilisation du modèle conjoint à classes latentes dans le domaine de la recherche clinique. Nous espérons également que les paramètres de stratification pourront être utilisés dans des essais cliniques sur des traitements de la *sclérose latérale amyotrophique*.

Les perspectives pour la suite de ce travail se divisent en 3 grands types : les perspectives méthodologiques, les perspectives cliniques et les perspectives de publication.

Parmi les **perspectives méthodologiques**, dans un premier temps une étude de simulations portant sur l'impact de certains facteurs sur les outcomes sera effectuée. Dans le cadre de cette étude les résultats obtenus par les différents modèles statistiques (linéaire mixte, survie classique, conjoint à effets aléatoires partagés et conjoint à classes latentes) seront comparés, permettant ainsi de vérifier si les résultats de l'étude TROPHOS, présentés dans le **Chapitre 4** sont généralisables.

Le second objectif est d'étudier la validité des règles concernant la taille d'échantillon, en ce basant, par exemple, sur les travaux de Peduzzi et al. (53). Ces règles pourraient être définies pour le nombre d'évènements par variables ou par paramètre à estimer en ce qui concerne le modèle de survie, et pour le nombre d'individus (ou pour le nombre de mesures répétées par patient) par variable (ou par paramètre) en ce qui concerne le modèle linéaire mixte.

Enfin, des travaux permettant de développer une méthode de calcul du nombre de sujets, basée sur des tests statistiques, devraient être initiés pour assurer une utilisation plus large du modèle. Un test de nullité du coefficient associé à l'effet du traitement pourrait être considéré dans un premier temps.

Concernant les **perspectives cliniques**, les analyses effectuées dans le **Chapitre 3**, c'est-à-dire la stratification des patients SLA et l'identification les facteurs associés aux profils d'évolution de la maladie, pourraient être répliquées sur une cohorte, différente de celle de TROPHOS, qui comporterait un taux de censure plus faible. Cette analyse nous permettrait de valider les résultats trouvés et de conseiller les cliniciens sur les nouveaux critères d'inclusion à prendre en compte pour cibler la thérapie.

Enfin, les **perspectives de publication** sont les suivantes. Un article didactique sur l'utilisation d'un modèle conjoint à classes latentes sera rédigé. Dans cet article les recommandations établies dans notre travail seront articulées dans le but d'aiguiller les statisticiens et les cliniciens dans l'utilisation de ce modèle. Un second article portant sur les simulations comparant les 4 modèles statistiques (linéaire mixte, survie classique, conjoint à effets aléatoires partagés et conjoint à classes latentes) sera rédigé dans le but d'attirer l'attention sur les limites d'utilisation des modèles classiques lorsque les données du biomarqueurs sont censurées par un évènement. Enfin, un article portant sur les résultats de l'application du modèle conjoint à classes latentes à l'étude TROPHOS et à une nouvelle cohorte sera préparé. Le but de cet article sera de proposer aux cliniciens les recommandations quant aux nouveaux critères d'inclusion identifiés.

Annexe A

Résultats des simulations : normalité des paramètres (Chapitre 2, section 2.2)

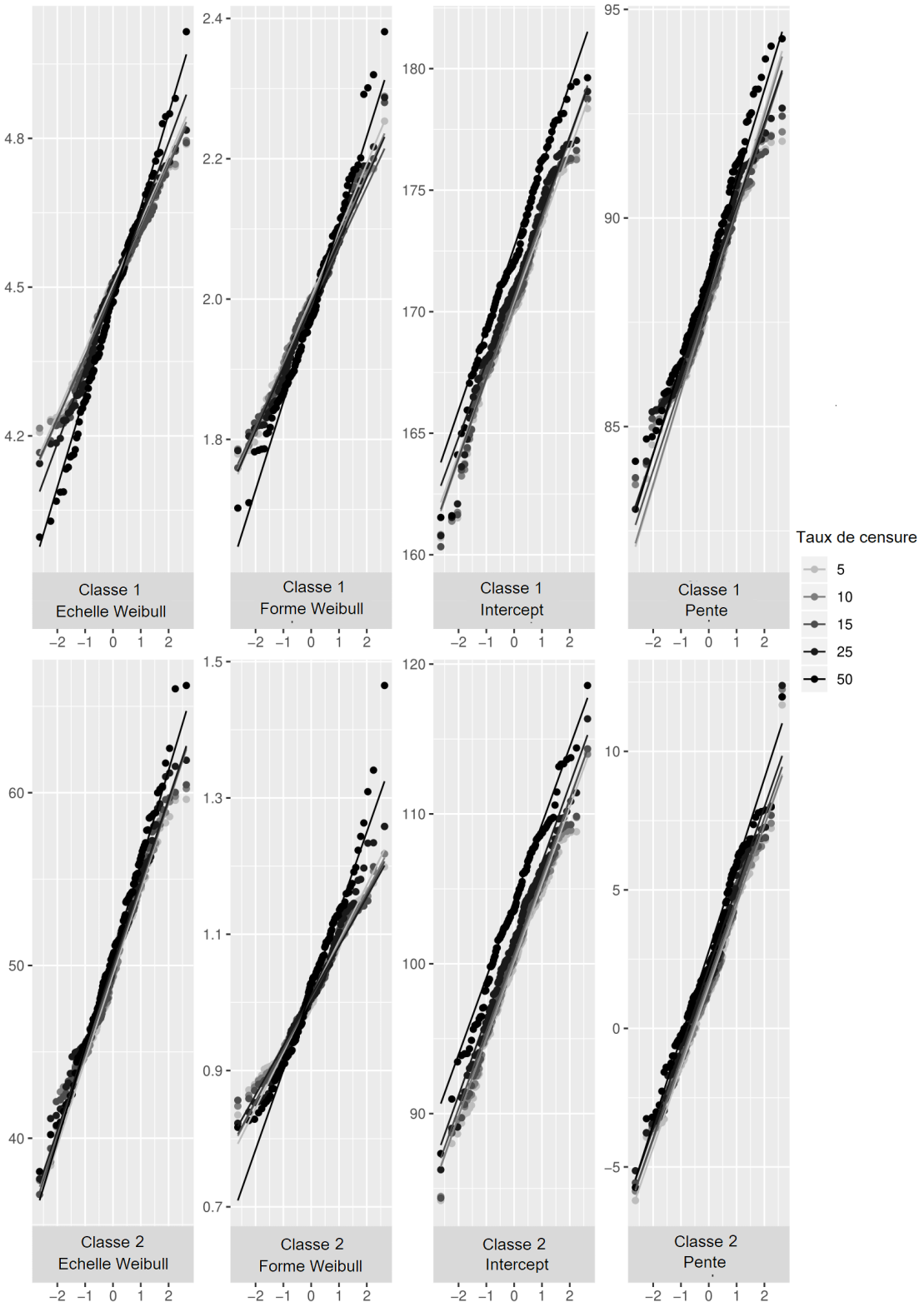


FIGURE A.1 – Résultats des simulations : diagramme quantile-quantile selon le taux de censure, n=500 patients et une *GS*. Les résultats pour l'échelle de Weibull, la forme de Weibull, l'intercept du modèle linéaire mixte et la pente du modèle linéaire mixte.

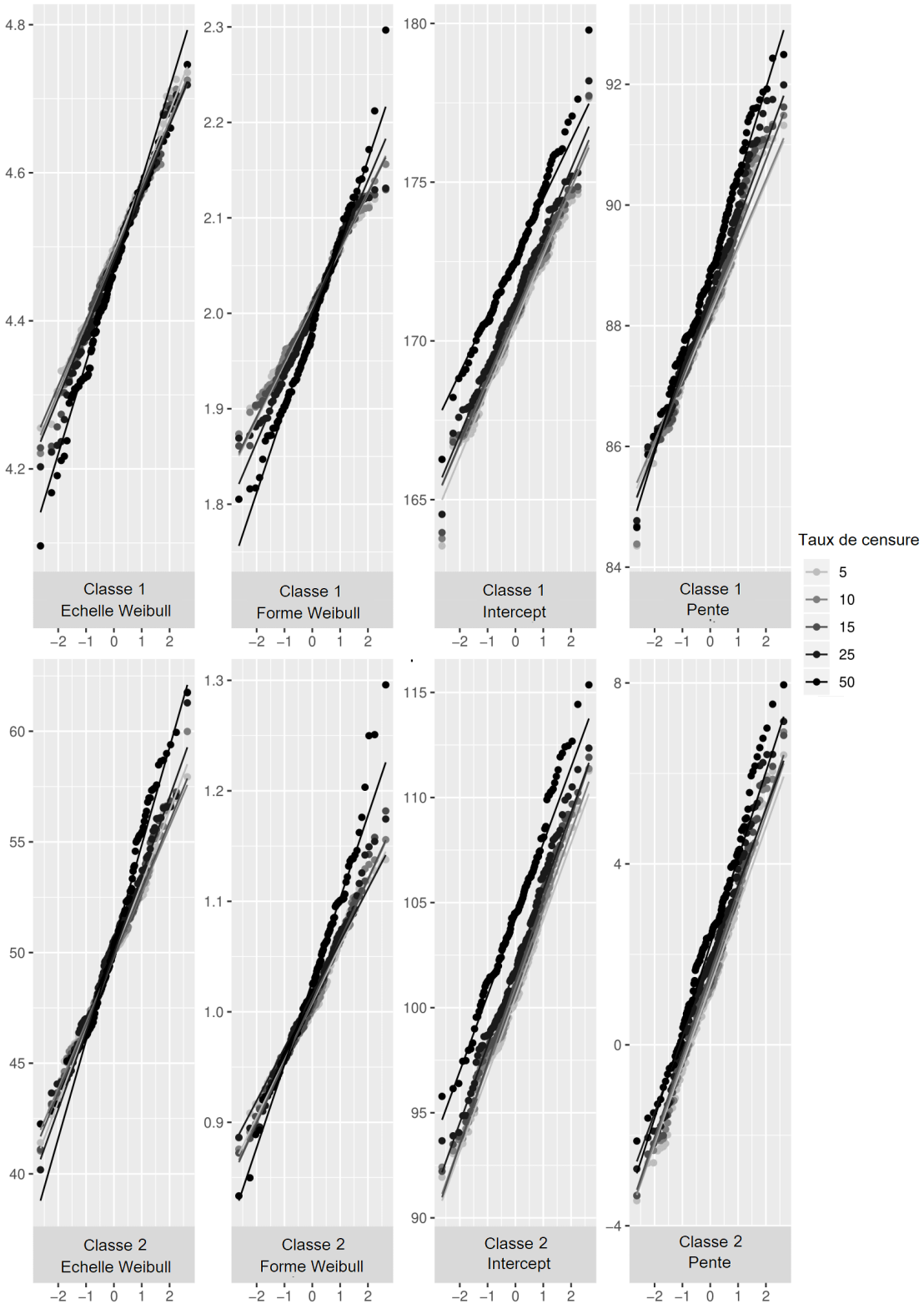


FIGURE A.2 – Résultats des simulations : diagramme quantile-quantile selon le taux de censure, n=1000 patients et une *GS*. Les résultats pour l'échelle de Weibull, la forme de Weibull, l'intercept du modèle linéaire mixte et la pente du modèle linéaire mixte.

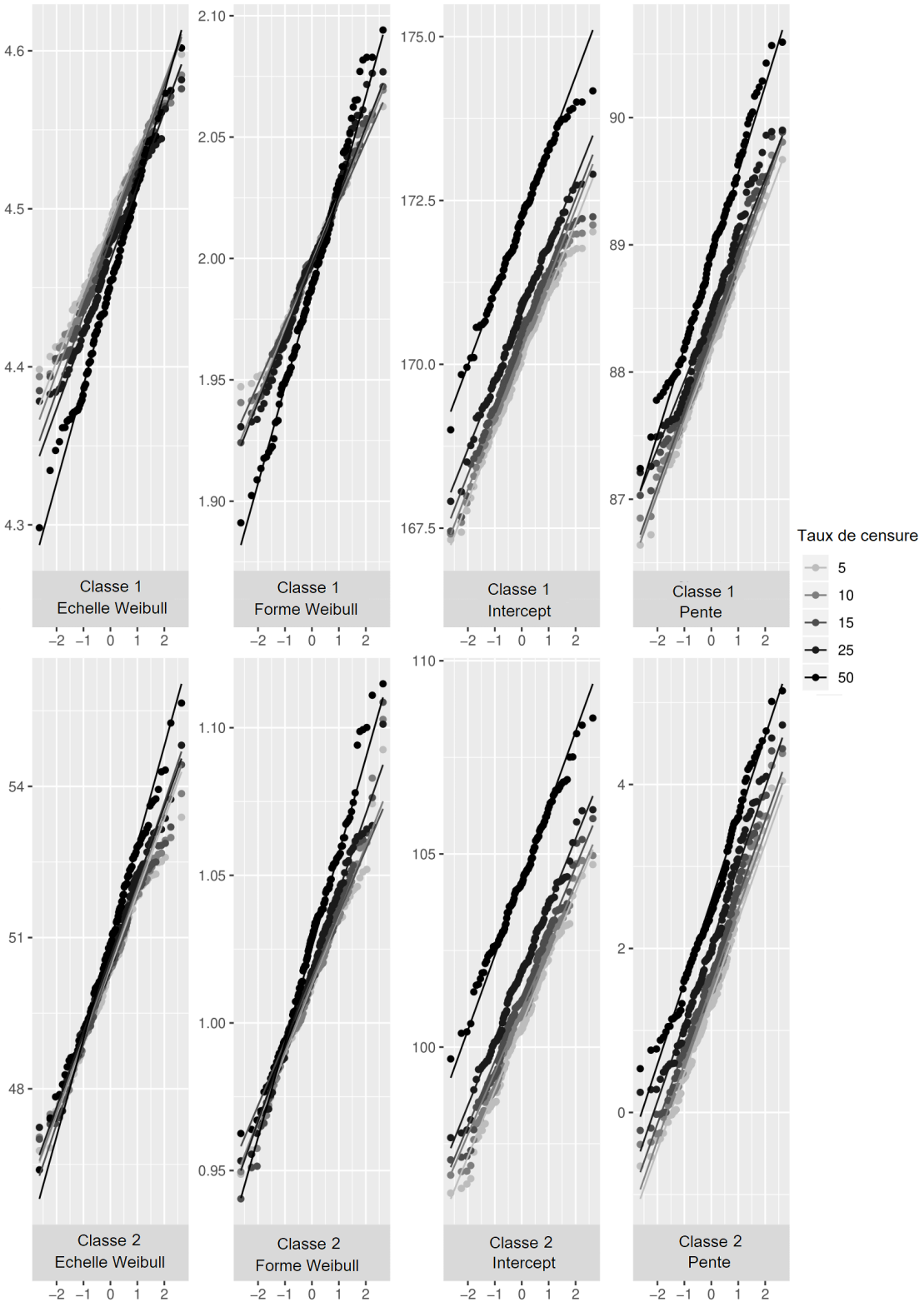


FIGURE A.3 – Résultats des simulations : diagramme quantile-quantile selon le taux de censure, n=5000 patients et une *GS*. Les résultats pour l'échelle de Weibull, la forme de Weibull, l'intercept du modèle linéaire mixte et la pente du modèle linéaire mixte.

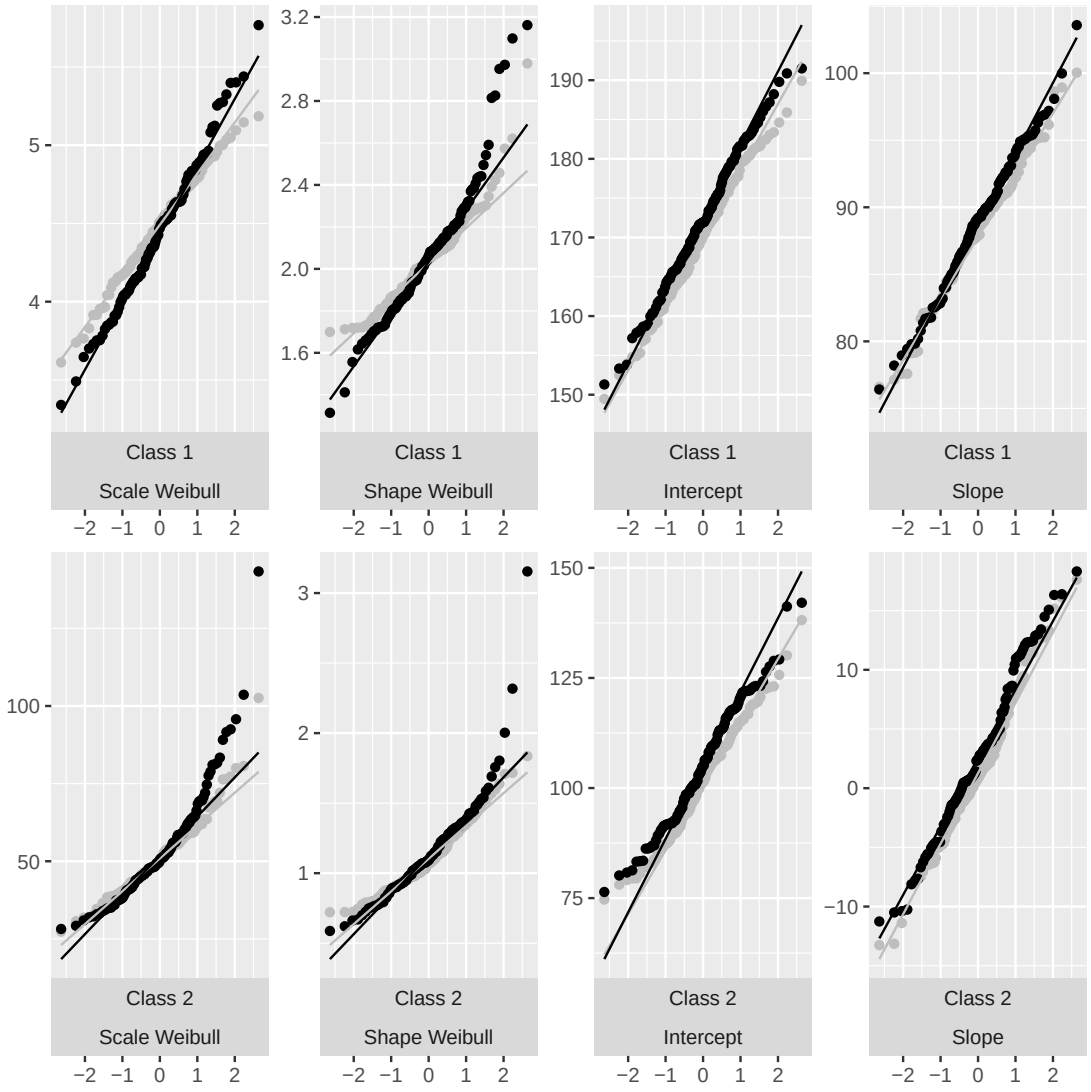


FIGURE A.4 – Résultats des simulations : diagramme quantile-quantile pour les estimations des paramètres, cas *Grande séparation*, $n = 100$. En noir : taux de censure $\tau = 0.5$, en gris clair : $\tau = 0.05$. Les résultats pour l'échelle de Weibull, la forme de Weibull, l'intercept du modèle linéaire mixte et la pente du modèle linéaire mixte.

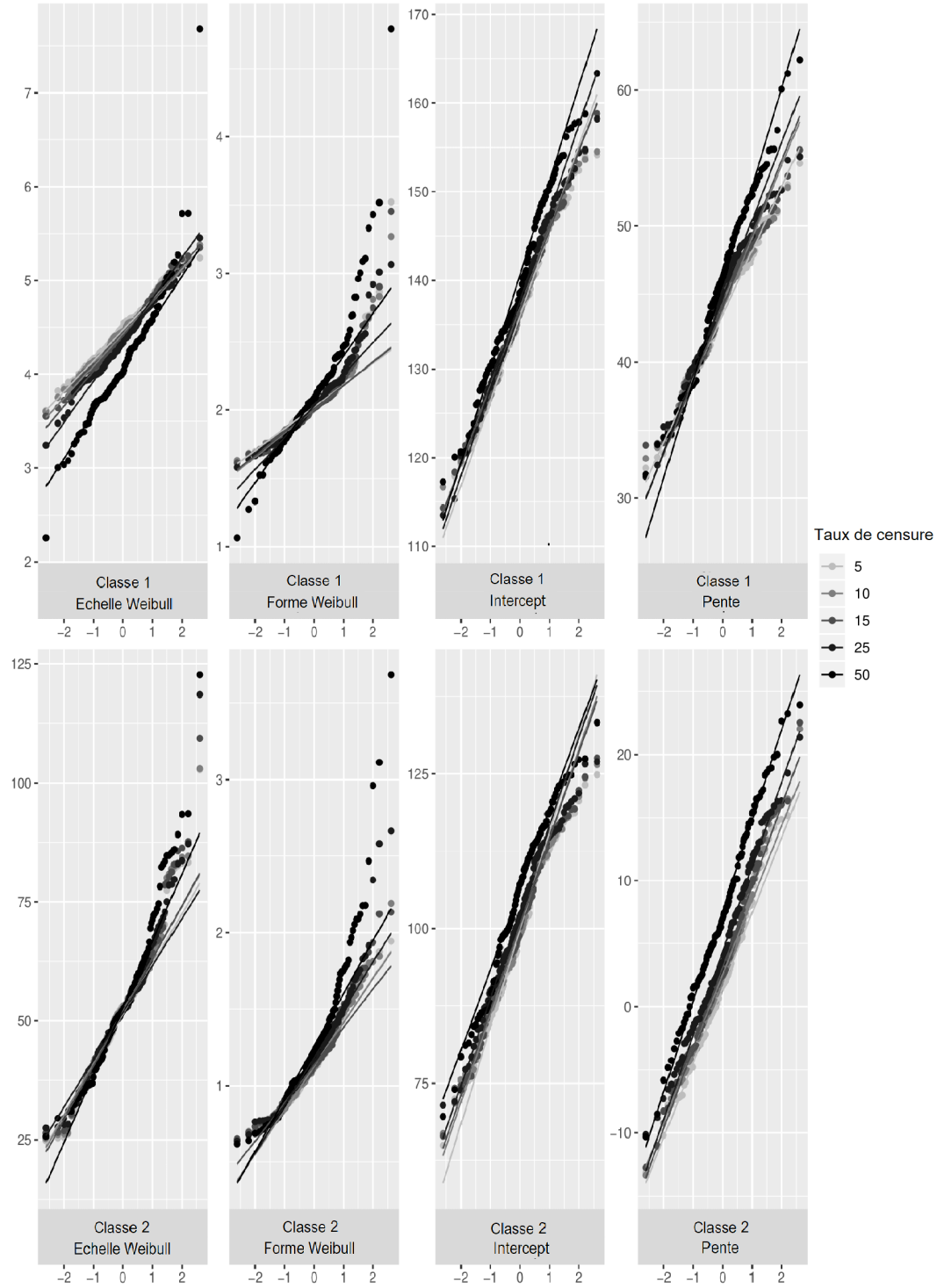


FIGURE A.5 – Résultats des simulations : diagramme quantile-quantile selon le taux de censure, n=100 patients et une *FS*.Les résultats pour l'échelle de Weibull, la forme de Weibull, l'intercept du modèle linéaire mixte et la pente du modèle linéaire mixte.

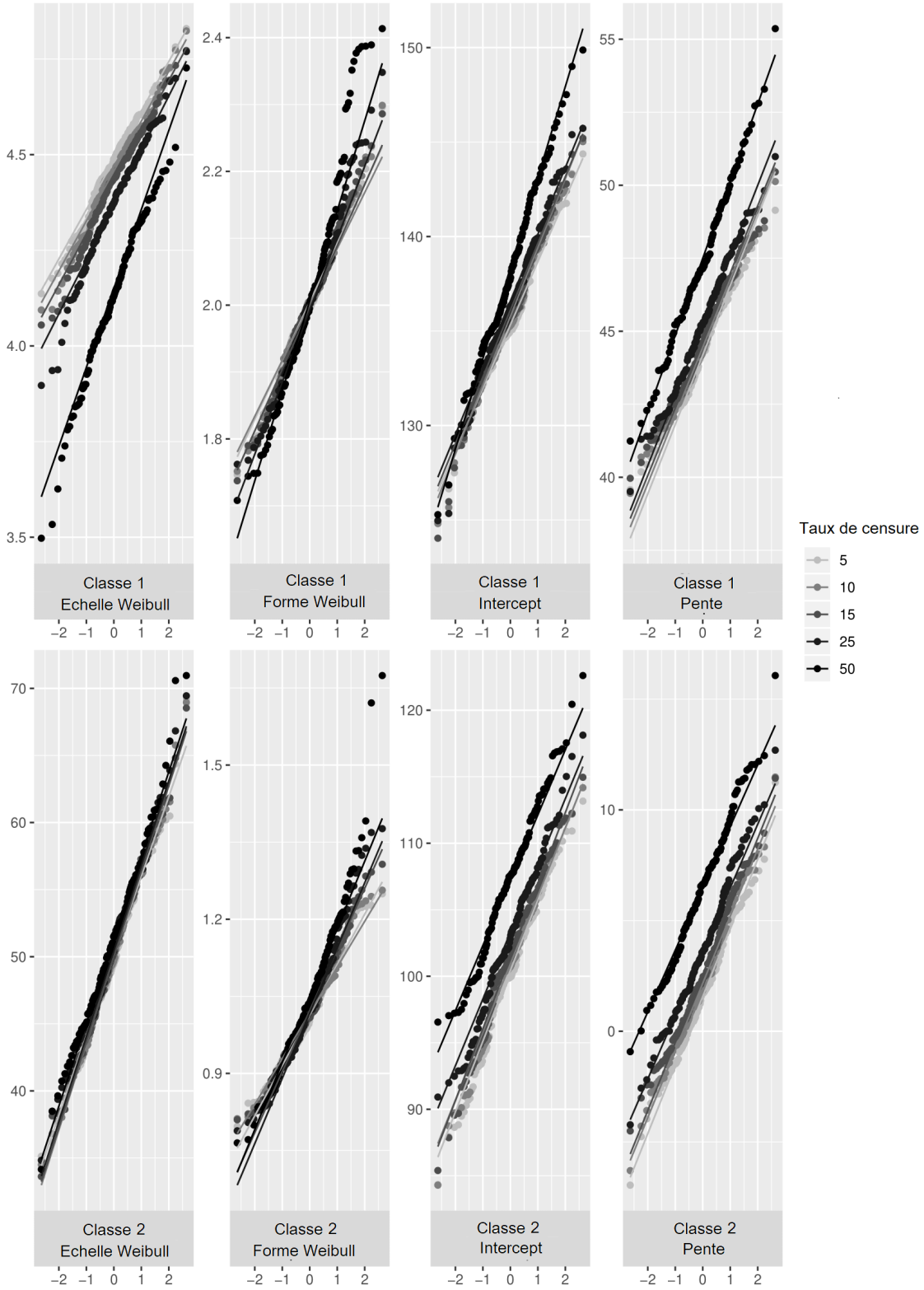


FIGURE A.6 – Résultats des simulations : diagramme quantile-quantile selon le taux de censure, n=500 patients et une *FS*.Les résultats pour l'échelle de Weibull, la forme de Weibull, l'intercept du modèle linéaire mixte et la pente du modèle linéaire mixte.

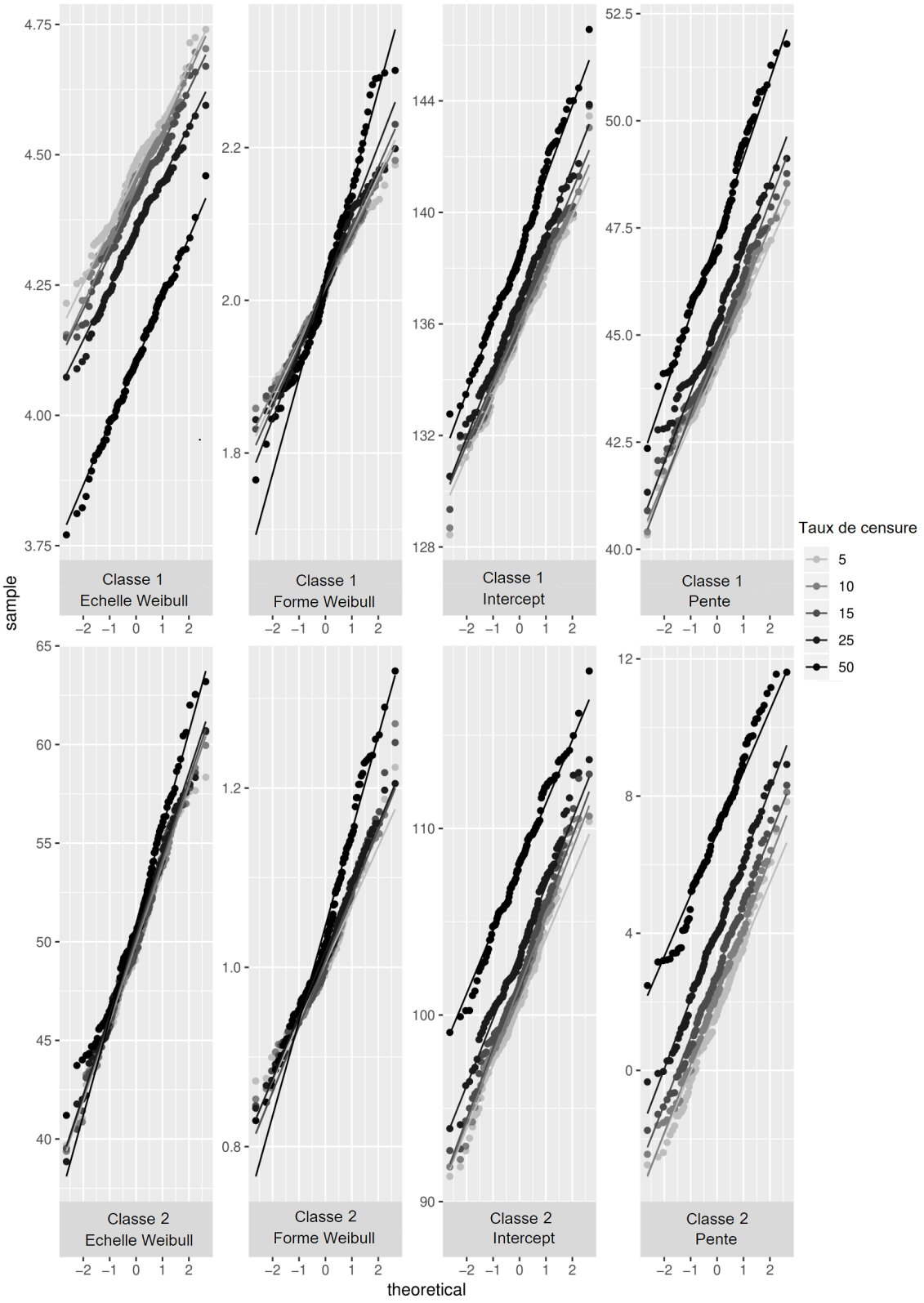


FIGURE A.7 – Résultats des simulations : diagramme quantile-quantile selon le taux de censure, n=1000 patients et une *FS*.Les résultats pour l'échelle de Weibull, la forme de Weibull, l'intercept du modèle linéaire mixte et la pente du modèle linéaire mixte.

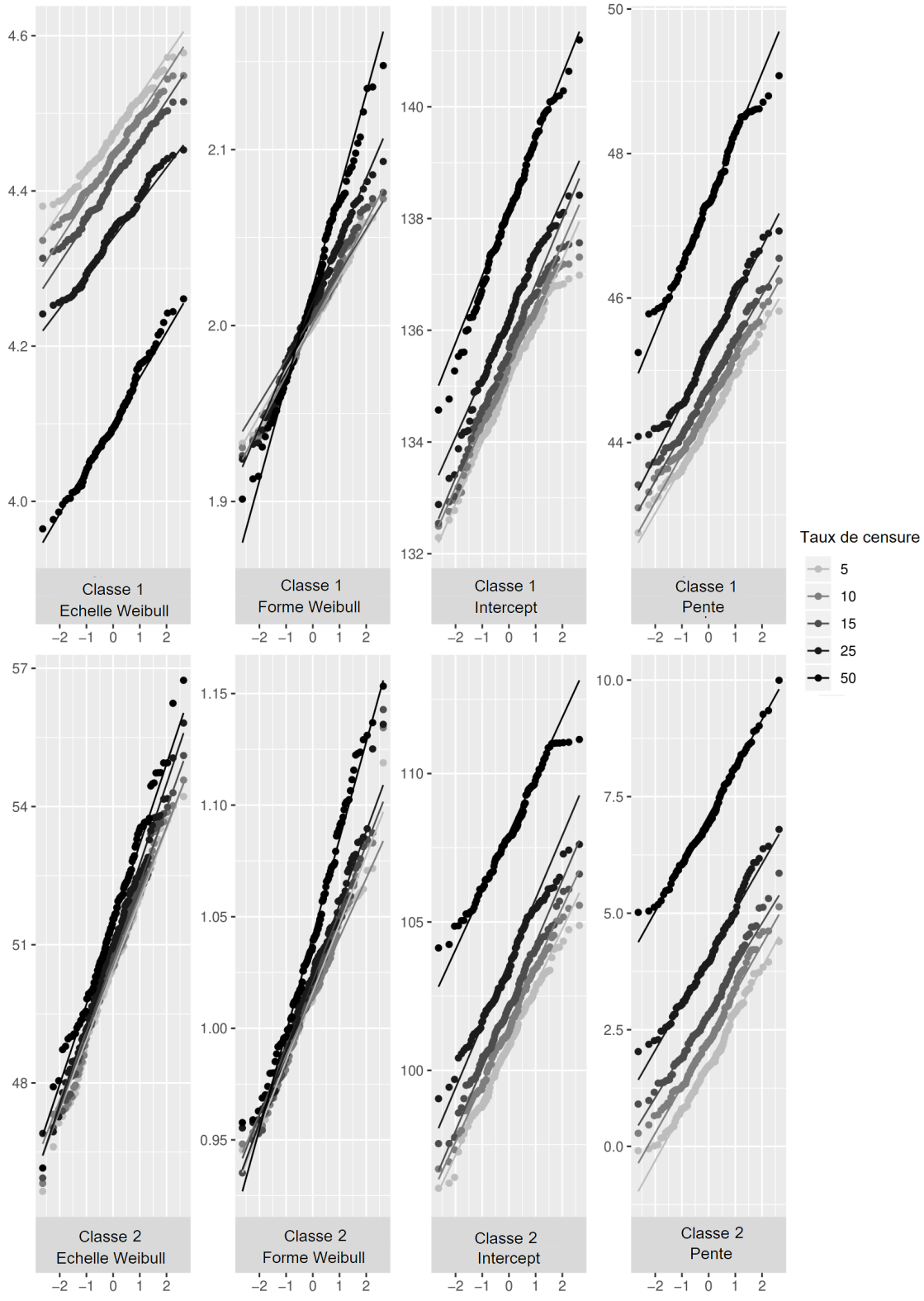


FIGURE A.8 – Résultats des simulations : diagramme quantile-quantile selon le taux de censure, n=5000 patients et une *GS*.Les résultats pour l'échelle de Weibull, la forme de Weibull, l'intercept du modèle linéaire mixte et la pente du modèle linéaire mixte.

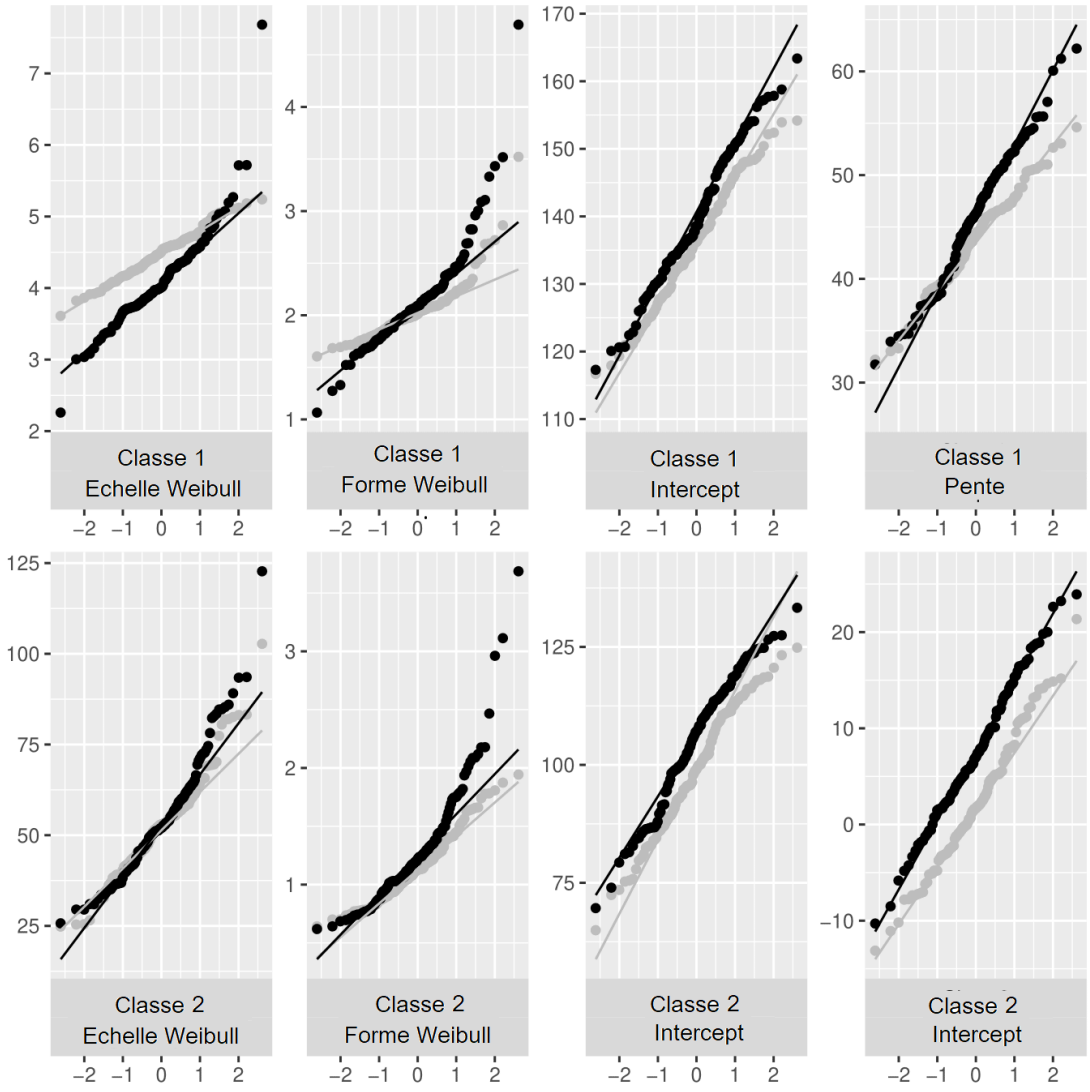


FIGURE A.9 – Résultats des simulations : diagramme quantile-quantile pour les estimations des paramètres, cas *Faible séparation*, $n = 100$. En noir : taux de censure $\tau = 0.5$, en gris clair : $\tau = 0.05$. Les résultats pour l'échelle de Weibull, la forme de Weibull, l'intercept du modèle linéaire mixte et la pente du modèle linéaire mixte.

Annexe B

Résultats des simulations : Biais relatif (Chapitre 2, Section 2.2)

TABLEAU B.1 – Résultats des simulations : biais relatif des paramètres du modèle pour le cas d'une *GS* en fonction du nombre d'individus, n, et du taux de censure, τ. Les estimations de l'erreur et des écarts-types de l'intercept aléatoire ($\hat{\sigma}_\epsilon$ et $\hat{\sigma}_b$ respectivement), de l'intercept et de la pente ($\hat{\beta}_{0g}$, $\hat{\beta}_{1g}$ respectivement) du sous-modèle longitudinal et de l'échelle et de la forme de Weibull du sous-modèle de survie ($\hat{\zeta}_{1g}$ et $\hat{\zeta}_{2g}$ respectivement) sont présentés. g : identification de l'appartenance à une classe.

		Sous-modèle linéaire mixte						Sous-modèle de survie			
n	τ	$\hat{\sigma}_b$	$\hat{\sigma}_\epsilon$	$\hat{\beta}_{0g}$		$\hat{\beta}_{1g}$		$\hat{\zeta}_{1g}$		$\hat{\zeta}_{2g}$	
				$g=1$	$g=2$	$g=1$	$g=2$	$g=1$	$g=2$	$g=1$	$g=2$
100	5	0,1053	21,1797	0,0429	1,3974	0,0281	10,7052	0,3018	3,2362	2,5023	9,8304
	10	0,0961	21,2395	0,1421	1,6954	0,0748	38,1642	0,2923	3,1830	2,4243	10,0367
	15	0,0997	21,3986	0,2846	1,9987	0,1936	59,0676	0,4148	3,6286	2,3160	10,5283
	25	0,0824	21,3810	0,4687	3,0567	0,4054	82,7586	0,4455	3,6646	2,2878	11,9084
	50	1,7125	21,8392	1,3486	5,3694	0,8271	120,1328	1,1201	7,5050	3,5401	14,0533
500	5	0,2267	21,4488	0,1534	0,1412	0,0685	25,4826	0,0591	0,5932	0,0332	0,2021
	10	0,2230	21,3820	0,2346	0,2261	0,1317	35,8174	0,1892	0,4996	0,1561	0,0727
	15	0,2062	21,4656	0,3380	0,4738	0,2027	48,1685	0,3062	0,0501	0,1232	0,0229
	25	0,2011	21,4412	0,5670	1,3390	0,4127	74,1988	0,2546	0,6707	0,4496	0,1703
	50	0,1945	21,6664	1,3876	3,7674	0,9627	116,6916	0,7377	1,6005	0,6217	1,4936
1000	5	0,0004	20,1935	0,3260	0,7563	0,2925	8,4122	0,1684	0,1125	0,4160	0,2180
	10	0,0015	20,2002	0,4197	1,0629	0,3599	19,7922	0,3399	0,0777	0,4994	0,0130
	15	0,0117	20,1928	0,5181	1,3744	0,4324	31,9018	0,3658	0,2015	0,3471	0,1781
	25	1,6848	20,2110	0,7436	2,1019	0,6690	60,0292	0,4954	0,6364	0,1559	0,4325
	50	0,0072	20,3934	1,5776	4,6004	1,0896	102,7405	0,8822	1,3124	0,5011	1,9393
5000	5	1,6342	19,9957	0,0242	0,5270	0,2185	16,5831	0,1929	0,6064	0,0605	0,2997
	10	1,6315	19,9945	0,1226	0,8526	0,3075	31,0770	0,2985	0,7797	0,0810	0,3320
	15	0,0380	19,9728	0,2215	1,2091	0,3905	44,6046	0,4162	0,8876	0,0823	0,5370
	25	0,0449	20,0112	0,4584	1,9183	0,5515	72,0024	0,6396	1,2441	0,0648	0,8037
	50	0,0400	20,2601	1,2738	4,3253	1,0406	113,4169	1,0629	1,8816	0,5238	1,7593

TABLEAU B.2 – Résultats des simulations : biais relatif des paramètres du modèle pour le cas d'une *FS* en fonction du nombre d'individus, n, et du taux de censure, τ. Les estimations de l'erreur et des écarts-types de l'intercept aléatoire ($\hat{\sigma}_\epsilon$ et $\hat{\sigma}_b$ respectivement), de l'intercept et de la pente ($\hat{\beta}_{0g}$, $\hat{\beta}_{1g}$ respectivement) du sous-modèle longitudinal et de l'échelle et de la forme de Weibull du sous-modèle de survie ($\hat{\zeta}_{1g}$ et $\hat{\zeta}_{2g}$ respectivement) sont présentés. g : identification de l'appartenance à une classe.

		Sous-modèle linéaire mixte						Sous-modèle de survie			
n	τ	$\hat{\sigma}_b$	$\hat{\sigma}_\epsilon$	$\hat{\beta}_{0g}$		$\hat{\beta}_{1g}$		$\hat{\zeta}_{1g}$		$\hat{\zeta}_{2g}$	
				$g=1$	$g=2$	$g=1$	$g=2$	$g=1$	$g=2$	$g=1$	$g=2$
100	5	0,2157	22,3405	0,9001	1,1061	0,6187	77,4912	0,1821	5,5733	2,6803	14,4064
	10	3,3939	22,1855	1,1485	0,5399	0,2169	129,4999	0,7945	5,9039	2,4429	14,9468
	15	0,2315	22,2708	1,3985	0,0946	0,4619	175,0008	1,4886	6,5439	2,4417	16,5623
	25	1,5321	22,1504	1,8233	1,7052	1,6413	273,7677	3,1293	6,4496	3,1511	20,3027
	50	1,3699	21,7546	3,6344	5,3290	4,6976	532,4839	8,3763	9,3616	7,9156	30,6630
500	5	0,2556	21,5565	0,2726	0,0707	0,6062	36,0688	0,4649	0,1000	0,0440	0,5662
	10	1,9185	21,3569	0,4126	0,8612	1,1048	79,5267	1,1380	0,1392	0,0640	0,9334
	15	3,4729	21,2975	0,6091	1,3911	1,5773	120,2866	1,6905	0,6533	0,1375	1,3348
	25	0,1728	20,9193	0,9847	3,0168	2,8914	212,8892	3,0037	1,7108	0,1707	2,2996
	50	1,7796	19,8756	2,3534	7,4185	7,8224	444,8984	8,3268	2,8825	0,9417	4,7449
1000	5	0,0184	20,1206	0,5111	0,9550	0,8944	44,8903	0,6416	0,3375	0,6727	0,2864
	10	0,0302	20,0392	0,6927	1,5469	1,3819	87,2876	1,2630	0,2644	0,8368	0,0338
	15	0,0424	19,8918	0,8955	2,1650	1,8632	139,1406	1,8765	0,0905	0,9058	0,2346
	25	0,0763	19,6487	1,3465	3,5027	3,2279	240,5047	3,3728	0,7920	1,0379	0,6998
	50	0,2146	18,4371	2,7442	8,0212	7,5645	482,3023	8,8312	2,2757	1,4414	3,6641
5000	5	1,6360	19,8545	0,0630	0,8358	0,6849	48,0499	0,6231	0,9167	0,0643	0,5431
	10	0,0529	19,7483	0,2447	1,4861	1,2312	95,3766	1,2469	1,1418	0,1531	0,6085
	15	1,5950	19,6407	0,4366	2,1381	1,7922	142,7663	1,9299	1,3603	0,2632	0,9193
	25	0,1003	19,3803	0,8741	3,5248	3,0292	240,2600	3,4624	1,9415	0,5656	1,3986
	50	0,2294	18,2465	2,2807	7,9142	7,5368	487,8892	8,8603	3,0965	0,9679	3,1725

Annexe C

Étude TROPHOS : matrices du modèle linéaire mixte (Chapitre 4, Section 1.1)

Notations :

— n_i =18 : nombre de mesures de l'individu i ;

— $p = 10$: nombre de covariabes à effets fixes (temps, 5 covariables, 3 interactions, intercept) ;

— $q = 2$: nombre de covariables à effets aléatoires (pente et intercept). Les effets aléatoires sont non-corrélés.

$$\underbrace{ALSFRS_{i\,1\times n_i}^{\intercal}}_{} = \begin{pmatrix} ALSFRS_i(T_0) & ALSFRS_i(T_1) & ALSFRS_i(T_3), ALSFRS_i(T_6) & ALSFRS_i(T_9) & ALSFRS_i(T_12) & ALSFRS_i(T_18) \end{pmatrix}$$

$$\underbrace{X_i^{\intercal}}_{n_i\times p} = \begin{pmatrix} 1 & T_{i0} & SO_i & BMI_i & MMT_i & CVL_i & MCV_i & SO_i\times T_{i0} & MMT_i\times T_{i0} & CVL_i\times T_{i0} \\ 1 & T_{i1} & SO_i & BMI_i & MMT_i & CVL_i & MCV_i & SO_i\times T_{i1} & MMT_i\times T_{i1} & CVL_i\times T_{i1} \\ 1 & T_{i3} & SO_i & BMI_i & MMT_i & CVL_i & MCV_i & SO_i\times T_{i3} & MMT_i\times T_{i3} & CVL_i\times T_{i3} \\ . & . & . & . & . & . & . & . & . & \\ . & . & . & . & . & . & . & . & . & \\ . & . & . & . & . & . & . & . & . & \\ 1 & T_{i18} & SO_i & BMI_i & MMT_i & CVL_i & MCV_i & SO_i\times T_{i18} & MMT_i\times T_{i18} & CVL_i\times T_{i18} \end{pmatrix}$$

$$\underbrace{\beta}_{p\times 1} = \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 \cdots \beta_9 \end{pmatrix}$$

$$\underbrace{Z_i^{\intercal}}_{n_i\times q} = \begin{pmatrix} 1 & T_{i0} \\ 1 & T_{i1} \\ 1 & T_{i3} \\ . & . \\ . & . \\ . & . \\ 1 & T_{i18} \end{pmatrix}, \underbrace{b}_{q\times 1} = \begin{pmatrix} b_0 & b_1 \end{pmatrix}, \underbrace{B}_{q\times q} = \begin{pmatrix} \sigma_{b0}^2 & 0 \\ 0 & \sigma_{b1}^2 \end{pmatrix}$$

Ainsi, l'Eq.(1.14) est écrit sous la forme matriciel de l'Eq.(4.1) :

$$ALSFRS_i = X_i^{\intercal}\beta + Z_i^{\intercal}b + \epsilon_i \tag{C.1}$$

Annexe D

Graphique des distributions des temps d'évènement non censurés selon les 2 classes latentes retouvées (Chapitre 4, section 2.1)

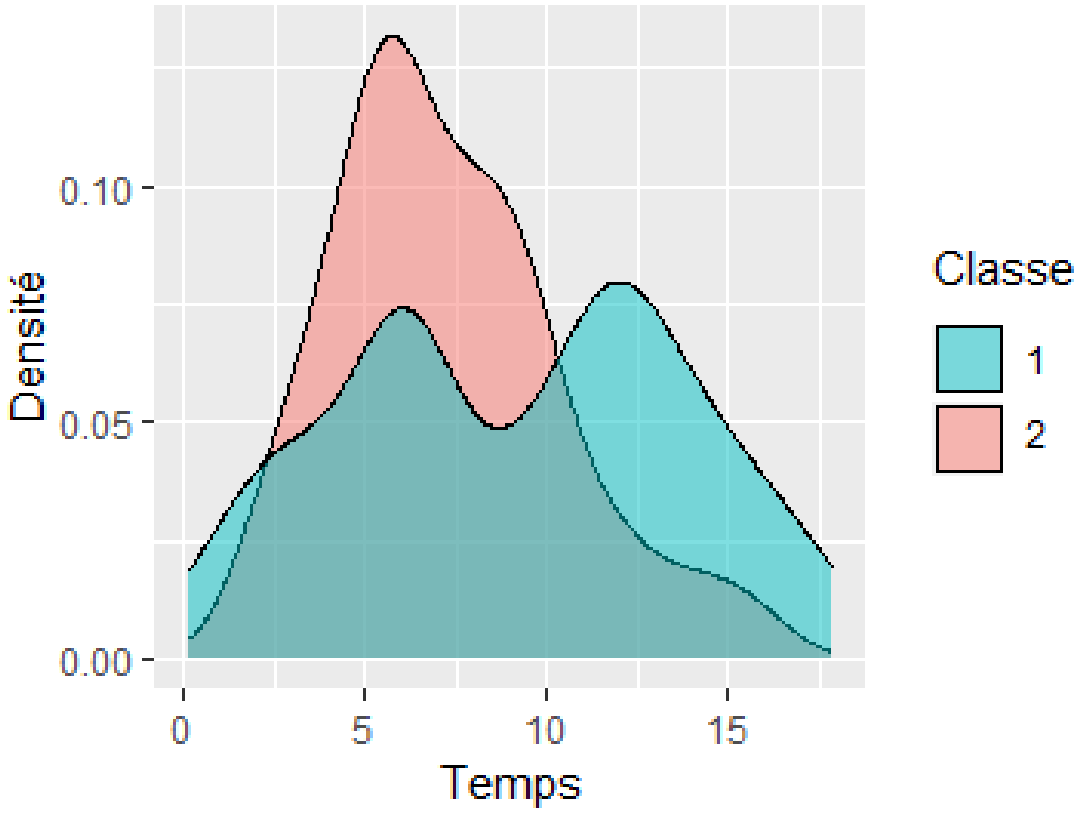


FIGURE D.1 – Distribution des temps d'évènement non censurés selon les deux classes latentes trouvées.

Annexe E

Article

Kyheng *et al. BMC Medical Research Methodology* (2021) 21:198
https://doi.org/10.1186/s12874-021-01377-9

BMC Medical Research Methodology

RESEARCH Open Access

Joint latent class model: Simulation study of model properties and application to amyotrophic lateral sclerosis disease

Maéva Kyheng[1,2*], Génia Babykina[1,2], Camille Ternynck[1,2], David Devos[3], Julien Labreuche[3] and Alain Duhamel[1,2]

Abstract

Background: In many clinical applications, evolution of a longitudinal marker is censored by an event occurrence, and, symmetrically, event occurrence can be influenced by the longitudinal marker evolution. In such frameworks joint modeling is of high interest. The Joint Latent Class Model (JLCM) allows to stratify the population into groups (classes) of patients that are homogeneous both with respect to the evolution of a longitudinal marker and to the occurrence of an event; this model is widely employed in real-life applications. However, the finite sample-size properties of this model remain poorly explored.

Methods: In the present paper, a simulation study is carried out to assess the impact of the number of individuals, of the censoring rate and of the degree of class separation on the finite sample size properties of the JLCM. A real-life application from the neurology domain is also presented. This study assesses the precision of class membership prediction and the impact of covariates omission on the model parameter estimates.

Results: Simulation study reveals some departures from normality of the model for survival sub-model parameters. The censoring rate and the number of individuals impact the relative bias of parameters, especially when the classes are weakly distinguished. In real-data application the observed heterogeneity on individual profiles in terms of a longitudinal marker evolution and of the event occurrence remains after adjusting to clinically relevant and available covariates;

Conclusion: The JLCM properties have been evaluated. We have illustrated the discovery in practice and highlights the usefulness of the joint models with latent classes in this kind of data even with pre-specified factors. We made some recommendations for the use of this model and for future research.

Keywords: Joint model, Latent classes, Survival analysis, Linear mixed model, MLE properties, Monte Carlo simulations, Amyotrophic lateral sclerosis

*Correspondence: m.kyheng.chr@gmail.com
[1]ULR 2694 - METRICS : évaluation des technologies de santé et des pratiques médicales, Univ. Lille, CHU Lille, Lille, France
[2]Département de Biostatistiques, CHU Lille, Lille, France
Full list of author information is available at the end of the article

Background

Joint models for longitudinal and time-to-event data are now widespread due to large cohort studies allowing collection of repeated measures of biomarkers and clinical events times [1]. The most popular way to analyze this kind of combined data are the *shared random effects models*, proposed by Wulfsohn and Tsiatis [2], where a function of random effects, issued from the model for longitudinal marker, is included as a covariate into the survival model. This approach allows to explain the relation between a longitudinal parameter and a time-to-event, assuming a homogeneous population. However, for certain diseases, the homogeneity assumption is not met and existence of different profiles of biomarker progression and/or of the time to-event should be accounted for in the model.

Mixture models are widely used in medical research. Different extensions allowing to account for the potential heterogeneity in population were proposed. Verbeke and Lesaffre [3] extended the mixture model to longitudinal data, assuming a latent profile of the biomarker progression (growth mixture model GMM). Muthén and Shedden [4] jointly studied longitudinal data with a binary outcome. Lin et al. [5] developed the joint latent class model (JLCM) replacing the binary outcome by a time-to-event. The JLCM allows firstly to account for the dependency between a longitudinal biomarker and a time-to-event by distinguishing between different profiles of biomarker progression associated with the risk of event. Secondly, it allows to analyze different profiles of longitudinal biomarker process censored by the event occurrence. Finally, the JLCM provides predictions for the risk of event conditional on the biomarker progression.

Very flexible, the JLCM remains quite complex. Indeed, it is composed of 3 sub-models (a multinomial logistic regression for latent classes, a linear mixed model for longitudinal process and a survival model for the time-to-event) and each of these sub-models can include covariates with effects specific or common to the latent classes.

To our knowledge, very few papers deal with studying the properties and the behaviour of the JLCM, for example Proust-Lima et al. [6], therefore it is rarely used in published clinical studies. Using a literature search of MEDLINE and WOS until december 2020, we found only 8 medical papers published since the model development in 2002 [5]. These papers appeared following a comprehensive methodology review concerning the JLCM [7] and have different objectives. These objectives can be summarized as follows: 1) to study the relationship between a longitudinal biomarker and the risk of event [8–11]; 2) to identify sub-groups of longitudinal biomarker progression censored by the event occurrence [12]; 3) to study the impact of different factors on the longitudinal biomarker progression censored by the event occurrence [13]; 4) to predict the risk of an event based on the longitudinal biomarker progression [14, 15]. Different implementations of the model were proposed to achieve a same objective. For example, for the first objective, Syrjälä et al. [8] search for the relation between childhood food consumption and the risk of advanced islet autoimmunity using a JLCM without covariates; Brilleman et al. [9] explore the relationship between the changes in body mass index and the risk of death and/or transplant in hemodialysis patients by means of the JLCM for competing risks, including the pre-specified covariates with a common effect on latent classes only in the survival sub-model; Ogata et al. [10] and Portegies et al. [11] analyze the association between fasting plasma glucose progression and the risk of cardiovascular disease and the association between the blood pressure trajectories and the risk of stroke respectively by including the pre-specified covariates with a latent class-specific effect into the linear mixed sub-model and into the survival sub-model. As other examples, for the fourth objective, [14] search to prevent Alzeihmer disease using MMSE (*Mini-Mental State Examination*) score progression and creating a predictive risk model with class-specific covariates in both linear mixed sub-model and in the survival sub-model; Stamenic et al. [15] defined latent classes to assess the impact of serum creatinine on graft failure risk with no covariates in JLCM, and performed a multivariable multinomial logistic analysis after defining these latent classes in order to analyze the factors associated to the classes.

A few simulation studies concerning the JLCM and its extensions (competing risks, interval censoring, multistate survival sub-model) were carried out [6, 16–18]. However, these simulations focus on the model usability and aim at validating the estimation procedure rather than exploring the general properties of the model and its finite-sample properties.

Thus the usage of the model is heterogeneous and its properties in terms of sample size and censoring rate are not comprehensively studied.

In this context, the objective of this paper is to empirically, by a simulation study, explore the asymptotic properties of the JLCM model, namely, the impact of the censoring rate and of the number of individuals on bias and normality of parameter estimates as well as on the quality of latent class identification. A real data application will also be carried out. Within this application, the impact of covariates omission and inclusion in the model on estimations and class membership prediction will be investigated.

Methods

Joint latent class model

The joint latent class model is composed of three sub-models: a multinomial logistic regression defining the probability of belonging to a latent class, a mixed linear model for each latent class describing the evolution of the longitudinal marker, and a survival model accounting for the time-to-event for each class. The sub-models are detailed as follows.

- **The multinomial logistic regression** is defined by $\pi_i g$, the probability of individual i to belong to a given latent class g, conditional on a covariate vector $\boldsymbol{X}_i$:

$$\pi_{ig} = P(c_i = g|\boldsymbol{X}_i) = \frac{e^{\xi_{0g}+\boldsymbol{X}_i^T\boldsymbol{\xi}_{1g}}}{\sum_{l=1}^{G} e^{\xi_{0l}+\boldsymbol{X}_i^T\boldsymbol{\xi}_{1l}}}, \quad (1)$$

  where c_i is the latent class for patient i, $c_i \in (1, \cdots, G)$, $\boldsymbol{X}_i^T$ is a vector of explanatory variables for i necessarily independent of time, $\boldsymbol{\xi}_{1g}$ the vector of coefficients associated to the covariates effects within class g. Note that $\boldsymbol{\xi}_{0G} = 0$ and $\boldsymbol{\xi}_{1G} = 0$ to assure the model identifiability. If no prior information about the latent class is available, it is possible to use the marginal probability of the class g, $\frac{e^{\boldsymbol{\xi}_{0g}}}{\sum_{l=1}^{G} e^{\boldsymbol{\xi}_{0l}}}$ in Eq. (1).
- **The mixed linear model** for a trajectory of a longitudinal marker of an individual i over time points t_{ij}, Y_{ij} in a latent class g is defined as:

$$Y_{ij}|(c_i = g) = \boldsymbol{X}_{1ij}^T\boldsymbol{\gamma} + \boldsymbol{X}_{2ij}^T\boldsymbol{\beta}_g + \boldsymbol{Z}_{ij}^T\boldsymbol{b}_{ig} + \epsilon_{ij}, \quad (2)$$

  where $\boldsymbol{X}_{1ij}^T$ is the vector of explanatory variables common to all latent classes and $\boldsymbol{\gamma}$ the corresponding vector of coefficients , $\boldsymbol{X}_{2ij}^T$ is the vector of class-specific explanatory variables with $\boldsymbol{\beta}_g$ the corresponding vector of coefficients, and $\boldsymbol{Z}_{ij}$ is the vector of explanatory variables associated with the random effects $\boldsymbol{b}_{ig} \sim \mathcal{N}(\boldsymbol{\mu}_g, \boldsymbol{B}_g)$ ($\boldsymbol{\mu}_g$ is a mean of random effects, $\boldsymbol{B}_g$ is a variance-covariance matrix of random effects, both of which can be common or specific to latent classes). Note that $\boldsymbol{X}_{1ij}^T$ and $\boldsymbol{X}_{2ij}^T$ have no variables in common.
- **The survival model** for an individual i over time is defined by its hazard function, $\alpha_i(t)$, within each latent class as:

$$\alpha_i(t)|\left(c_i = g\right) = \alpha_0\left(t, \boldsymbol{\zeta}_g\right) \exp\left(\boldsymbol{X}_{1i}^T\boldsymbol{\vartheta} + \boldsymbol{X}_{2i}^T\boldsymbol{\eta}_g\right) \quad (3)$$

  with $\alpha_0(\cdot)$ the baseline risk function in latent class g, parametrized by vector $\boldsymbol{\zeta}_g$, $\boldsymbol{X}_{1i}^T$ is the vector of explanatory variables and $\boldsymbol{\vartheta}$ the associated parameters common to all latent classes, $\boldsymbol{X}_{2i}^T$ is the vector of class-specific explanatory variables and $\boldsymbol{\eta}_g$ the corresponding class-specific parameters of the model.

We denote by T_i the observed time to a clinical event of interest for individual i. In the framework of JLCM, it is important to note that the measures of the longitudinal marker after T_i, if there exist, are excluded from the observed data. Indeed, the objective is to describe the link between the risk of the event and the marker change over time preceding the event. The observed duration $T_i = \min(T_i^\star, C_i)$, where $T_i^\star$ corresponds to the real time-to-event (possibly not observed) and C_i corresponds to the right-censored duration. The survival function corresponding to the hazard of Eq. (3), is defined as:

$$S(t) = \exp\left(-\int_0^t \alpha(u)du\right) \quad (4)$$

Note that the individual covariate vectors $\boldsymbol{X}_i^T$ can be different in each of the three sub-models (Eqs. (1)-(3)), but have same notations for simplicity.

Likelihood

The parameters of the model can be estimated by the maximum likelihood method. The log-likelihood of the model defined for G latent classes is defined by Commenges and Jacqmin-Gadda [19] as:

$$L(\boldsymbol{\theta}_G) = \sum_{i=1}^{N} \log\Big(\sum_{g=1}^{G} \pi_{ig} f_{\boldsymbol{y}_i|c_i}(\boldsymbol{Y_i}|c_i = g) \alpha_i(T_i|c_i = g)^{\delta_i} S_i(T_i|c_i = g)\Big), \quad (5)$$

where π_{ig} is the probability of belonging to class g (Eq. (1)), $f_{\boldsymbol{y}_i|c_i}(\boldsymbol{Y_i}|c_i = g)$ is the probability density function of the longitudinal marker data in class g, defined in Eq. (2), $\alpha_i(T_i|c_i = g)$ is the hazard function defined in Eq. (3), $S_i(T_i|c_i = g)$ is the corresponding survival function. The event indicator δ_i for each individual is defined as:

$$\delta_i = \begin{cases} 1, \text{ if } T_i^\star < C_i. \\ 0, \text{ otherwise.} \end{cases} \quad (6)$$

The model parameters are estimated using the maximum likelihood estimator (MLE); the log-likelihood function is maximized by Newton-Raphson-like algorithm [20].

The optimal number of latent classes, G, is defined following Tofighi and Enders [21] by the BIC (Bayesian information criterion): the number of classes corresponding to the minimum value of BIC is preferred. However, the choice of G is also based on the number of patients per class and the concordance between the *a posteriori* classification derived from the model and expert opinion.

Class prediction and goodness-of-fit

Model goodness-of-fit can be assessed by a measure of class prediction accuracy. The class membership can be identified by computing the posterior probability of

belonging to a class g for each subject, based on the estimated model parameters. This probability is conditional on the observed covariate vector, i.e. the longitudinal data $\boldsymbol{Y}$ and the event times $\boldsymbol{T}$, and is defined in Eq. (7):

$$\begin{aligned}\pi_{ig}{}^{Y,T} &= P(c_i = g|\boldsymbol{Y_i}, T_i, \delta_i; \hat{\boldsymbol{\theta}}_G) \\ &= \frac{\hat{\pi}_{ig} f_{\boldsymbol{Y_i}|c_i}(\boldsymbol{Y_i}|c_i = g; \hat{\theta}_G)\alpha_i(T_i|c_i = g; \hat{\theta}_G)^{\delta_i} S_i(T_i|c_i = g; \hat{\theta}_G)}{\sum_{l=1}^{G} \hat{\pi}_{il} f_{\boldsymbol{Y_i}|c_i}(\boldsymbol{Y_i}|c_i = l; \hat{\theta}_G)\alpha_i(T_i|c_i = l; \hat{\theta}_G)^{\delta_i} S_i(T_i|c_i = l; \hat{\theta}_G)}.\end{aligned} \tag{7}$$

The subject i is assigned to a class g corresponding to the maximum estimated *a posteriori* probability π_{ig}.

Other approaches to goodness-of-fit can be employed, in particular those based on different types of residuals corresponding to different sub-models. These approaches will not be developed in the present paper.

Results

Simulation study

In the present study the properties of the JLCM are assessed by Monte-Carlo simulations. Simulations focus on the general model properties, on the model robustness to the number of individuals and the number of events, and on the quality of class separation.

The general framework for the simulation study is presented below.

Simulations design

The simulations are carried out for different settings in terms of the number of individuals n, $n = \{100, 500, 1000, 5000\}$, and in terms of the censoring rate τ, $\tau = \{0.05, 0.10, 0.15, 0.25, 0.50\}$, allowing to explore both possible asymptotic directions: the number of individuals and number of observed events [22]. The capacity of the model to distinguish between the latent classes is investigated by considering two different settings in terms of class separation: *high separation* (the classes are very different in terms of longitudinal marker evolution) and *low separation* (the classes are quite similar). The censoring mechanism was independent from the event process and no covariates were included in simulated models. Given the complex likelihood function, the optimisation algorithm may not always converge. That's why for each setting in terms of n, τ and class separation, 120 datasets were generated to assure obtaining at least 100 results in each setting. The distribution of each of the estimated parameters was then analyzed in terms of normality, relative bias and coverage rate. The normality was assessed graphically by quantile-quantile plots. Indeed, normality tests would often reject the null hypothesis due to outliers in parameter estimations (this situation is probable due to the likelihood complexity; it results in local maxima, but is rare in practice) and/or to high test power. The relative bias for a parameter θ is calculated as:

$$RB(\theta, n) = \left| \frac{\frac{1}{K}\sum_{h=1}^{K} \hat{\theta}_{n,h} - \theta}{\theta} \right|,$$

with $\frac{1}{K}\sum_{h=1}^{K} \hat{\theta}_{n,h}$ the average parameter estimation from the sample of n individuals over K Monte-Carlo runs, and θ the real parameter value. The absolute value will be considered.

The coverage rate was calculated for each model parameter as the percentage of coverage of the real value by the estimated confidence interval.

The capacity of the model to distinguish the latent classes is assessed by the percentage of correctly predicted class memberships.

Data generation

The real parameters were chosen to mimic the real data, described in Stamenic et al. paper [15], dealing with a prognostic tool for individualized prediction of graft failure risk within ten years after kidney transplantation, using serum creatinine progression as a longitudinal marker. Following Eqs. (1 - 4), the generated data were governed by the following general model:

$$\begin{cases} \pi_{i1} = Constant \\ \quad \text{for a 2-class model} \xi_{01} = \ln\left(\frac{\pi_{i1}}{1-\pi_{i1}}\right), \text{ see Eq. (1)} \\ Y_{ij}|(c_i = g) = \beta_{0g} + \beta_{1g} t_{ij} + b_{ig} + \epsilon_{ig} \\ \quad b_{ig} \sim \mathcal{N}\left(0, \sigma^2_{b,g}\right), \epsilon_{ig} \sim \mathcal{N}\left(0, \sigma^2_{\epsilon,g}\right) \\ S(t)|(c_i = g) = \exp\left(-\left(\frac{t}{\zeta_{1g}}\right)^{\zeta_{2g}}\right) \\ \quad T^{\star} \sim \mathcal{W}eibull\left(\zeta_{1g}, \zeta_{2g}\right) \\ M(t)|(c_i = g) = \exp\left(-\left(\frac{t}{\tilde{\zeta}_{1g}}\right)^{\tilde{\zeta}_{2g}}\right) \\ \quad C \sim \mathcal{W}eibull\left(\tilde{\zeta}_{1g}, \tilde{\zeta}_{2g}\right), \end{cases}$$

$M(t)$ being the survival function of the censoring distribution and C the censoring time. Note that the fact that there is no covariate in logistic model for class membership implies constant probability for each class membership. The considered longitudinal model is a random intercept mixed model and it implies that in Eq. (2), $\boldsymbol{X}_{1ij}^T$ is a zero matrix (no common covariates) and $\boldsymbol{X}_{2ij}^T = \begin{pmatrix} 1 & t_{ij} \end{pmatrix}$. The considered survival and censoring distributions imply that the survival and censoring times are Weibull random variables. The parameters of the censoring distribution were chosen empirically to meet the required censoring rate given the corresponding survival distribution. These nuisance parameters are not presented in the article.

The time points for repeated measures of the longitudinal marker are fixed to 1, 3, 6, 12, 18 and 24 months, following Stamenic et al. [15]. The parameters vector for

a 2-classes model, with class common random effect and error variance of mixed sub-model is as follows:

$$\boldsymbol{\theta} = \left(\xi_{01}, \beta_{01}, \beta_{11}, \beta_{02}, \beta_{12}, \sigma_b^2, \sigma_\epsilon^2, \zeta_{11}, \zeta_{21}, \zeta_{12}, \zeta_{22}\right). \quad (8)$$

The real values for the parameters were chosen as follows:

1 *High separation* framework.
This setting is directly derived from Stamenic et al. [15], resulted in $\beta_{01} = 170$, $\beta_{02} = 100$, $\beta_{11} = 88$ by year, $\beta_{12} = 1.2$ by year, $\sigma^2_{b,1} = \sigma^2_{b,2} = 50$ and $\sigma^2_{\epsilon,1} = \sigma^2_{\epsilon,2} = 60$, $\zeta_{11} = 4.5$, $\zeta_{21} = 2$, $\zeta_{12} = 50$, $\zeta_{22} = 1.01$.
2 *Low separation* framework.
In this setting the values of the mixed model from *high separation* are divided by 2 to obtain quiet similar classes in terms of longitudinal marker evolution; survival model as well as random parameters were not modified, resulting in $\beta_{01} = 135$, $\beta_{02} = 100$, $\beta_{11} = 44$ by year, $\beta_{12} = 1.2$ by year.

In both settings, the shape parameter for the Weibull distribution for censoring was fixed to 1.5, inspired from real life, where more censoring occurs with time. The scale parameter for this distribution was empirically derived to meet the required censoring rate. The probability of class 1 membership was set to 0.3 in both settings, resulting in the logistic model parameter from Eq. (1) $\xi_{01} = -0.84$. The examples of simulated trajectories for the *high separation* and *low separation* settings are illustrated in Fig. 1; the observed longitudinal trajectories are rather confounded in the *low separation* setting in comparison with the *high separation*.

Normality assessment

The normality of the estimated parameters is assessed by plotting quantile-quantile plots for each setting in terms of classes, the number of individuals n and of the censoring rate τ.

Figure 2 illustrates the results for the mixed and the survival sub-models, for 100 individuals, censoring rate 0.05 and 0.5 in the *high separation* setting. For small censoring rate (0.05) the normality of all the parameters is globally respected; heavy censoring (0.5) implies deviations from normality for the parameters of the survival sub-model.

Similar trends are observed for the other settings in terms of n and τ (results not presented). Note that the normality of the longitudinal sub-model parameters is not heavily impacted by small sample size and/or heavy censoring. Also, the MLE's normality is not considerably influenced by the degree of class separation according to the present simulation study (results not presented). However, this conclusion should be considered with caution, since it can be different for different separation degrees.

As expected, departures from normality decrease with increasing number of individuals (see Fig. 3 for the Weibull scale and shape parameters, heavy censoring) regardless of heavy censoring. Note that most of papers dealing with asymptotic properties of survival models are focused on the regression coefficients.Very few papers focus on the Weibull distribution parameters. Sirvanci and Yang [23] derives the asymptotic normality of the Weibull model parameters for Type I censoring data (fixed length of follow-up). However, in our study, empirically the departures from normality are reported for small sample size in terms of the number of events and/or the number of individuals (simulation results not presented here); in this sense, the normality problem is not specific to the joint latent class model, but is rather inherited from survival analysis.

Relative bias assessment

The relative bias (RB) of class-specific parameters estimates is illustrated in Figs. 4 and 5 for the *high separation* setting and in Figs. 6 and 7 for the *low separation* setting. The detailed numerical results are provided in Tables 1 and 2 for the *high* and *low separation* settings respectively.

The general trends for the RB range and for its evolution according to the sample size and to the censoring rate depend on model parameter and on degree of class separation. Concerning the variance parameters (the variance of error and of the random effect in the mixed sub-model) there is no clear trend in their RB evolution; the following trends are revealed for the remaining parameters:

- As for the **absolute values**, in the *high separation* setting (Fig. 6 and Table 1), the RB is the most important for two parameters of class 2: 1) the survival sub-model Weibull shape parameter (RB over 10% for small number of individuals) and 2) the mixed sub-model slope parameter (RB varies from 10% to 120% depending on number of individuals and on the censoring rate, the mean number of longitudinal markers in the worse case (100 patients and a censor of 50%) is 5.1). For the remaining parameters the RB does not exceed 10%. The trend is quite similar for the *low separation* setting (Fig. 6 and Table 2), but to a higher extent: the RB varies from over 30% to 530% in the worst setting (small n and high τ).
- As for the impact of the **censoring rate**, the RB increases linearly for a given number of individuals according to the decreasing number of events (increasing censoring rate). This trend is the same for both settings in terms of degree of class separation,

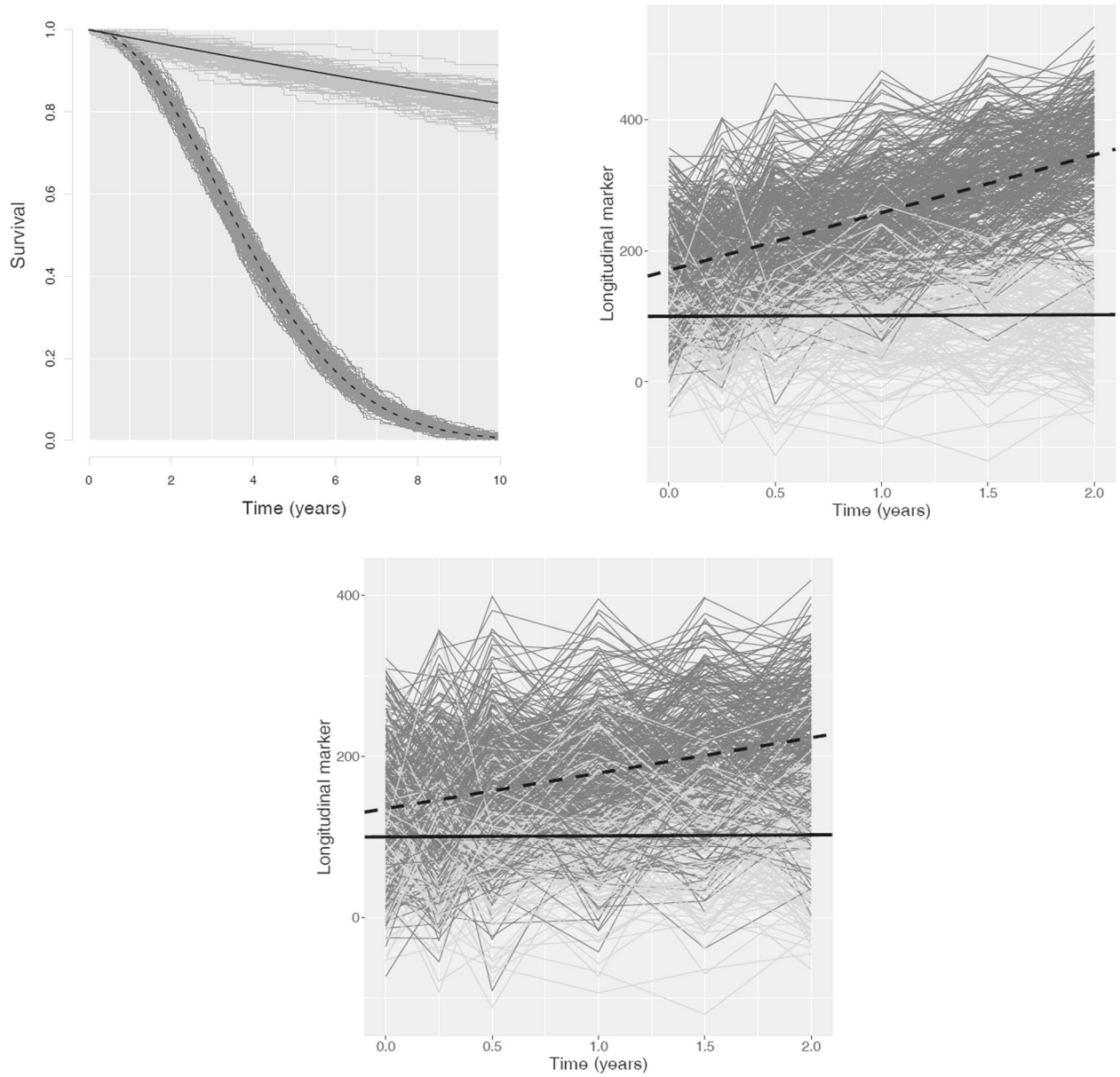


Fig. 1 Simulated survival curves and longitudinal marker trajectories that mimic the real data from Stamenic et al. [15]. The number of individuals $n = 500$; the censoring rate $\tau = 0.05$. Class 1: individual trajectories in darkgray, dashed line for mean trajectory; class 2: individual trajectories in lightgray, solid line for the mean trajectory. Figure at the top left: Generated survival curves for two classes and resulted examples of individual trajectories (same results for *high separation* and *low separation* settings). Figure on the top right: Simulated longitudinal marker evolution curves and the resulted examples of individual trajectories for the *high separation* setting. Bottom figure: Simulated longitudinal marker evolution curves and the resulted examples of individual trajectories for the *low separation* setting

but, in the same manner that the RB absolute values, in a higher extent for the *low separation* setting. Precisely, in the *high separation* setting the RB decreases by around 1% for the parameters of class 1 (2-8% in the *low separation* case) and for around 3-5% (2-15% in the *low separation* case) for the parameters of class 2, for the exception of the mixed model slope: 100% decrease in the RB in the *high separation* (respectively 400% in the *low separation* setting) for τ decreasing from 50% to 5%). Note that the linear trend for RB evolution in terms of τ is not always respected for small n.

- As for the impact of the **number of individuals**, the increasing n does not seem to strongly impact the RB. Moreover, the Weibull shape parameter is more influenced than the Weibull scale. Also, the *low*

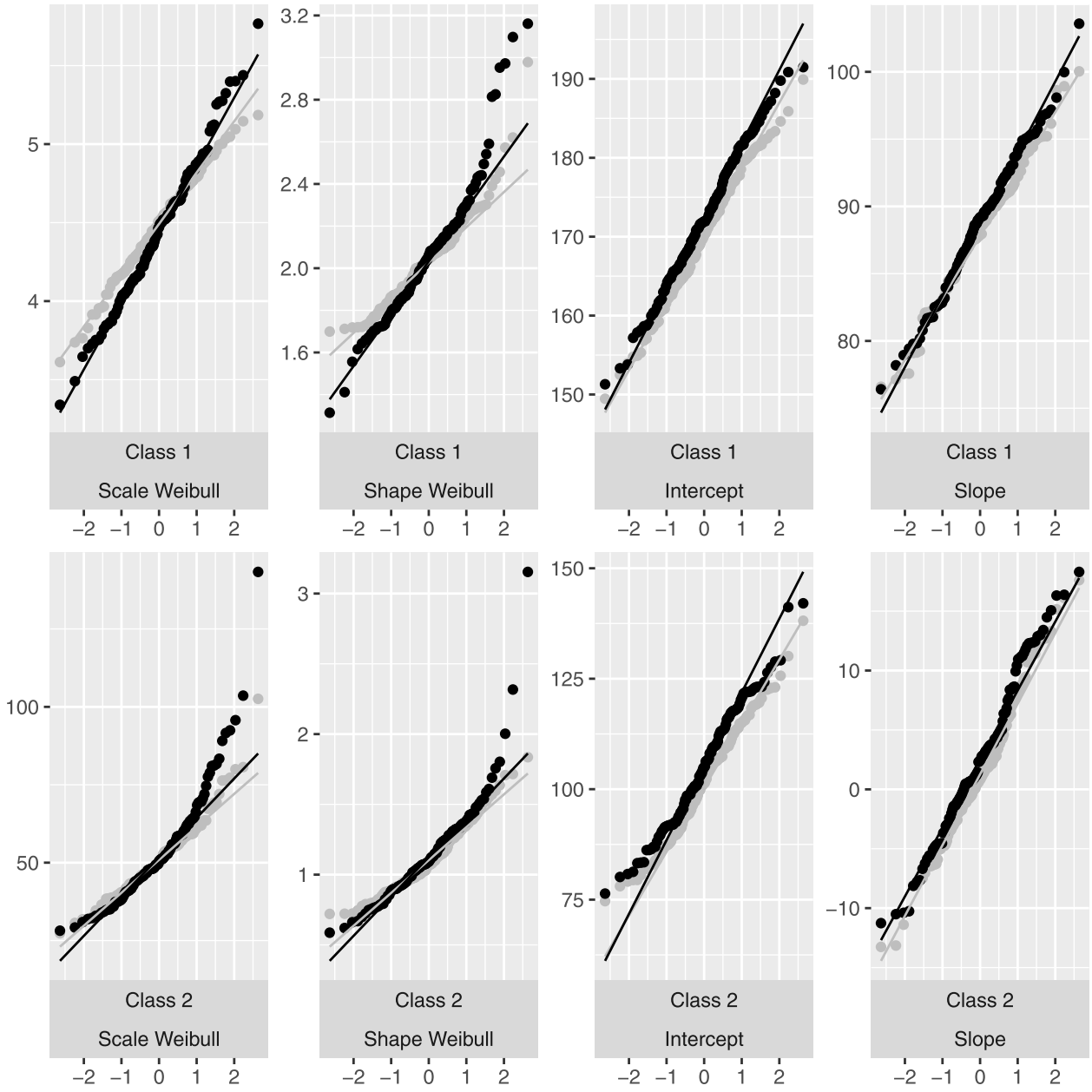


Fig. 2 Simulations results : quantile-quantile plot for the parameter estimations, *high separation* setting, $n = 100$. In black: censoring rate $\tau = 0.5$, in lightgray: $\tau = 0.05$. Results for Weibull scale, Weibull shape, mixed model intercept and mixed model slope, (ζ_{1g}, ζ_{2g}, β_{0g}, β_{1g} respectively form Eq. (8)) are presented. $g = \{1, 2\}$

separation setting is more influenced than the *high separation* setting.

Note that class 2 has the least number of patients with a lower risk of death; therefore the parameters of this class are more affected by the censoring rate. Also, the high bias for the class 2 slope parameter is explained by the small theoretical value for this parameter ($\beta_{12} = 1.2$).

Coverage rate assessment

The coverage rate is globally satisfactory (refer to Tables 3 and 4 for the 95% coverage rates in the *high* and the *low separation* settings respectively). However, the large sample size in terms of the number of individuals results in smaller confidence intervals, entailing lower empirical coverage rate. This trend is especially visible for heavy censoring. Departures from normality already

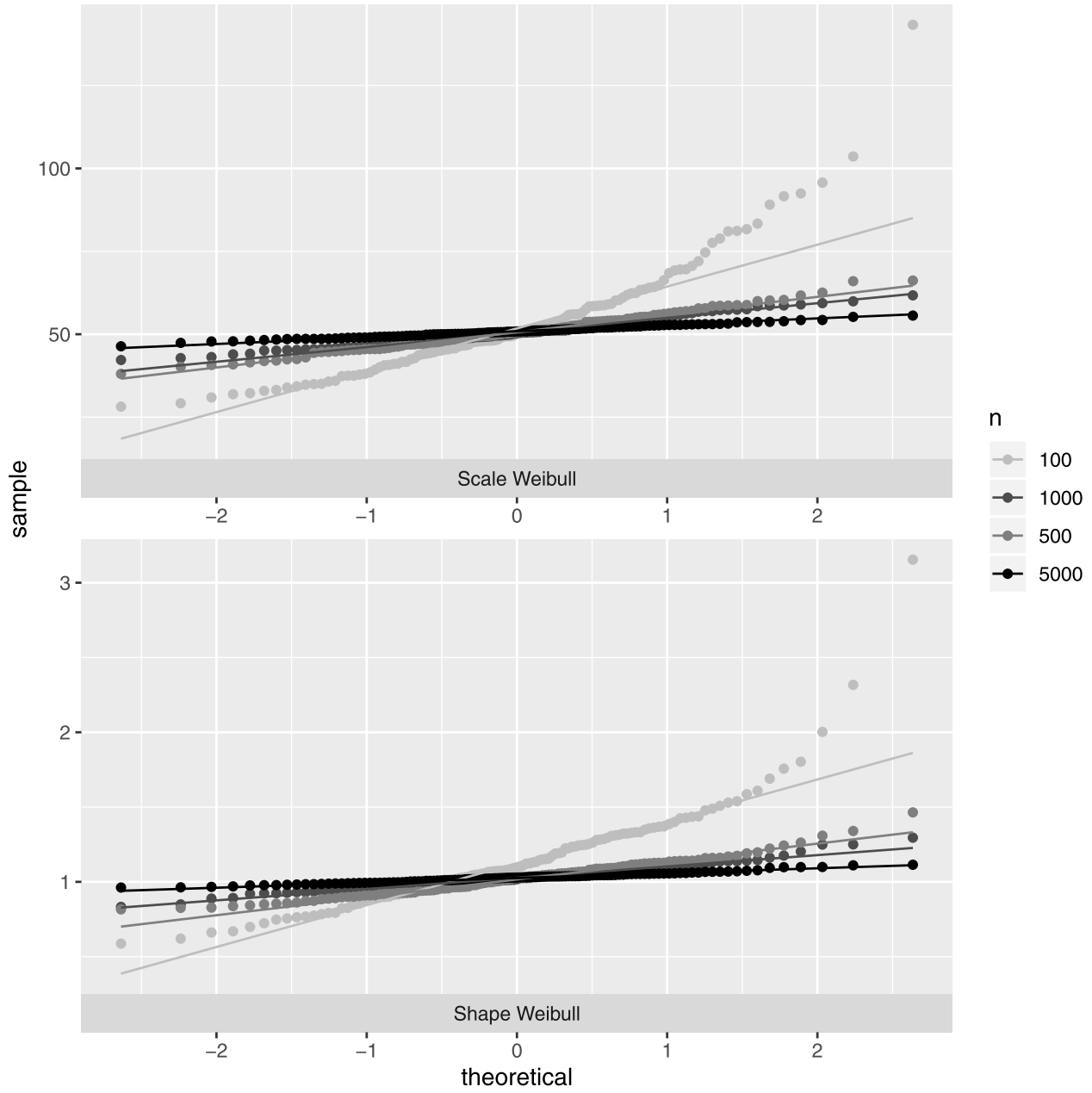


Fig. 3 Simulations results : quantile-quantile plot for the parameter estimations, *high separation* setting, censoring rate $\tau = 0.5$. Results for Weibull scale and Weibull shape parameters of class 2 (ζ_{12} and ζ_{22} respectively form Eq. (8)) according to number of individuals n are presented

mentioned for these settings can also be a cause of this phenomenon.

Class membership prediction assessment

The quality of the class membership prediction is globally satisfactory (Table 5): it is over 90% for the majority of settings in terms of n and τ. However, this quality is globally weaker for the *low separation* setting (less than 95% comparing to a rate higher than 95% for the *high separation* setting) and for heavy censoring (83-85% for the *low separation* setting, censoring rate 0.5). A decreasing censoring rate results in a 1% to 3% of the class identification improvement for all n, for the exception of heavy censoring τ. The sample size n does not considerably influence the quality of predictions, and in the *low separation* setting the prediction accuracy is around 3-6% weaker compared to the *high separation* setting, for the exception of heavy censoring cases.

The obtained simulations results can be summarized as follows: in general the MLE properties of the model

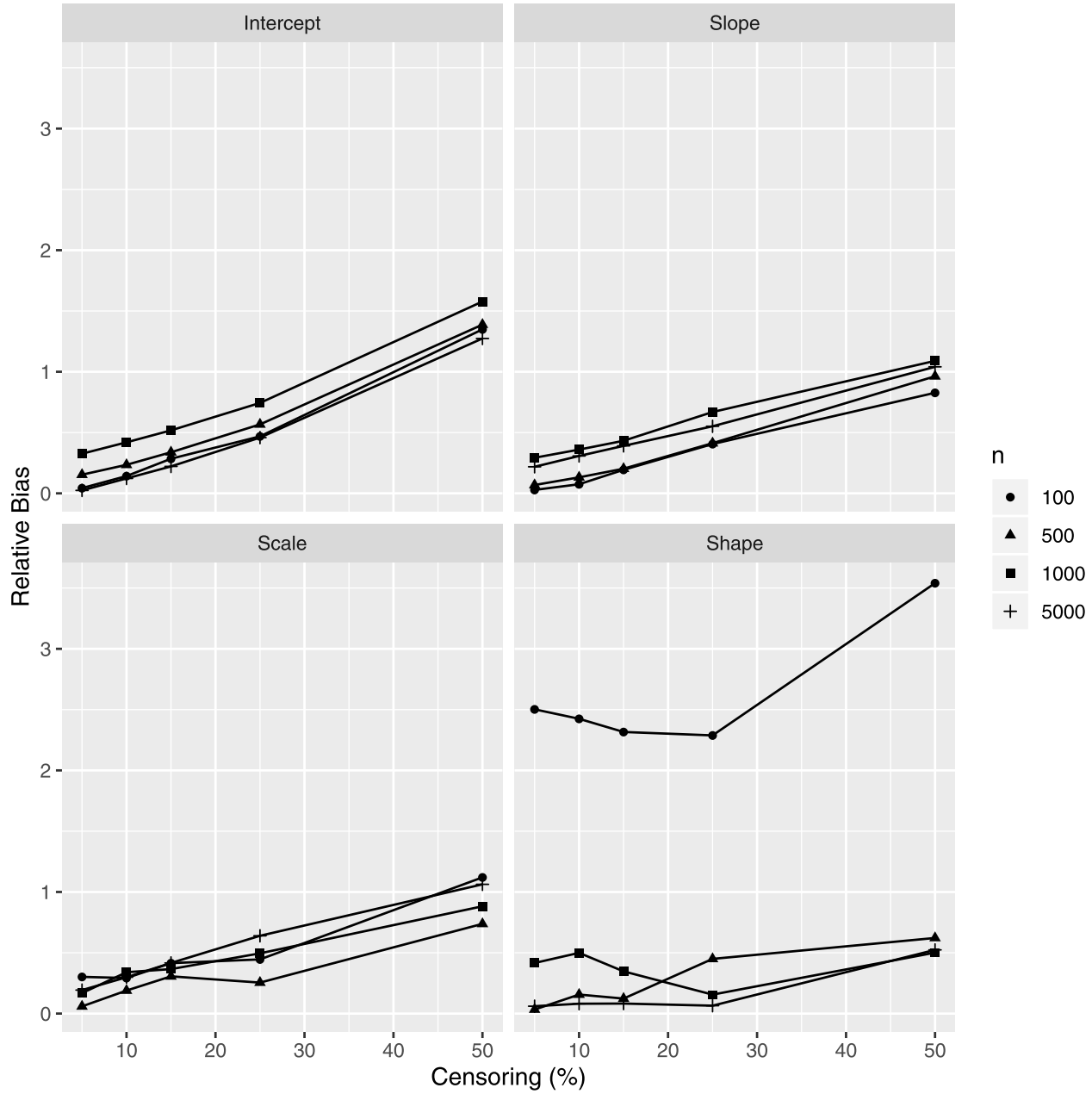


Fig. 4 Simulations results: relative bias of the class 1 parameter estimations according to the censoring rate τ and to the number of individuals n, *high separation* setting. Results for mixed model intercept, mixed model slope, Weibull scale and Weibull shape, (β_{01}, β_{11}, ζ_{11}, ζ_{21}, respectively form Eq. (8)) are presented. Same vertical scale is used for the four figures

parameters are impacted by the number of individuals as well as by the number of observed events and the number of longitudinal observations, which are both governed by the censoring rate. Note that the frequency of longitudinal marker observations also determines the number of observed measures, although this parameter is left fixed in the present study.

The quality of class membership identification depends on the number of observed events rather than on the number of observed individuals. The degree of class separation, determined by the class-specific slope of the longitudinal model, influences the bias and the normality of the MLE as well as the class identification accuracy. The assessment of the model properties was carried out after removing simulations with estimation convergence problems. The convergence problems are principally due to initial parameter values used in numerical estimation procedure. Such situations are quite rare : 1/120 (0.8%) for the

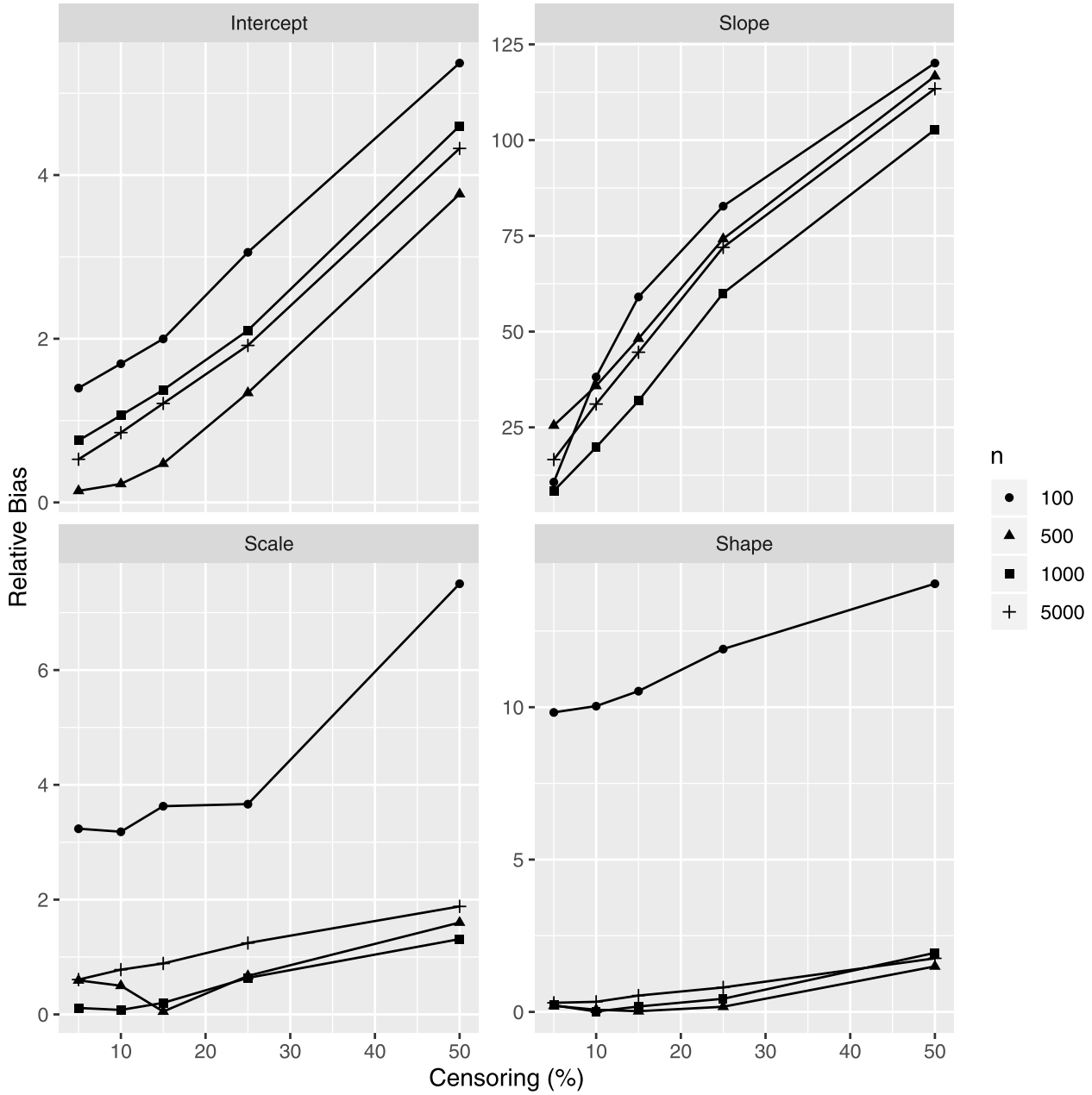


Fig. 5 Simulations results: relative bias of the class 2 parameter estimations according to the censoring rate τ and to the number of individuals n, *high separation* setting. Results for mixed model intercept, mixed model slope, Weibull scale and Weibull shape, (β_{02}, β_{12}, ζ_{12}, ζ_{22}, respectively form Eq. (8)) are presented. A specific vertical scale is used for each figure

setting $n = 100$ in *high separation* case and 9/120 times (7.5%) for the setting $n = 100$ in *low separation* case. Other settings were not impacted.

Real data application

In the present section, the analysis of the *Amyotrophic Lateral Sclerosis* (ALS) progression using a joint latent class model is presented.

ALS is a rapidly progressive and ultimately fatal neurodegenerative disease with an average life expectancy of 3–5 years from symptoms onset. However, longer than 10-years survival has been reported in 5–10% of patients [24, 25]. Despite numerous clinical trials dealing with treatments aimed at survival increase, only *riluzole* exhibited moderate efficacy [26]. One of the reasons which can explain the negative results of clinical trials is a strong

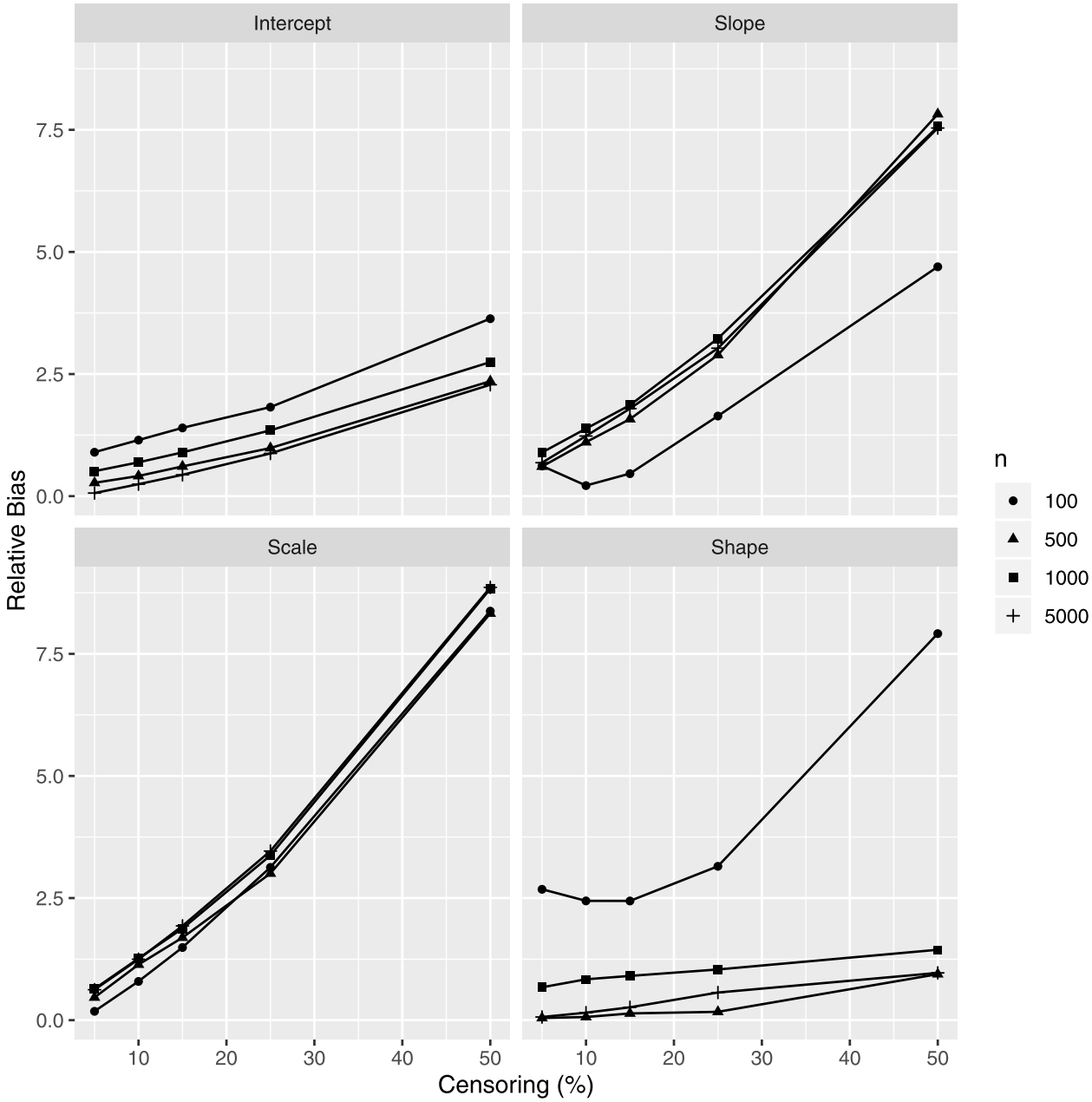


Fig. 6 Simulations results: relative bias of the class 1 parameter estimations according to the censoring rate τ and to the number of individuals *n*, *low separation* setting. Results for mixed model intercept, mixed model slope, Weibull scale and Weibull shape, (β_{01}, β_{11}, ζ_{11}, ζ_{21}, respectively form Eq. (8)) are presented. Same vertical scale is used for the four figures

heterogeneity of ALS patients in terms of the disease progression. The disease progression is generally measured at specific time points, resulting in a longitudinal marker. In this context, the joint latent class model, allowing to capture the patients heterogeneity and to simultaneously account for a longitudinal marker and a survival time, is better suited to analyze the ALS data.

The objective of our application is two-fold. Firstly, it is focused on capturing and describing the profiles of ALS patients in terms of the survival probability, the disease progression and clinical characteristics, described by covariates. Secondly, it aims at exploring the results in the light of model properties revealed by the simulation study.

Data collection

The data were collected in the framework of the *Trophos prospective cohort study* (TRO19622), a multicenter, randomized, placebo controlled, phase II/III clinical trial,

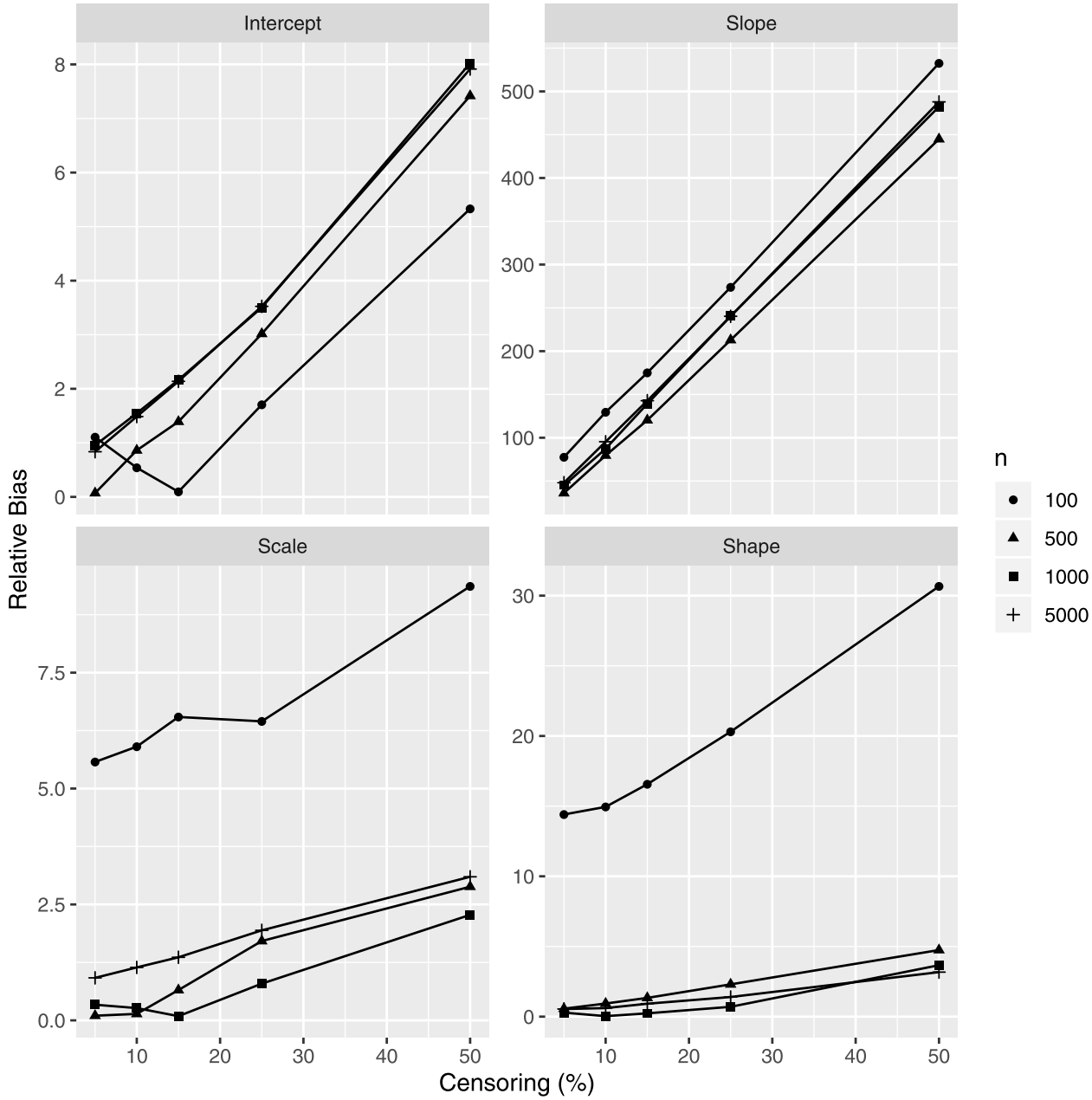


Fig. 7 Simulations results: relative bias of the class 2 parameter estimations according to the censoring rate τ and to the number of individuals n, *low separation* setting. Results for mixed model intercept, mixed model slope, Weibull scale and Weibull shape, (β_{02}, β_{12}, ζ_{12}, ζ_{22}, respectively form Eq. (8)) are presented. A specific vertical scale is used for each figure

which showed no efficacy of *olesoxime* in ALS [27]. The cohort consisted of 512 patients recruited across 15 European centres during the three-years period (2009–2011). The study time scale is the time since inclusion. The mean age of patients was 56 ($sd = 11.2$) years at inclusion and 55 ($sd = 11.2$) years at symptoms onset, with 331 (64.6%) men and 181 (35.4%) women. The diagnosis was definite in 107 patients (20.9%) and probable in 404 patients (79.1%) [28]; 101 (19.8%) patients suffered from bulbar form. The disease duration spanned between 6 and 36 months. Patients were treated with *50mg riluzole* twice a day for at least one month and had a baseline slow vital capacity (SVC) of 70%.

All patients were examined at inclusion and every 3 months thereafter for a maximum of 18 months for clinical, biochemical and hematological parameters. The disease-specific functional rating scale, revised ALSFRS (ALSFRS-R), was also assessed 1 month post-inclusion

Table 1 Simulations results: relative bias of model parameters for *high separation* setting according to the number of individuals, n, and to the censoring rate, τ

		Longitudinal sub-model						Survival sub-model			
n	τ	$\hat{\sigma}_b$	$\hat{\sigma}_\epsilon$	$\hat{\beta}_{0g}$		$\hat{\beta}_{1g}$		$\hat{\zeta}_{1g}$		$\hat{\zeta}_{2g}$	
				$g=1$	$g=2$	$g=1$	$g=2$	$g=1$	$g=2$	$g=1$	$g=2$
100	5	0.1053	21.1797	0.0429	1.3974	0.0281	10.7052	0.3018	3.2362	2.5023	9.8304
	10	0.0961	21.2395	0.1421	1.6954	0.0748	38.1642	0.2923	3.1830	2.4243	10.0367
	15	0.0997	21.3986	0.2846	1.9987	0.1936	59.0676	0.4148	3.6286	2.3160	10.5283
	25	0.0824	21.3810	0.4687	3.0567	0.4054	82.7586	0.4455	3.6646	2.2878	11.9084
	50	1.7125	21.8392	1.3486	5.3694	0.8271	120.1328	1.1201	7.5050	3.5401	14.0533
500	5	0.2267	21.4488	0.1534	0.1412	0.0685	25.4826	0.0591	0.5932	0.0332	0.2021
	10	0.2230	21.3820	0.2346	0.2261	0.1317	35.8174	0.1892	0.4996	0.1561	0.0727
	15	0.2062	21.4656	0.3380	0.4738	0.2027	48.1685	0.3062	0.0501	0.1232	0.0229
	25	0.2011	21.4412	0.5670	1.3390	0.4127	74.1988	0.2546	0.6707	0.4496	0.1703
	50	0.1945	21.6664	1.3876	3.7674	0.9627	116.6916	0.7377	1.6005	0.6217	1.4936
1000	5	0.0004	20.1935	0.3260	0.7563	0.2925	8.4122	0.1684	0.1125	0.4160	0.2180
	10	0.0015	20.2002	0.4197	1.0629	0.3599	19.7922	0.3399	0.0777	0.4994	0.0130
	15	0.0117	20.1928	0.5181	1.3744	0.4324	31.9018	0.3658	0.2015	0.3471	0.1781
	25	1.6848	20.2110	0.7436	2.1019	0.6690	60.0292	0.4954	0.6364	0.1559	0.4325
	50	0.0072	20.3934	1.5776	4.6004	1.0896	102.7405	0.8822	1.3124	0.5011	1.9393
5000	5	1.6342	19.9957	0.0242	0.5270	0.2185	16.5831	0.1929	0.6064	0.0605	0.2997
	10	1.6315	19.9945	0.1226	0.8526	0.3075	31.0770	0.2985	0.7797	0.0810	0.3320
	15	0.0380	19.9728	0.2215	1.2091	0.3905	44.6046	0.4162	0.8876	0.0823	0.5370
	25	0.0449	20.0112	0.4584	1.9183	0.5515	72.0024	0.6396	1.2441	0.0648	0.8037
	50	0.0400	20.2601	1.2738	4.3253	1.0406	113.4169	1.0629	1.8816	0.5238	1.7593

The estimations of the error and the random intercept standard deviations ($\hat{\sigma}_\epsilon$ and $\hat{\sigma}_b$ respectively), of the intercept and the slope ($\hat{\beta}_{0g}$, $\hat{\beta}_{1g}$ respectively) from the longitudinal sub-model and of Weibull scale and shape from the survival sub-model ($\hat{\zeta}_{1g}$ and $\hat{\zeta}_{2g}$ respectively) are presented. g: class membership identification

and then every 3 months until 18 months maximum. Survival time was defined as the duration between the date of disease onset and the date of a composite end-point: ALS-related death, tracheotomy, beginning of the non-invasive positive pressure ventilation (NIPPV) over 23 hours per day for 14 consecutive days or the date when last known to be alive.

Model construction

In terms of class identification, from 1 to 4 latent classes were considered. A quadratic trend for the longitudinal marker evolution was specified, and the corresponding mixed model was specific to each class, meaning that the quadratic terms were eliminated if not significantly different from 0, leading to a linear trend. The model performance in terms of class identification was assessed by the BIC.

To assess the impact of the sample size on parameter estimations, the estimations were carried out for the *whole sample* (512 patients) and for a subset of 100 randomly chosen patients. The results from the *reduced sample* appeared to be slightly different (results not presented here), reflecting the potential bias, revealed by the simulation study.

To better understanding of latent classes, modeling with and without covariates was performed. The covariates were included into the survival and the mixed sub-models, whereas the logistic regression, describing the probability of belonging to a class, was defined without covariates in all settings.

- **A model without covariates** (Eq. 9) includes a random-intercept mixed model with a class-specific quadratic function of time specified for the longitudinal marker evolution Y_{ij}; the variances of the random effect (σ_b^2) and of the error (σ_ϵ^2) were considered common to all classes. Survival curves are also considered as class-specific.
  The originally interval-censored survival times, collected at baseline and at months 1, 3, 6, 9, 12, 15 and 18, were imputed from a Weibull distribution of these interval-censored dates to obtain the exact event times. The imputation was carried out in order to obtain the setting close to that used in simulations.

Table 2 Simulations results: relative bias of model parameters for *low separation* setting according to the number of individuals, *n*, and to the censoring rate, τ

		Longitudinal sub-model						Survival sub-model			
n	τ	$\hat{\sigma}_b$	$\hat{\sigma}_\epsilon$	$\hat{\beta}_{0g}$		$\hat{\beta}_{1g}$		$\hat{\zeta}_{1g}$		$\hat{\zeta}_{2g}$	
				$g=1$	$g=2$	$g=1$	$g=2$	$g=1$	$g=2$	$g=1$	$g=2$
100	5	0.2157	22.3405	0.9001	1.1061	0.6187	77.4912	0.1821	5.5733	2.6803	14.4064
	10	3.3939	22.1855	1.1485	0.5399	0.2169	129.4999	0.7945	5.9039	2.4429	14.9468
	15	0.2315	22.2708	1.3985	0.0946	0.4619	175.0008	1.4886	6.5439	2.4417	16.5623
	25	1.5321	22.1504	1.8233	1.7052	1.6413	273.7677	3.1293	6.4496	3.1511	20.3027
	50	1.3699	21.7546	3.6344	5.3290	4.6976	532.4839	8.3763	9.3616	7.9156	30.6630
500	5	0.2556	21.5565	0.2726	0.0707	0.6062	36.0688	0.4649	0.1000	0.0440	0.5662
	10	1.9185	21.3569	0.4126	0.8612	1.1048	79.5267	1.1380	0.1392	0.0640	0.9334
	15	3.4729	21.2975	0.6091	1.3911	1.5773	120.2866	1.6905	0.6533	0.1375	1.3348
	25	0.1728	20.9193	0.9847	3.0168	2.8914	212.8892	3.0037	1.7108	0.1707	2.2996
	50	1.7796	19.8756	2.3534	7.4185	7.8224	444.8984	8.3268	2.8825	0.9417	4.7449
1000	5	0.0184	20.1206	0.5111	0.9550	0.8944	44.8903	0.6416	0.3375	0.6727	0.2864
	10	0.0302	20.0392	0.6927	1.5469	1.3819	87.2876	1.2630	0.2644	0.8368	0.0338
	15	0.0424	19.8918	0.8955	2.1650	1.8632	139.1406	1.8765	0.0905	0.9058	0.2346
	25	0.0763	19.6487	1.3465	3.5027	3.2279	240.5047	3.3728	0.7920	1.0379	0.6998
	50	0.2146	18.4371	2.7442	8.0212	7.5645	482.3023	8.8312	2.2757	1.4414	3.6641
5000	5	1.6360	19.8545	0.0630	0.8358	0.6849	48.0499	0.6231	0.9167	0.0643	0.5431
	10	0.0529	19.7483	0.2447	1.4861	1.2312	95.3766	1.2469	1.1418	0.1531	0.6085
	15	1.5950	19.6407	0.4366	2.1381	1.7922	142.7663	1.9299	1.3603	0.2632	0.9193
	25	0.1003	19.3803	0.8741	3.5248	3.0292	240.2600	3.4624	1.9415	0.5656	1.3986
	50	0.2294	18.2465	2.2807	7.9142	7.5368	487.8892	8.8603	3.0965	0.9679	3.1725

The estimations of the error and the random intercept standard deviations ($\hat{\sigma}_\epsilon$ and $\hat{\sigma}_b$ respectively), of the intercept and the slope ($\hat{\beta}_{0g}$, $\hat{\beta}_{1g}$ respectively) from the longitudinal sub-model and of Weibull scale and shape from the survival sub-model ($\hat{\zeta}_{1g}$ and $\hat{\zeta}_{2g}$ respectively) are presented. *g*: class membership identification

Specifically, a Weibull distribution was first fitted to the interval-censored dates, and then the exact event times were sampled from this distribution truncated by the limits of the observed intervals for each patient.

$$\begin{cases} \pi_{ig} = \frac{e^{\xi_{0g}}}{\sum_{l=1}^{G} e^{\xi_{0l}}}, \text{ from Eq. (1)} \\ \\ Y_{ij}|(c_i = g) = \beta_{0g} + \beta_{1g} t_{ij} + \beta_{2g} t_{ij}^2 + b_{0i} + b_{1i} t_{ij} \\ +b_{2i} t_{ij}^2 + \epsilon_i, \\ \quad \boldsymbol{b}_i \sim \mathcal{N}(0, \boldsymbol{B}), \epsilon_{ij} \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right), \text{ from Eq. (2)} \\ \\ S(t_i)|(c_i = g) = \exp\left(-\left(\frac{t_i}{\zeta_{1g}}\right)^{\zeta_{2g}}\right), \\ \quad T^\star \sim \mathit{Weibull}\left(\zeta_{1g}, \zeta_{2g}\right), \text{ from Eq. (4)} \end{cases} \tag{9}$$

with $\boldsymbol{B}$ covariance matrix of random effects.

- **A model with covariates** (Eq. 10, the hazard function is specified for easier interpretation) was specified based on clinical expertise and a preliminary unpublished study. This model includes baseline individual characteristics in the random intercept mixed sub-model and in the survival sub-model; the impact of these characteristics is specified common to all classes, following the clinical considerations. The quadratic term of time for the mixed sub-model appeared to be not significantly different from 0 for this model and is thus removed. Baseline covariates and their interactions with time were as well chosen from clinical expertise.
The following abbreviations are used: *AO* (Age at onset), *SO* (Symptom Onset), *BMI* (Body Mass Index), *MUSC* (Muscular capacity), *SVC* (Slow vital capacity), *MCV* (Mean corposcular volume).

Table 3 Simulations results: empirical coverage rates of estimated 95% confidence intervals according to number of individuals, n, and to the censoring rate, τ, *high separation* setting

n	τ	$\hat{\beta}_{0g}$		$\hat{\beta}_{1g}$		$\hat{\xi}_{1g}$		$\hat{\xi}_{2g}$	
		$g=1$	$g=2$	$g=1$	$g=2$	$g=1$	$g=2$	$g=1$	$g=2$
100	5	0.9664	0.9496	0.9328	0.9496	0.8824	0.8655	0.9580	0.9160
	10	0.9664	0.9496	0.9328	0.9412	0.8571	0.8655	0.9496	0.9076
	15	0.9664	0.9496	0.9412	0.9496	0.8824	0.8908	0.9412	0.9328
	25	0.9580	0.9496	0.9412	0.9580	0.8487	0.7899	0.9496	0.9076
	50	0.9328	0.9076	0.9748	0.9328	0.8403	0.7899	0.8824	0.8571
500	5	0.9667	0.9667	0.9833	0.9500	0.9250	0.9167	0.9333	0.9500
	10	0.9667	0.9750	0.9833	0.9417	0.9583	0.9250	0.9250	0.9500
	15	0.9667	0.9667	0.9750	0.9750	0.9250	0.9083	0.9417	0.9333
	25	0.9583	0.9500	0.9750	0.9417	0.8667	0.7750	0.9083	0.9167
	50	0.8750	0.9167	0.9333	0.9083	0.9000	0.8333	0.9333	0.9250
1000	5	0.9833	0.9333	0.9500	0.9250	0.8667	0.8750	0.9500	0.9417
	10	0.9750	0.9167	0.9500	0.9333	0.8833	0.9417	0.9583	0.9417
	15	0.9750	0.9167	0.9333	0.9333	0.8833	0.8917	0.9833	0.9583
	25	0.9500	0.9000	0.9167	0.9083	0.8917	0.8917	0.9417	0.9000
	50	0.8667	0.8000	0.8917	0.9083	0.9000	0.7000	0.9083	0.9167
5000	5	0.9750	0.9083	0.9333	0.8750	0.8917	0.9000	0.9667	0.9500
	10	0.9833	0.9083	0.9417	0.8583	0.9000	0.9250	0.9500	0.9417
	15	0.9667	0.8667	0.9083	0.8333	0.8250	0.8750	0.9417	0.9333
	25	0.9333	0.7917	0.9000	0.8250	0.8167	0.8667	0.8917	0.9333
	50	0.5000	0.2833	0.8000	0.7167	0.7750	0.7750	0.8500	0.9000

The results for the intercept and the slope from the longitudinal sub-model ($\hat{\beta}_{0g}$, $\hat{\beta}_{1g}$ respectively) and for the Weibull scale and shape from the survival sub-model ($\hat{\xi}_{1g}$ and $\hat{\xi}_{2g}$ respectively) are presented. g: class identification

$$
\begin{cases}
\pi_{ig} & = \frac{e^{\xi_{0g}}}{\sum_{l=1}^{G} e^{\xi_{0l}}}, \text{ from Eq. (1)} \\
Y_{ij}|(c_i=g) & = \beta_{0g} + \beta_{1g} t_{ij} + \gamma_1 SO_i + \gamma_2 BMI_i + \\
& \quad \gamma_3 MUSC_i + \gamma_4 SVC_i + \gamma_5 MCV_i + \\
& \quad t_{ij} \times (\gamma_6 SO_i + \gamma_7 MUSC_i + \gamma_8 SVC_i) + \\
& \quad b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}, \\
& \quad b_i \sim \mathcal{N}\left(0, \sigma_b^2\right), \\
& \quad \epsilon_{ij} \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right) \text{ from Eq. (2)} \\
\alpha_i(t)|(c_i=g) & = \underbrace{\zeta_{1g}^{\zeta_{2g}} \zeta_{2g} t^{\zeta_{2g}-1}}_{\alpha_0(t)} \exp(\vartheta_1 SO_i + \vartheta_2 BMI_i + \\
& \quad \vartheta_3 MUSC_i + \vartheta_4 SVC_i + \\
& \quad \vartheta_5 AO_i), \text{ from Eq. (3)}
\end{cases}
\tag{10}
$$

Real data analysis results

According to the BIC, 4 latent classes were retained for the model without covariates (BIC=15110 for 1 latent class, 14974 for 2 classes, 14911 for 3 latent classes and 14901 for 4 latent classes)and 2 latent classes for the model with covariates (BIC=14517 for 1 latent class, 14408 for 2 classes, 14410 for 3 latent classes and 14420 for 4 latent classes). Estimation results are presented in Table 6 and in Table 7 for the two models respectively. Models without and with covariates using the complete cases sample included 511 and 497 patients respectively. The difference in the number of patients is caused by missing covariates. Estimated survival curves and predicted ALSFRS evolution profiles are illustrated in Figs. 8 and 9 for the two considered models respectively.

The resulting latent classes are quite distinct both for the 4-classes no covariate model and for the 2-classes model including the covariates. The classes are characterized by a degree of ALSFRS decline and by the survival probability: a more rapid ALSFRS evolution is associated to a worse survival prognosis (refer to Figs. 8 and 9). In particular, the latent classes identified within the *model without covariates* can be interpreted in the following manner (refer to Fig. 8 for illustration).

- Classes 1 and 4 from the model without covariates are each composed of 5.1% of population. They represent patients with the most rapid decrease of ALSFRS and the highest risk of death, with a median

Table 4 Simulations results: empirical coverage rates of estimated 95% confidence intervals according to number of individuals, *n*, and to the censoring rate, τ, *low separation* setting

n	τ	$\hat{\beta}_{0g}$		$\hat{\beta}_{1g}$		$\hat{\zeta}_{1g}$		$\hat{\zeta}_{2g}$	
		$g=1$	$g=2$	$g=1$	$g=2$	$g=1$	$g=2$	$g=1$	$g=2$
100	5	0.9664	0.9496	0.9328	0.9496	0.8824	0.8655	0.9580	0.9160
	10	0.9664	0.9496	0.9328	0.9412	0.8571	0.8655	0.9496	0.9076
	15	0.9664	0.9496	0.9412	0.9496	0.8824	0.8908	0.9412	0.9328
	25	0.9580	0.9496	0.9412	0.9580	0.8487	0.7899	0.9496	0.9076
	50	0.9328	0.9076	0.9748	0.9328	0.8403	0.7899	0.8824	0.8571
500	5	0.9667	0.9667	0.9833	0.9500	0.9250	0.9167	0.9333	0.9500
	10	0.9667	0.9750	0.9833	0.9417	0.9583	0.9250	0.9250	0.9500
	15	0.9667	0.9667	0.9750	0.9750	0.9250	0.9083	0.9417	0.9333
	25	0.9583	0.9500	0.9750	0.9417	0.8667	0.7750	0.9083	0.9167
	50	0.8750	0.9167	0.9333	0.9083	0.9000	0.8333	0.9333	0.9250
1000	5	0.9833	0.9333	0.9500	0.9250	0.8667	0.8750	0.9500	0.9417
	10	0.9750	0.9167	0.9500	0.9333	0.8833	0.9417	0.9583	0.9417
	15	0.9750	0.9167	0.9333	0.9333	0.8833	0.8917	0.9833	0.9583
	25	0.9500	0.9000	0.9167	0.9083	0.8917	0.8917	0.9417	0.9000
	50	0.8667	0.8000	0.8917	0.9083	0.9000	0.7000	0.9083	0.9167
5000	5	0.9750	0.9083	0.9333	0.8750	0.8917	0.9000	0.9667	0.9500
	10	0.9833	0.9083	0.9417	0.8583	0.9000	0.9250	0.9500	0.9417
	15	0.9667	0.8667	0.9083	0.8333	0.8250	0.8750	0.9417	0.9333
	25	0.9333	0.7917	0.9000	0.8250	0.8167	0.8667	0.8917	0.9333
	50	0.5000	0.2833	0.8000	0.7167	0.7750	0.7750	0.8500	0.9000

The results for the intercept and the slope from the longitudinal sub-model ($\hat{\beta}_{0g}$, $\hat{\beta}_{1g}$ respectively) and for the Weibull scale and shape from the survival sub-model ($\hat{\zeta}_{1g}$ and $\hat{\zeta}_{2g}$ respectively) are presented. *g*: class identification

survival around 7 months and 14 months for class 1 and 4 respectively.

- Class 2 is the largest (68.5% of patients) and is characterized by the slowest evolution of ALSFRS and the highest survival rate (median survival over 20 months).
- Class 3 is composed of 21.3% of population and represents an "average" class with an ALSFRS progression similar to that in class 1 but with a lower baseline value: from Table 6 we observe the baseline value of 37 in class 3 *vs* 39 for class 2. The survival probability in class 3 is lower than that in class 2, with a median survival around 15 months.

The latent classes identified within the *model with covariates* can be interpreted in the following manner (refer to Fig. 9 for illustration).

- Class 1 is the largest (92.6% of patients), is characterized by a moderate ALSFRS progression (-2.3 point by months) and by a better survival prognosis (over 20 months median survival compared to around 8 months for class 2, for a patient with the average covariates vector).
- Class 2 is composed only of 37 patients (7.4%) and describes a specific patient profile, worsening and dying very quickly.

Note that after adjustment on the pre-specified factors from literature, known to be associated to ALSFRS progression and survival, two latent patient profiles are identified by the model, indicating a lack of explanatory capacity of these factors and motivating the use of the latent class model. This remaining latency in the model with covariates confirms the interest of using the JLCM to analyze this kind of data, and suggests a need for further clinical analysis of the disease progression.

Discussion

Several general considerations and recommendations concerning the use of the joint latent class model can be derived from the results of simulations.

To summarize, the departures from **normality** are particularly present for the survival sub-model parameters, and these departures disappear for a large enough number of observed events (small censoring rate) and/or large enough sample size (from 500 individuals normality is generally respected even for heavy censoring).

Table 5 Simulations results: class identification accuracy, calculated as the rate of correctly predicted class memberships, according to the number of individuals, *n*, and to the censoring rate, τ. The difference between the rates of the *high* and the *low separation* settings is provided

n	τ	*High separation*	*Low separation*	Difference
100	5	0.9760	0.9418	-0.0342
	10	0.9767	0.9347	-0.0420
	15	0.9748	0.9248	-0.0500
	25	0.9689	0.9039	-0.0650
	50	0.9556	0.8335	-0.1221
500	5	0.9790	0.9440	-0.0350
	10	0.9778	0.9376	-0.0402
	15	0.9764	0.9321	-0.0443
	25	0.9720	0.9148	-0.0572
	50	0.9586	0.8458	-0.1128
1000	5	0.9814	0.9477	-0.0337
	10	0.9798	0.9419	-0.0379
	15	0.9782	0.9354	-0.0428
	25	0.9745	0.9186	-0.0559
	50	0.9605	0.8488	-0.1017
5000	5	0.9817	0.9480	-0.0337
	10	0.9801	0.9417	-0.0384
	15	0.9784	0.9348	-0.0436
	25	0.9748	0.9189	-0.0559
	50	0.9618	0.8504	-0.1114

In terms of the **relative bias**, the trends are more complex. The parameters of the survival sub-model are also more impacted, especially for a small sample size *n*. The large number of individuals does not compensate for heavy censoring, as it was the case for normality. There is no particular trend in terms of *n*, except for the survival sub-model parameters, whose bias is considerably increased for $n = 100$. The bias decreases quasi linearly for almost all parameters with increasing number of observed events (decreasing censoring rate). The estimations in the *low separation* case are less robust to the sample size and to the censoring rate than in the *high separation* case.

Finally, the **class identification accuracy** is slightly higher for the *high separation* setting and for smaller censoring, but not considerably influenced by the number of individuals, except for the case of heavy censoring; in the *low separation* setting the class identification accuracy is quite poor.

In the light of the obtained results, several remarks can be formulated concerning the general model usability.

Concerning implementation, the *low separation* setting, i.e., the small difference in the longitudinal model slopes, the likelihood optimization procedure is more likely to converge to a local maxima. Thus, several estimations with different initial parameter values should be carried out to assure that the obtained estimation is the global maxima.

Concerning the general model properties, the following should be accounted for.

Table 6 Real data results: parameter estimates with standard errors and *p*-values from the four-latent classes model **without covariates**

number of observations		**2591**	
number of patients		**511**	
average number of longitudinal measure		**5**	
number of events		**132**	
censoring rate		**0.74**	
Sub-model	Parameter	Estimate (se)	*p*-value
Multinomial logistic regression	ξ_{01}	-0.29 (0.49)	0.55
	ξ_{02}	2.26 (0.44)	< 0.001
	ξ_{03}	1.21 (0.53)	0.022
Weibull survival model	ζ_{11}	0.37 (0.02)	$< .001$
	ζ_{21}	1.52 (0.13)	< 0.001
	ζ_{12}	0.12 (0.02)	< 0.001
	ζ_{22}	1.25 (0.14)	< 0.001
	ζ_{13}	0.24 (0.01)	< 0.001
	ζ_{23}	1.56 (0.12)	< 0.001
	ζ_{14}	0.26 (0.01)	< 0.001
	ζ_{24}	2.02 (0.31)	< 0.001
Linear mixed model : fixed effects	β_{01}	35.22 (1.11)	< 0.001
	β_{11}	-5.29 (0.32)	< 0.001
	β_{21}	0.31 (0.04)	< 0.001
	β_{02}	39.37 (0.34)	< 0.001
	β_{12}	-0.52 (0.06)	< 0.001
	β_{22}	-0.01 (0.00)	0.007
	β_{03}	37.44 (0.66)	< 0.001
	β_{13}	-1.92 (0.16)	< 0.001
	β_{23}	0.04 (0.01)	< 0.001
	β_{04}	39.00 (1.30)	< 0.001
	β_{14}	-0.36 (0.20)	< 0.001
	β_{24}	-0.10 (0.02)	< 0.001
Linear mixed model : random effects	$\sigma^2_{b_0}$	22.93	
	$\sigma^2_{b_1}$	0.20	
	$\sigma^2_{b_2}$	0.00	
	$\sigma^2_{\epsilon,1}$	1.67	

Table 7 Real data results: parameter estimates with standard errors and *p*-values from the two-latent classes model **with covariates**

number of observations		**2525**	
number of patients		**497**	
average number of longitudinal measure		**5**	
number of events		**129**	
censoring rate		**0.74**	
Sub-model	Parameter	Estimate (se)	*p*-value
Multinomial logistic regression	ξ_{01}	2.22 (0.31)	< 0.001
Weibull model	ζ_{11}	0.48 (0.17)	0.004
	ζ_{21}	1.48 (0.08)	< 0.001
	ζ_{12}	0.68 (0.21)	0.001
	ζ_{22}	1.64 (0.12)	< 0.001
	ϑ_1	-0.05 (0.01)	0.008
	ϑ_2	-0.05 (0.03)	0.079
	ϑ_3	-0.03 (0.01)	< 0.001
	ϑ_4	-0.41 (0.12)	< 0.001
	ϑ_5	0.04 (0.01)	< 0.001
Linear mixed model : fixed effects	$\hat{\beta}_{01}$	9.79 (4.02)	0.015
	β_{11}	-2.32 (0.27)	< 0.001
	β_{02}	7.83 (4.12)	0.057
	β_{12}	-4.06 (0.39)	< 0.001
	γ_1 (SO)	-0.06 (0.02)	< 0.001
	γ_2 (BMI)	-0.13 (0.05)	0.009
	γ_3 (MUSC)	0.16 (0.00)	< 0.001
	γ_4 (SVC)	1.04 (0.18)	< 0.001
	γ_5 (MCV)	0.10 (0.04)	0.007 1
	γ_6 (SO$\times t_j$)	0.02 (0.00)	< 0.001
	γ_7 (MUSC$\times t_j$)	0.01 (0.00)	< 0.001
	γ_8 (SVC$\times t_j$)	0.06 (0.02)	0.018
Linear mixed model : random effects	$\sigma^2_{b_0}$	14.10 (0.00)	
	σ^2_{b1}	0.18	
	$\sigma^2_{\epsilon,1}$	1.97	

Note: the following covariates and their interactions with time (if significant) are presented: SO (Symptom Onset), BMI (Body Mass Index), MUSC (Muscular capacity), SVC (Slow vital capacity), MCV (Mean corposcular volume)

1 Small sample size in terms of the number of individuals results in deviations from normality, especially for the survival model parameters. The provided confidence intervals may not be valid.
2 Heavy censoring implies bias in parameter estimation, especially in case of weak separation between latent classes. This bias is not compensated by large sample size.
3 Heavy censoring gives poor class identification accuracy, especially in the case of weak separation between latent classes.
4 The model parameters are generally more sensible to censoring rate than to the number of individuals in terms of bias, thus, increasing the time of observation is more beneficial for the accuracy of estimates than increasing the sample size in terms of the number of individuals.
5 In case of poor separation between latent classes, the bias increases and the class predictions accuracy decreases, the results should be interpreted with caution.
6 Small latent groups with few events (heavy censoring) should be characterized with caution, since the parameter estimations can be considerably biased.

As for the real data application results, using the joint latent class model for the described data is beneficial. Indeed, the latency remains in data after adjustment on covariates known from clinical expertise. Note however that the observed ALSFRS profiles are rather distinguished, i.e. the observed data are close to the *high separation* setting, implying better general results. As shown by simulations, in case of lower separation, it could be more difficult to obtain and interpret the latent classes. Moreover, the results obtained from the *whole* and *reduced* samples differ (results not presented). Thus, care should be taken when interpreting the parameters derived from small samples due to possible bias and inference problems resulting from departures from normality.

In the present paper, we focus on JLCM as the approach to account for unobserved heterogeneity when modelling censored longitudinal outcomes. Other alternatives to this approach exist as the mixed latent Markov models proposed by Bartolucci et al.

Conclusion

The JLCM properties have been evaluated. We have illustrated the discovery in practice and highlights the usefulness of the joint models with latent classes in this kind of data even with pre-specified factors. We made some recommendations for the use of this model and for future research. Further work is needed to assess the role of covariates, their place in different sub-models of the JLCM, and the impact of their omission on parameter estimations and class membership identification. Also, precise recommendations concerning a minimum number of events or individuals needed to obtain satisfactory results within the JLCM can be formulated. Impact of longitudinal observation frequency on parameter estimations and latent classes identification can also be study considered in further work.

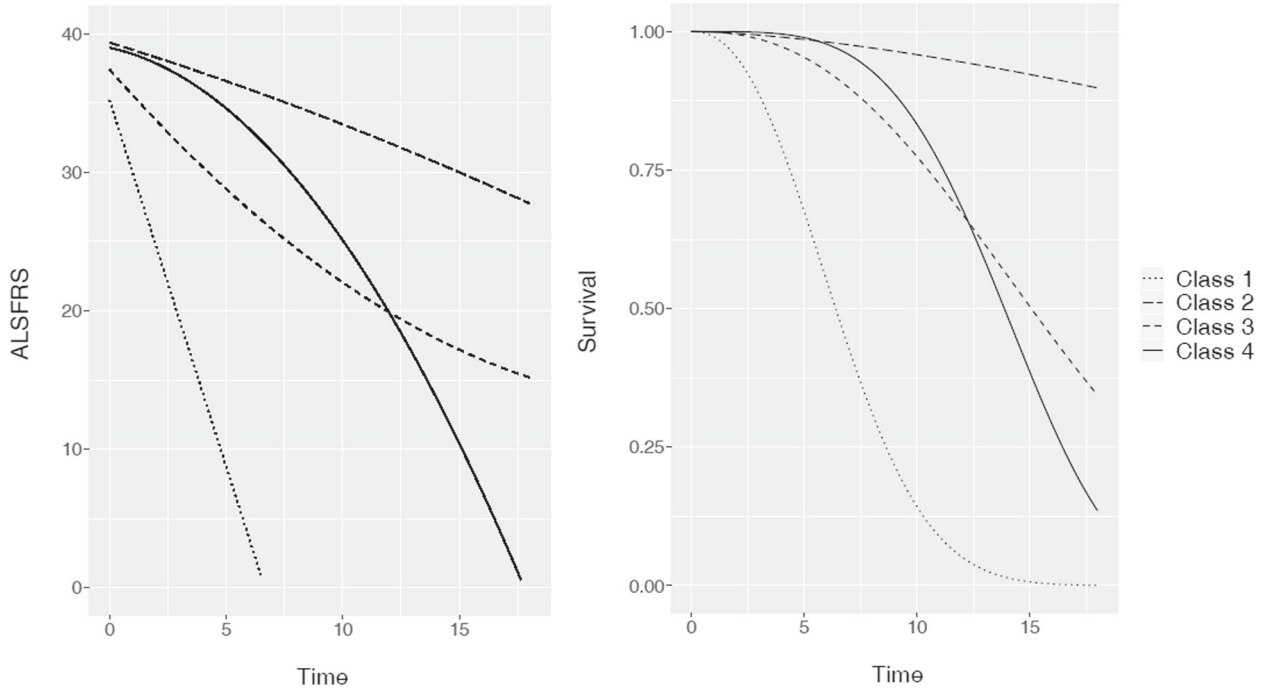


Fig. 8 Real data results: class-specific estimations of the mean ALSFRS evolution and of the survival probability for the model without covariates

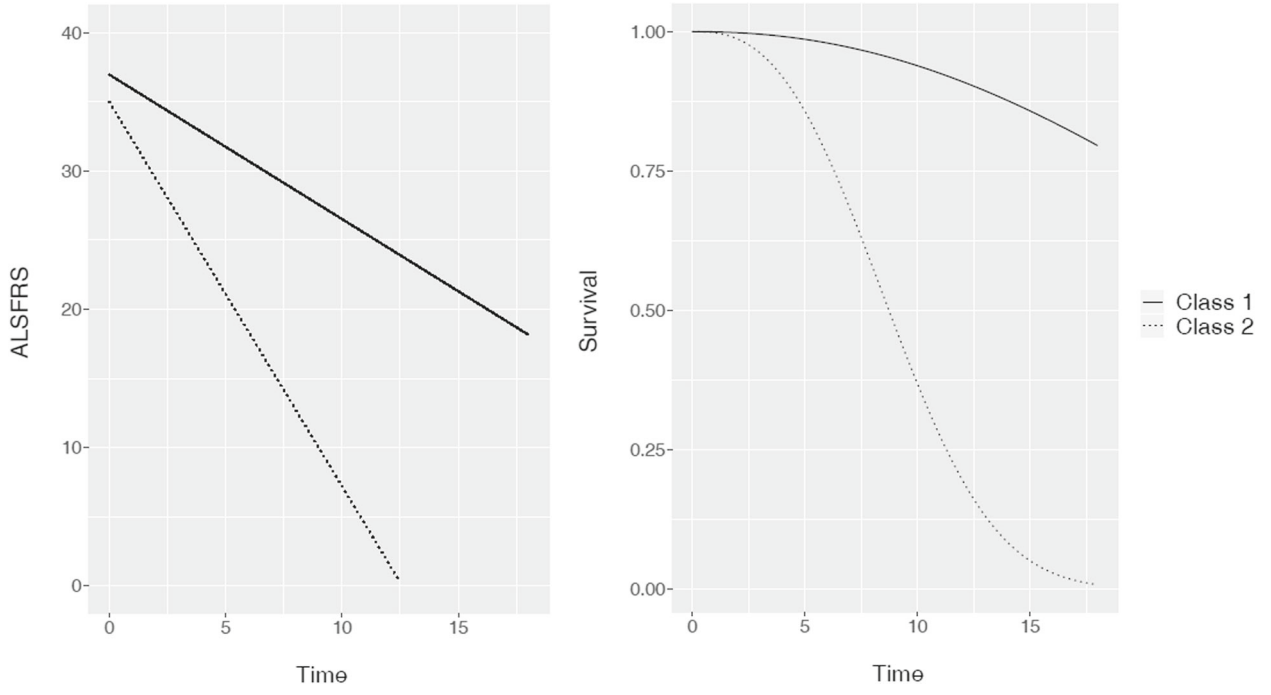


Fig. 9 Real data results: class-specific estimations of the ALSFRS evolution and of the survival probability for a model with covariates. The mean values of explanatory variables were used for illustration

Abbreviations
JLCM: Joint latent class model; MLE: Maximum likelihood estimations; GMM: growth mixture model; MMSE: Mini-mental state examination; ALS: amyotrophic lateral sclerosis; SVC: slow vital capacity; ALSFRS-r: amyotrophic lateral sclerosis functional rating scale revised, NIPPV non-invasive positive pressure ventilation; BIC: bayesian information criterion

Acknowledgements
The authors wish to acknowledge support from the ARSLA charity (Association pour la Recherche sur la Sclérose Latérale Amyotrophique et autres maladies du motoneurones). We thank Valerie Cuvier, Pierre-François Pradat, Vincent Meininger for the of data of the Trophos prospective cohort study (TRO19622). The study has been funded by ARLSA charity.

Authors' contributions
MK, AD, GB, JL participated in the design and conduct of the study. MK and GB writing the manuscript. MK, GB performed the statistical analysis and revised the manuscript. CT helped the statistical simulation and analysis. DD helped in the design of application method. MK, AD, GB, JL, CT and DD conceived of the design and coordination of the study and helped revising the manuscript. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The TROPHOS dataset analysed during the current study is publically available from https://pubmed.ncbi.nlm.nih.gov/27713255/

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Author details
[1]ULR 2694 - METRICS : évaluation des technologies de santé et des pratiques médicales, Univ. Lille, CHU Lille, Lille, France. [2]Département de Biostatistiques, CHU Lille, Lille, France. [3]Expert center for ALS, Expert center for Parkinson, Medical Pharmacology, Univ. Lille, Lille Neuroscience & Cognition, Inserm, UMR-S1172, Lille, France.

Received: 7 April 2021 Accepted: 20 August 2021
Published online: 30 September 2021

References

1. Rizopoulos D. Joint models for longitudinal and time-to-event data: With applications in R. London: Chapman & Hall; 2012.
2. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. Biometrics. 1997;53:330–9.
3. Verbeke G, Lesaffre E. A linear mixed-effects model with heterogeneity in the random-effects population. J Am Stat Assoc. 1996;91(433):217–21.
4. Muthén B, Shedden K. Finite mixture modeling with mixture outcomes using the em algorithm. Biometrics. 1999;55(2):463–9.
5. Lin H, Turnbull BW, McCulloch CE, Slate EH. Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. J Am Stat Assoc. 2002;97(457):53–65.
6. Proust-Lima C, Dartigues J-F, Jacqmin-Gadda H. Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: a latent process and latent class approach. Stat Med. 2016;35(3):382–98.
7. Proust-Lima C, Séne M, Taylor JM, Jacqmin-Gadda H. Joint latent class models for longitudinal and time-to-event data: A review. Stat Methods Med Res. 2014;23(1):74–90.
8. Syrjälä E, Nevalainen J, Peltonen J, Takkinen H-M, Hakola L, Åkerlund M, Veijola R, Ilonen J, Toppari J, Knip M, Virtanen SM. A joint modeling approach for childhood meat, fish and egg consumption and the risk of advanced islet autoimmunity. Sci Rep. 2019;9(1):7760. https://doi.org/10.1038/s41598-019-44196-1. Accessed 02 Jul 2019.
9. Brilleman SL, Moreno-Betancur M, Polkinghorne KR, McDonald SP, Crowther MJ, Thomson J, Wolfe R. Changes in body mass index and rates of death and transplant in hemodialysis patients: a latent class joint modeling approach. Epidemiology. 2019;30(1):38–47. https://doi.org/10.1097/EDE.0000000000000931. Accessed 10 May 2019.
10. Ogata S, Watanabe M, Kokubo Y, Higashiyama A, Nakao YM, Takegami M, Nishimura K, Nakai M, Kiyoshige E, Hosoda K, Okamura T, Miyamoto Y. Longitudinal trajectories of fasting plasma glucose and risks of cardiovascular diseases in middle age to elderly people within the general Japanese population: the Suita Study. J Am Heart Assoc. 2019;8(3):010628. https://doi.org/10.1161/JAHA.118.010628.
11. Portegies MLP, Mirza SS, Verlinden VJA, Hofman A, Koudstaal PJ, Swanson SA, Ikram MA. Mid-to late-life trajectories of blood pressure and the risk of stroke: the Rotterdam Study. Hypertension (Dallas, Tex.: 1979). 2016;67(6):1126–32. https://doi.org/10.1161/HYPERTENSIONAHA.116.07098.
12. Jiang G, Luk AOY, Tam CHT, Xie F, Carstensen B, Lau ESH, Lim CKP, Lee HM, Ng ACW, Ng MCY, Ozaki R, Kong APS, Chow CC, Yang X, Lan H-Y, Tsui SKW, Fan X, Szeto CC, So WY, Chan JCN, Ma RCW, Hong Kong Diabetes Register TRS Study Group. Progression of diabetic kidney disease and trajectory of kidney function decline in Chinese patients with Type 2 diabetes. Kidney Int. 2019;95(1):178–87. https://doi.org/10.1016/j.kint.2018.08.026.
13. Marioni RE, Proust-Lima C, Amieva H, Brayne C, Matthews FE, Dartigues J-F, Jacqmin-Gadda H. Cognitive lifestyle jointly predicts longitudinal cognitive decline and mortality risk. Eur J Epidemiol. 2014;29(3):211–9. https://doi.org/10.1007/s10654-014-9881-8.
14. Qin Y, Tian Y, Han H, Liu L, Ge X, Xue H, Wang T, Zhou L, Liang R, Yu H. Risk classification for conversion from mild cognitive impairment to Alzheimer's disease in primary care. Psychiatry Res. 2019;278:19–26. https://doi.org/10.1016/j.psychres.2019.05.027. Accessed 02 Jul 2019.
15. Stamenic D, Rousseau A, Essig M, Gatault P, Buchler M, Filloux M, Marquet P, Prémaud A. A prognostic tool for individualized prediction of graft failure risk within ten years after kidney transplantation. J Transplant. 2019;2019:7245142. https://doi.org/10.1155/2019/7245142.
16. Proust-Lima C, Joly P, Dartigues J-F, Jacqmin-Gadda H. Joint modelling of multivariate longitudinal outcomes and a time-to-event: a nonlinear latent class approach. Comput Stat Data Anal. 2009;53(4):1142–54.
17. 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. 2016;35(22):3933–48.
18. Rouanet A, Joly P, Dartigues J-F, Proust-Lima C, Jacqmin-Gadda H. Joint latent class model for longitudinal data and interval-censored semi-competing events: Application to dementia. Biometrics. 2016;72(4):1123–35.
19. Commenges D, Jacqmin-Gadda H. Modèles Biostatistiques Pour L'épidémiologie. France: De Boeck Superieur; 2015.
20. Proust C, Jacqmin-Gadda H. Estimation of linear mixed models with a mixture of distribution for the random effects. Comput Methods Prog Biomed. 2005;78(2):165–73.
21. Tofighi D, Enders CK. Identifying the correct number of classes in growth mixture models. Adv Latent Variable Mixture Model. 2008;2007:317–41.
22. Babykina G, Couallier V. Empirical assessment of the maximum likelihood estimator quality in a parametric counting process model for recurrent events. Comput Stat Data Anal. 2012;56(2):297–315.
23. Sirvanci M, Yang G. Estimation of the weibull parameters under type i censoring. J Am Stat Assoc. 1984;79(385):183–7.
24. Yates E, Rafiq M. Prognostic factors for survival in patients with amyotrophic lateral sclerosis: analysis of a multi-centre clinical trial. J Clin Neurosci. 2016;32:51–6.
25. Chio A, Logroscino G, Hardiman O, Swingler R, Mitchell D, Beghi E, Traynor BG, Consortium E, et al. Prognostic factors in ALS: a critical review. Amyotroph Lateral Scler. 2009;10(5-6):310–23.
26. Zinman L, Cudkowicz M. Emerging targets and treatments in amyotrophic lateral sclerosis. Lancet Neurol. 2011;10(5):481–90.

27. Lenglet T, Lacomblez L, Abitbol J, Ludolph A, Mora J, Robberecht W, Shaw P, Pruss R, Cuvier V, Meininger V, et al. A phase II- III trial of olesoxime in subjects with amyotrophic lateral sclerosis. Eur J Neurol. 2014;21(3):529–36.
28. Brooks BR, Miller RG, Swash M, Munsat TL. El escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis. Amyotroph Lateral Scler Other Motor Neuron Disorders. 2000;1(5):293–9.

Publisher's Note

Bibliographie

[1] Molina JM, Chêne G, Ferchal F, Journot V, Pellegrin I, Sombardier MN, et al. The ALBI trial : a randomized controlled trial comparing stavudine plus didanosine with zidovudine plus lamivudine and a regimen alternating both combinations in previously untreated patients infected with human immunodeficiency virus. The Journal of infectious diseases. 1999 ;180(2) :351-8.

[2] Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982 :963-74.

[3] Commenges D, Jacqmin-Gadda H. Modèles biostatistiques pour l'épidémiologie. De Boeck Superieur ; 2015.

[4] Verbeke G, Lesaffre E. A linear mixed-effects model with heterogeneity in the random-effects population. Journal of the American Statistical Association. 1996 ;91(433) :217-21.

[5] Muthén B, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics. 1999 ;55(2) :463-9.

[6] Proust C, Jacqmin-Gadda H. Estimation of linear mixed models with a mixture of distribution for the random effects. Computer methods and programs in biomedicine. 2005 ;78(2) :165-73.

[7] Elliott MR, Gallo JJ, Ten Have TR, Bogner HR, Katz IR. Using a Bayesian latent growth curve model to identify trajectories of positive affect and negative events following myocardial infarction. Biostatistics. 2005 ;6(1) :119-43.

[8] Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society : Series B (Methodological). 1972 ;34(2) :187-202.

[9] Rizopoulos D, Molenberghs G, Lesaffre EM. Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. Biometrical Journal. 2017 ;59(6) :1261-76.

[10] Tsiatis AA, Degruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. Journal of the American Statistical Association. 1995 ;90(429) :27-37.

[11] Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. Biometrics. 1997 :330-9.

[12] Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. Biostatistics. 2000 ;1(4) :465-80.

[13] Xu J, Zeger SL. The evaluation of multiple surrogate endpoints. Biometrics. 2001 ;57(1) :81-7.

[14] Song X, Davidian M, Tsiatis AA. An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. Biostatistics. 2002 ;3(4) :511-28.

[15] Lin H, McCulloch CE, Mayne ST. Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. Statistics in Medicine. 2002 ;21(16) :2369-82.

[16] Faucett CL, Thomas DC. Simultaneously modelling censored survival data and repeatedly measured covariates : a Gibbs sampling approach. Statistics in medicine. 1996 ;15(15) :1663-85.

[17] Yu M, Taylor JMG, Sandler HM. Individual prediction in prostate cancer studies using a joint longitudinal survival–cure model. Journal of the American Statistical Association. 2008 ;103(481) :178-87.

[18] Jacqmin-Gadda H, Thiébaut R, Dartigues JF. Modélisation conjointe de données longitudinales quantitatives et de délais censurés : Joint modeling of quantitative longitudinal data and censored survival time. Revue d'épidémiologie et de santé publique. 2004 ;52(6) :502-10.

[19] Rizopoulos D, Verbeke G, Lesaffre E. Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. Journal of the Royal Statistical Society : Series B (Statistical Methodology). 2009 ;71(3) :637-54.

[20] Böhning D. Computer-assisted analysis of mixtures and applications : meta-analysis, disease mapping and others. vol. 81. CRC press ; 1999.

[21] Jacqmin-Gadda H, Proust-Lima C, Taylor JM, Commenges D. Score test for conditional independence between longitudinal outcome and time to event given the classes in the joint latent class model. Biometrics. 2010 ;66(1) :11-9.

[22] Marquardt DW. An algorithm for least-squares estimation of nonlinear parameters. Journal of the society for Industrial and Applied Mathematics. 1963 ;11(2) :431-41.

[23] Proust-Lima C, Taylor JM. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA : a joint modeling approach. Biostatistics. 2009 ;10(3) :535-49.

[24] Lin H, McCulloch CE, Rosenheck RA. Latent pattern mixture models for informative intermittent missing data in longitudinal studies. Biometrics. 2004 ;60(2) :295-305.

[25] Zhang JJ, Wang M. Latent class joint model of ovarian function suppression and DFS for premenopausal breast cancer patients. Statistics in medicine. 2010 ;29(22) :2310-24.

[26] Garre FG, Zwinderman AH, Geskus RB, Sijpkens YW. A joint latent class changepoint model to improve the prediction of time to graft failure. Journal of the Royal Statistical Society : Series A (Statistics in Society). 2008 ;171(1) :299-308.

[27] Proust-Lima C, Séne M, Taylor JM, Jacqmin-Gadda H. Joint latent class models for longitudinal and time-to-event data : a review. Statistical methods in medical research. 2014 ;23(1) :74-90.

[28] Proust-Lima C, Philipps V, Liquet B. Estimation of extended mixed models using latent classes and latent processes : the R package lcmm. arXiv preprint arXiv :150300890. 2015.

[29] Marioni RE, Proust-Lima C, Amieva H, Brayne C, Matthews FE, Dartigues JF, et al. Cognitive lifestyle jointly predicts longitudinal cognitive decline and mortality risk. European journal of epidemiology. 2014 ;29(3) :211-9.

[30] Portegies ML, Mirza SS, Verlinden VJ, Hofman A, Koudstaal PJ, Swanson SA, et al. Mid-to late-life trajectories of blood pressure and the risk of stroke : the Rotterdam Study. Hypertension. 2016 ;67(6) :1126-32.

[31] Stamenic D, Rousseau A, Essig M, Gatault P, Buchler M, Filloux M, et al. A prognostic tool for individualized prediction of graft failure risk within ten years after kidney transplantation. Journal of Transplantation. 2019 ;2019.

[32] Ogata S, Watanabe M, Kokubo Y, Higashiyama A, Nakao YM, Takegami M, et al. Longitudinal trajectories of fasting plasma glucose and risks of cardiovascular diseases in middle age to elderly people within the general Japanese population : the Suita study. Journal of the American Heart Association. 2019 ;8(3) :e010628.

[33] Qin Y, Tian Y, Han H, Liu L, Ge X, Xue H, et al. Risk classification for conversion from mild cognitive impairment to Alzheimer's disease in primary care. Psychiatry Research. 2019 ;278 :19-26.

[34] Jiang G, Luk AOY, Tam CHT, Xie F, Carstensen B, Lau ESH, et al. Progression of diabetic kidney disease and trajectory of kidney function decline in Chinese patients with type 2 diabetes. Kidney international. 2019 ;95(1) :178-87.

[35] Brilleman SL, Moreno-Betancur M, Polkinghorne KR, McDonald SP, Crowther MJ, Thomson J, et al. Changes in body mass index and rates of death and transplant in hemodialysis patients : a latent class joint modeling approach. Epidemiology. 2019 ;30(1) :38-47.

[36] Syrjälä E, Nevalainen J, Peltonen J, Takkinen HM, Hakola L, Åkerlund M, et al. A joint modeling approach for childhood meat, fish and egg consumption and the risk of advanced islet autoimmunity. Scientific reports. 2019 ;9(1) :1-10.

[37] Carrier H, Cortaredona S, Philipps V, Jacqmin-Gadda H, Tournier M, Verdoux H, et al. Long-term risk of hip or forearm fractures in older occasional users of benzodiazepines. British journal of clinical pharmacology. 2020 ;86(11) :2155-64.

[38] Naymagon L, Zubizarreta N, Feld J, van Gerwen M, Alsen M, Thibaud S, et al. Admission D-dimer levels, D-dimer trends, and outcomes in COVID-19. Thrombosis research. 2020 ;196 :99-105.

[39] Raghavan S, Liu WG, Berkowitz SA, Barón AE, Plomondon ME, Maddox TM, et al. Association of Glycemic Control Trajectory with Short-Term Mortality in Diabetes Patients with High Cardiovascular Risk : a Joint Latent Class Modeling Study. Journal of general internal medicine. 2020 ;35(8) :2266-73.

[40] Khorashadizadeh F, Tabesh H, Parsaeian M, Esmaily H, Foroushani AR. Predicting the survival of AIDS patients using two frameworks of statistical joint modeling and comparing their predictive accuracy. Iranian Journal of Public Health. 2020 ;49(5) :949.

[41] Peter RS, Meyer ML, Mons U, Schöttker B, Keller F, Schmucker R, et al. Long-term trajectories of anxiety and depression in patients with stable coronary heart disease and risk of subsequent cardiovascular events. Depression and anxiety. 2020 ;37(8) :784-92.

[42] Graillon T, Ferrer L, Siffre J, Sanson M, Peyre M, Peyrière H, et al. Role of 3D volume growth rate for drug activity evaluation in meningioma clinical trials : the example of the CEVOREM study. Neuro-oncology. 2021 ;23(7) :1139-47.

[43] Tiruneh F, Chewaka L, Abdissa D. Statistical Joint Modeling for Predicting the Association of CD4 Measurement and Time to Death of People Living with HIV Who Enrolled in ART, Southwest Ethiopia. HIV/AIDS (Auckland, NZ). 2021 ;13 :73.

[44] Yamanouchi M, Furuichi K, Hoshino J, Toyama T, Shimizu M, Yamamura Y, et al. Two-year longitudinal trajectory patterns of albuminuria and subsequent rates of end-stage kidney disease and all-cause death : a nationwide cohort study of biopsy-proven diabetic kidney disease. BMJ Open Diabetes Research and Care. 2021 ;9(1) :e002241.

[45] Zheng H, Cui Y, Li K, Zhang J, Qu J, Shi H, et al. The association between the pattern of change in N-terminal pro-B-type natriuretic peptide and short-term outcomes in children undergoing surgery for congenital heart disease. Interactive CardioVascular and Thoracic Surgery. 2021 ;32(4) :601-6.

[46] Proust-Lima C, Joly P, Dartigues JF, Jacqmin-Gadda H. Joint modelling of multivariate longitudinal outcomes and a time-to-event : a nonlinear latent class approach. Computational statistics & data analysis. 2009 ;53(4) :1142-54.

[47] 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. Statistics in medicine. 2016 ;35(22) :3933-48.

[48] Proust-Lima C, Dartigues JF, Jacqmin-Gadda H. Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death : a latent process and latent class approach. Statistics in medicine. 2016 ;35(3) :382-98.

[49] Rouanet A, Joly P, Dartigues JF, Proust-Lima C, Jacqmin-Gadda H. Joint latent class model for longitudinal data and interval-censored semi-competing events : Application to dementia. Biometrics. 2016 ;72(4) :1123-35.

[50] Beyersmann J, Allignol A, Schumacher M. Competing risks and multistate models with R. Springer Science & Business Media ; 2011.

[51] Babykina G, Couallier V. Empirical assessment of the maximum likelihood estimator quality in a parametric counting process model for recurrent events. Computational statistics & data analysis. 2012 ;56(2) :297-315.

[52] Sirvanci M, Yang G. Estimation of the Weibull parameters under type I censoring. Journal of the American Statistical Association. 1984 ;79(385) :183-7.

[53] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. Journal of clinical epidemiology. 1996 ;49(12) :1373-9.

[54] Rizopoulos D. Joint models for longitudinal and time-to-event data : With applications in R. CRC press ; 2012.

[55] ARSLA A. SLA EN CHIFFRES : Données épidémiologiques arsla ; 2016. Available from : `https://www.arsla.org/la-sla-en-chiffres/`.

[56] Cedarbaum JM, Stambler N, Malta E, Fuller C, Hilt D, Thurmond B, et al. The ALSFRS-R : a revised ALS functional rating scale that incorporates assessments of respiratory function. Journal of the neurological sciences. 1999 ;169(1-2) :13-21.

[57] Hosmer D. Lemeshow S. Applied logistic regression. USA : John Wiley and Sons ; 2000.

[58] Youden WJ. Index for rating diagnostic tests. Cancer. 1950 ;3(1) :32-5.

[59] Lin H, Turnbull BW, McCulloch CE, Slate EH. Latent class models for joint analysis of longitudinal biomarker and event process data : application to longitudinal prostate-specific antigen readings and prostate cancer. Journal of the American Statistical Association. 2002 ;97(457) :53-65.

[60] Kurland BF, Johnson LL, Egleston BL, Diehr PH. Longitudinal data with follow-up truncated by death : match the analysis method to research aims. Statistical science : a review journal of the Institute of Mathematical Statistics. 2009 ;24(2) :211.

[61] Rizopoulos D, Lesaffre E. Introduction to the special issue on joint modelling techniques. Statistical methods in medical research. 2014 ;23(1) :3-10.