Modèles conjoints avec variance résiduelle hétéroscédastique : application à l'étude de l'impact de la variabilité de la pression artérielle sur des événements de santé compétitifs

Léonie Courcoul

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEURE
DE L'UNIVERSITÉ DE BORDEAUX

Ecole Doctorale Sociétés, Politique, Santé Publique

Spécialité Santé Publique, option Biostatistique

Par **Léonie COURCOUL**

Modèles conjoints avec variance résiduelle hétéroscédastique : application à l'étude de l'impact de la variabilité de la pression artérielle sur des événements de santé compétitifs

Sous la direction de : **Hélène JACQMIN-GADDA**

Soutenue le 26 novembre 2024

Membres du jury :

Mme. Nicola COLEY	Épidémiologiste	CHU de Toulouse	Examinatrice
M. Jérémie GUEDJ	Directeur de Recherche	INSERM U1137, Paris	Rapporteur
Mme. Hélène JACQMIN-GADDA	Directrice de Recherche	INSERM U1219, Bordeaux	Directrice de thèse
Mme. Catherine LEGRAND	Professeure	Université Catholique de Louvain	Rapporteur
Mme. Cécile PROUST-LIMA	Directrice de Recherche	INSERM U1219, Bordeaux	Présidente

Membre invité :

M. Antoine BARBIERI	Maître de Conférences	Université de Bordeaux	Co-encadrant de thèse

Remerciements

À l'image d'un roman, une thèse est portée par bien plus que son auteur. Derrière chaque idée, chaque ligne, chaque avancée, se cachent des échanges et un soutien inestimable. Avant de tourner la première page de ce travail, il me tient à cœur de saluer ces compagnons de route, sans qui cette histoire n'aurait pas existé.

À mes directeurs de thèse, Hélène Jaqmin-Gadda et Antoine Barbieri

Je commence naturellement par vous remercier, Hélène et Antoine, car sans vous cette thèse n'aurait jamais vu le jour. Vous avez formé un duo complémentaire, et je n'aurais pas pu espérer de meilleur encadrement pendant ces trois années. Merci pour ces moments de travail intenses, entre longues réunions, parfois épuisantes mais toujours enrichissantes. Je vous suis reconnaissante pour votre disponibilité, votre bienveillance et vos encouragements constants face à l'adversité des simulations ou des reviews. Vous avez toujours répondu présent pour me rassurer dans les moments de doute, et ce jusqu'au dernier moment.

Hélène, j'ai eu le privilège de bénéficier de ton immense expertise en statistiques, et j'ai énormément appris à tes côtés. Merci d'avoir toujours pris le temps de me guider, malgré ton emploi du temps très chargé. Antoine, bien que ton rôle de co-encadrant ne t'ait pas permis d'être visible pour l'administration, tu occupes une place entière dans ces remerciements. Merci pour tes nombreuses relectures, pour les séances d'entraînements aux présentations orales et tes messages d'encouragements avant chaque conférence.

Aux membres du jury,

Je tiens à remercier chaleureusement Mme Catherine Legrand et M. Jérémie Guedj, pour la qualité de leurs rapports et le temps qu'ils ont consacré à la lecture de ce manuscrit. Merci à Mme Nicola Coley d'apporter un regard épidémiologique dans l'examination de ce travail. Enfin, merci à Cécile Proust-Lima de me faire l'honneur de présider cette soutenance. Ta

rigueur scientifique et l'étendue de tes connaissances sont une véritable source d'inspiration.

À ceux qui ont accompagné ce travail,

Je remercie sincèrement tous ceux qui ont suivi et soutenu ce projet tout au long de ces trois années. Un grand merci au professeur Christophe Tzourio, à Hugues de Courson et à Karen Leffondré pour vos perspectives cliniques et épidémiologiques, qui ont été précieuses pour enrichir ce travail. À M. Jérémie Guedj, je suis reconnaissante pour toutes les discussions stimulantes sur les aspects statistiques lors des comités de suivi de thèse.

À Jérémie Riou,

Je te remercie sincèrement pour m'avoir accueillie en stage il y a maintenant quatre ans et pour m'avoir fait découvrir les modèles conjoints. C'est grâce à toi que j'ai rencontré le BPH et l'équipe Biostat.

À l'équipe Biostat,

J'ai eu la chance de réaliser cette thèse dans une équipe exceptionnelle, où règnent bienveillance, entraide et sérieux. Un immense merci à tous les membres de l'équipe, présents ou passés, pour avoir contribué à créer un environnement stimulant et enrichissant.

Merci à celles et ceux qui ont partagé mon quotidien, certains devenus des amis, du bureau S156 à l'ensemble des trois petits cochons, en passant par le S157.
À mes copines de thèse : Manel, ces trois années passées à partager le même bureau ont été riches en soutien mutuel. Merci pour tes précieux conseils. Tiphaine, on a commencé et on termine ensemble cette aventure ; merci pour ta présence pendant mes moments de doute. Kateline, que ce soit pour le sport, la course ou la thèse, tu as toujours su me guider. Merci pour ton écoute et nos échanges. Ariane, merci pour tes activités manuelles, qui m'ont permis de découvrir de nouveaux talents (je partais de très loin !). Léa et Valentine, vous avez apporté une vague de bonne humeur avec vos anecdotes et vos aventures (bonbons et manioc inclus). Lisa, ton énergie et ta capacité à travailler sans relâche m'ont toujours impressionnée. Sara, merci pour tous ces moments passés en dehors du boulot.
Federico, merci pour tes questions statistiques toujours intéressantes, ta bonne humeur et ta musique. À Justine, Blandine, Corentin, Marius, Adrien et Louis, merci pour tous les bons moments partagés.

À mes amis,

Il y a ceux qu'on rencontre en thèse, ceux qui partagent le quotidien, et puis il y a ceux qui, bien que plus éloignés, sont tout aussi essentiels. Ces amis qui nous rappellent qu'il y a une vie à côté, en dehors du code, des simulations et des articles. Léna, Sarah, Alan et Thibaut, merci pour tous les bons moments, à l'école et au-delà : pour nos visios, vacances et week-ends dans les quatre coins de la France. Ines et Audrey, merci pour nos escapades parisiennes et bordelaises. Arthur, Delphine, Louis et Ludivine, il y a dix ans, on jouait au tarot dans l'Agora. Aujourd'hui, même si se retrouver est devenu plus compliqué, chaque réunion reste un vrai plaisir, entre chansons et jeux. J'en profite pour remercier Aline et Arnaud pour leur accueil toujours très chaleureux et tous les moments de partages depuis bientôt deux ans.

À ma famille,

Je ne peux pas conclure ces remerciements sans exprimer ma gratitude envers toute ma famille. Même si vous ne comprenez pas toujours ce que je fais, votre soutien discret et constant durant ces trois années a été inestimable. Sans vous, je ne serais pas arrivée jusqu'ici. À mes parents, merci de m'avoir permis de suivre mes études et d'avoir toujours été là pour me soutenir, tant moralement que logistiquement. Merci également à mes deux petites sœurs, Aline et Ninon, je suis fière de vous voir grandir et trouver chacune votre propre voie.

Enfin, Gabriel, merci pour tout : pour ton soutien malgré la distance, ton humour, l'amour que tu m'apportes et la manière dont tu gères mon stress (ce n'est pas une mince affaire). À tes côtés j'ai trouvé ma musique du dimanche.

Valorisations scientifiques

Publications

- Courcoul L, Tzourio C, Woodward M, Barbieri A and Jacqmin-Gadda H. (2024) A location-scale joint model for studying the link between the time-dependent subject-specific variability of blood pressure and competing events. *En révision, arXiv :2306.16785.*
- Courcoul L, Helmer C, Barbieri A and Jacqmin-Gadda H. (2024) Joint model for interval-censored semi-competing events and longitudinal data with subject-specific within and between visits variabilities. *Soumis, arXiv :2408.06769.*
- Courcoul L, Jacqmin-Gadda H, Barbieri A. (2024) The R package LSJM for fitting location-scale joint models for a longitudinal marker and complex survival data. *En préparation.*

Packages R

- Courcoul L, Jacqmin-Gadda H, Barbieri A. (2024) `LSJM` : *Location-scale joint models for a longitudinal marker and complex survival data.* GitHub development version. https://github.com/LeonieCourcoul/LSJM
- Courcoul L, Jacqmin-Gadda H, Barbieri A. (2023) `FlexVarJM` : *Estimate Joint Models with Subject-Specific Variance.*
  R package version : 0.1.0. https://CRAN.R-project.org/package=FlexVarJM

Présentations orales en conférences internationales

- Courcoul L, Tzourio C, Woodward M, Barbieri A and Jacqmin-Gadda H. A flexible location-scale joint model to study the effect of blood pressure variability on competing events. *Fifth International Workshop on Statistical Analyses of Multi-Outcome Data - July 2024 - Salzburg, Austria.*
- Courcoul L, Tzourio C, Barbieri A and Jacqmin-Gadda H. A joint model for competing risks and longitudinal marker with a time-dependent subject-specific variance. *The 9th Survival Analysis for Junior Researchers conference - September 2023 - Ulm, Germany.*
- Courcoul L, Tzourio C, Barbieri A and Jacqmin-Gadda H. A location-scale joint model to study the effect of individual time-dependent variability of blood pressure on competing events. *44th Annual Conference of the International Society for Clinical Biostatistics - August 2023 - Milan, Italy.*
- Courcoul L, Barbieri A, Tzourio C and Jacqmin-Gadda H. Joint model with heterogeneous variance for studying the risk of stroke associated with blood pressure variability. *31st International Biometric Conference - July 2022 - Riga, Latvia.*

Présentations orales en conférences nationales

- Courcoul L, Barbieri A, Tzourio C and Jacqmin-Gadda H. Variabilités intra et intervisites de la pression artérielle et risque de démence : un modèle conjoint avec variance résiduelle individuelle. *55èmes Journées de Statistiques - Mai 2024 - Bordeaux.*
- Courcoul L, Barbieri A, Tzourio C and Jacqmin-Gadda H. Joint model with heterogeneous variance for studying the risk of stroke associated with blood pressure variability *Journées des Biostatistiques - Novembre 2022 - Rennes.*
- Courcoul L, Barbieri A, Tzourio C and Jacqmin-Gadda H. Impact de la variabilité de la pression artérielle sur le risque d'AVC. *53èmes Journées de Statistiques - Juin 2022 - Lyon.*

Poster

- Courcoul L, Barbieri A, Woodward M, Tzourio C and Jacqmin-Gadda H. Impact of blood pressure variability on the risk of stroke and premature death. *International Society Hypertension Kyoto 2022 Meeting - Octobre 2022 - Japan (à distance).*

Communication invitée en séminaire

- Courcoul L, Barbieri A, and Jacqmin-Gadda H. Modèles conjoints avec variance résiduelle flexible et hétéroscédastique : application à l'étude de l'impact de la variabilité de la pression artérielle sur des événements de santé compétitifs. *Rencontre d'échanges en biostatistiques du Centre de recherche du CHU de Québec - Université Laval - Juin 2024.*

Récompense scientifique

- Student Conference Award of the 2023 International Society for Clinical Biostatistics Conference, Milan, Italy, 2023

Liste des abréviations

- AIC : *Akaike Information Criteria*
- AUC : Area Under the Curve
- AVC : Accident Vasculaire Cérébral
- CCVD : *Cardio and Cerebro-Vascular Disease*
- DSM-IV : (*Diagnostic and Statistical Manual of Mental Disorders*)
- EM : *Expectation-Maximisation*
- INLA : *Integrated Nested Laplace Approximation*
- LOCF : *Last Observation Carried Forward*
- LSJM : *Location Scale Joint Model*
- LSMM : *Location Scale Mixed Model*
- OMS : Organisation Mondiale de la Santé
- PA : Pression Artérielle
- PAD : Pression Artérielle Diastolique
- PAM : Pression Artérielle Moyenne
- PAS : Pression Artérielle Systolique
- QMC : *Quasi-Monte-Carlo*

Table des matières

Chapitre I

Introduction

Sommaire

Le rapport de l'Organisation Mondiale de la Santé (OMS) sur l'hypertension artérielle commence par ces quatre mots : "*High blood pressure kills*". Il décrit les actions à mener comme une "course contre un tueur silencieux" responsable de nombreuses complications de santé comme les événements cardio et cérébrovasculaires ou la démence (WHO, 2023). Ce facteur de risque, reconnu depuis de nombreuses décennies, suscite toujours un important intérêt à tel point que de récentes recherches se sont intéressées plus particulièrement à l'impact de la variabilité de la pression artérielle sur de tels événements (Rothwell et al., 2010; Mehlum et al., 2018; Ma et al., 2020; De Courson, 2022). Mon projet de thèse s'inscrit dans ce dernier axe de recherche et vise à proposer et développer des méthodes statistiques robustes afin d'étudier l'impact de la variabilité de la pression artérielle sur le risque d'événements de santé majeurs tels que les événements cardiovasculaires et cérébrovasculaires, la démence mais aussi le décès.

I.1 Les événements cardio et cérébrovasculaires

I.1.1 Définition

Dans cette thèse, on entendra par événements cardio et cérébrovasculaires les maladies coronariennes avec survenue d'un infarctus du myocarde, les décès vasculaires ainsi que les accidents vasculaires cérébraux (AVC). On utilisera par la suite CCVD (*Cardio and Cerebro-Vascular Disease*) pour parler des événements ou maladies cardio et cérébrovasculaires.

La maladie coronarienne apparaît lorsque du cholestérol s'accumule, sous forme de plaques d'athéromes sur les parois des artères coronaires, ce qui engendre un durcissement de ces dernières et une sous-alimentation du coeur en oxygène. Il peut arriver qu'une plaque se détache, se déplace et enfin, s'immobilise dans une artère coronaire. Il y a alors formation d'un caillot de sang autour de cette plaque, interrompant ainsi l'apport en sang et privant par conséquent le coeur d'oxygène, entraînant par la suite la destruction d'une partie plus ou moins grande du muscle cardiaque. C'est ce qu'on appelle l'infarctus du myocarde (Figure I.1).

Les accidents vasculaires cérébraux peuvent soit être de type ischémique soit de type hémorragique. Les premiers, les plus fréquents, sont provoqués par l'arrêt du flux sanguin dans un vaisseau du cerveau. En général, cet arrêt survient suite à la formation d'un caillot. Le tissu cérébral est alors privé d'oxygène, conduisant à une nécrose des cellules. Les deuxièmes sont définis par un saignement intracrânien suite à la rupture d'un vaisseau (Figure I.2).

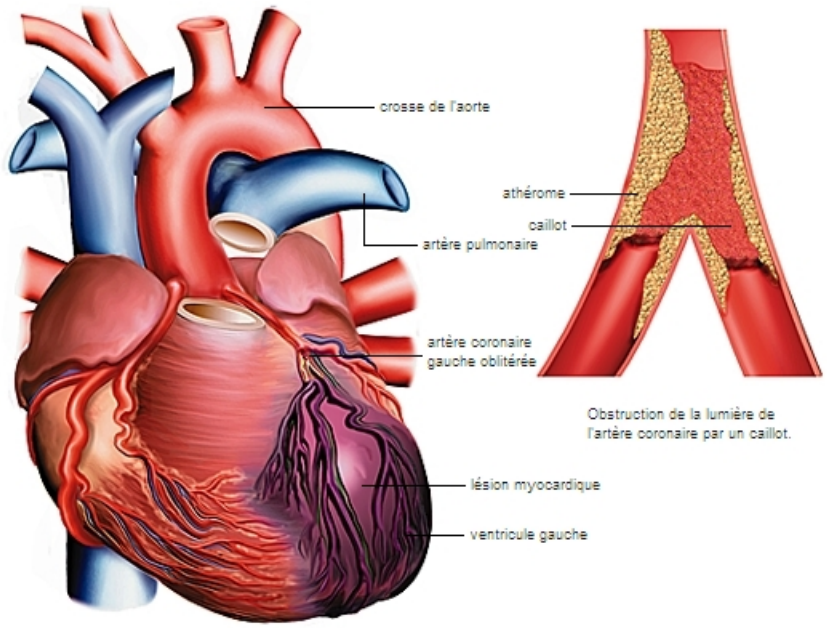

FIGURE I.1 – Infarctus du myocarde (`https://www.larousse.fr/encyclopedie/images/Infarctus_du_myocarde/1002756`)

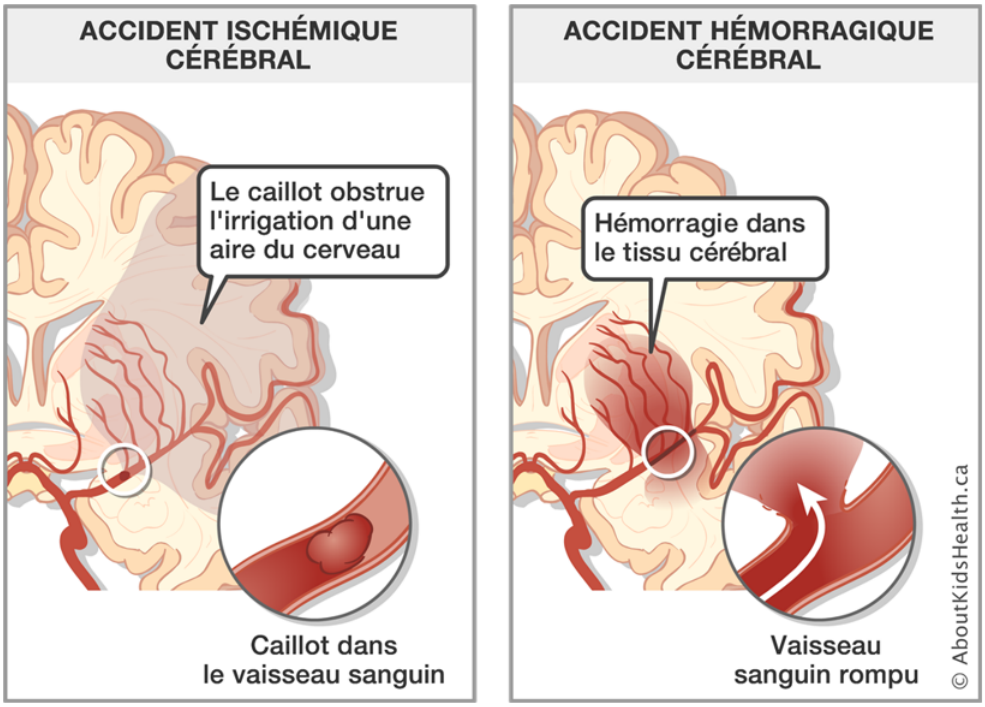

FIGURE I.2 – Les deux types d'AVC (`https://www.ffn-neurologie.fr/maladies/accident-vasculaire-cerebral-avc/`)

Les décès vasculaires correspondent aux décès par AVC ou tout autre événement vascu-

laire.

I.1.2 Épidémiologie

D'après l'OMS, les maladies cardiovasculaires et cérébrovasculaires représentent la première cause de décès, avec plus de 17 millions de morts chaque année. En particulier, près de 80% des décès liés à un CCVD surviennent lors d'une crise cardiaque ou d'un accident vasculaire cérébral. Ces décès touchent également les sujets jeunes puisqu'un tiers de ceux-ci surviennent chez des personnes de moins de 70 ans. En particulier, Salari et al. (2023) ont mis en évidence une prévalence globale de l'infarctus du myocarde à respectivement 3.8% et 9.5% chez les moins de 60 ans et plus de 60 ans. Les accidents vasculaires cérébraux représentent quant à eux la première cause de handicap acquis chez l'adulte avec 12.2 millions d'AVC chaque année (eClinicalMedicine, 2023). On estime qu'une personne sur quatre, âgée de 25 ans ou plus, sera victime d'un AVC durant sa vie. Cela engendre un coût non négligeable de dépense de santé. De Pouvourville (2016) estime en effet que la collectivité française a financé près de 8.6 milliards d'euros en 2007 pour les victimes d'AVC avec un coût pour les nouveaux cas incidents sur une année variant entre 7 839€ et 41 437€ en fonction de la gravité de l'AVC.

Les principaux facteurs de risque connus des maladies coronariennes sont l'âge, l'hypertension artérielle, le poids, la sédentarité et la consommation excessive d'alcool ou de tabac.

I.2 La démence

I.2.1 Définition

La démence est une maladie cognitive chronique et évolutive caractérisée par un déclin cognitif plus rapide que celui de personnes du même âge. Elle est définie selon les critères du DSM-IV (*Diagnostic and Statistical Manual of Mental Disorders*) par l'altération de la mémoire et d'au minimum une autre fonction cognitive, par exemple le langage ou les fonctions exécutives, de façon suffisamment importante pour entraver les activités de la vie quotidienne, voir l'état de santé général. La perte d'autonomie conduit dans la plupart des cas à l'institutionnalisation des patients puis au décès. La démence peut se retrouver sous trois formes : la maladie d'Alzheimer qui représente plus de 60% des cas (AD, 2022), la démence vasculaire et la démence avec corps de Lewy. Cependant, la distinction entre ces trois origines de démence est difficile et les deux premières peuvent souvent coexister.

La maladie d'Alzheimer est une maladie neurodégénérative causée par une lente dégénérescence des neurones liée à la modification de deux molécules, le peptide bêta amyloïde et la protéine tau. La première est naturellement présente dans le cerveau mais lorsqu'elle s'accumule anormalement au point de former des plaques, appelées plaques amyloïdes, cela devient toxique pour les cellules nerveuses. La deuxième est une protéine de structure des neurones. Chez les patients atteints par la maladie, elle est modifiée, provoquant successivement la désorganisation des neurones, une accumulation de filaments à l'intérieur de ces derniers puis la mort des cellules nerveuses.

La démence vasculaire est une perte de la fonction cognitive due à la destruction du tissu cérébral lorsque son apport en sang est réduit ou complètement bloqué, généralement causé par un AVC.

La démence à corps de Lewy est également une maladie neurodégénérative dans laquelle la dégénérescence des neurones est liée à la création de dépôts anormaux de la protéine alpha-synucléine à l'intérieur des cellules cérébrales.

I.2.2 Epidémiologie

L'*Alzheimer's society* estime à près de 55 millions le nombre de personnes atteintes de démence dans le monde avec près de 10 millions de nouveaux cas chaque année. La démence représente la septième cause de mortalité dans le monde et l'une des causes majeurs d'handicap et de dépendance parmi les personnes âgées. Par ailleurs, en 2019, le coût lié à la démence dans le monde était estimé à près de 1300 milliards de dollars (Wimo et al., 2023).

De nombreux facteurs de risques favorisent la démence, notamment l'âge et la génétique. En effet, en particulier dans la maladie d'Alzheimer, plusieurs gènes seraient liés à un risque accru de la maladie (gènes du métabolisme du peptide amyloïde, gènes impliqués dans l'inflammation, gènes entraînant la communication entre neurones etc.), tandis que certains protégeraient contre la maladie. Cependant, selon Livingston et al. (2020), 40% des cas de démences seraient dus à douze facteurs de risques modifiables :

- un niveau scolaire faible,
- l'hypertension,
- la déficience auditive,
- le tabac,

- l'obésité,
- la dépression,
- l'inactivité physique,
- le diabète,
- l'isolement social,
- la consommation excessive d'alcool,
- les traumatismes crâniens,
- la pollution atmosphérique.

De plus, il n'existe actuellement aucun traitement curatif contre la démence. Quatre médicaments existent sur le marché français pour ralentir l'évolution de la maladie mais étant données leur faible efficacité et leur mauvaise tolérance, la Haute Autorité de Santé a estimé qu'ils n'avaient plus leur place dans le traitement de la maladie. Ils sont toujours commercialisés mais non remboursés.

Ainsi, étant donnée l'importance des facteurs de risques modifiables cités précédemment et le manque de traitement, le potentiel de la prévention est très important et ce, à n'importe quel âge (Livingston et al., 2020). Enfin, de nombreuses actions peuvent être mises en place au quotidien pour maintenir, autant que possible, la qualité de vie des patients et leur bien-être, comme l'activité physique ou les interactions sociales afin de stimuler leur cerveau.

I.3 La pression artérielle

Les deux événements de santé présentés précédemment ayant pour facteur de risque commun l'hypertension artérielle, il apparaît comme nécessaire de s'y intéresser davantage.

Le cœur pompe le sang riche en oxygène afin de le distribuer, par l'intermédiaire des veines et des artères, à l'ensemble du corps. Lorsque le sang traverse le cœur, il exerce une force sur les parois artérielles. Cette force correspond à la pression artérielle (PA). Plus précisément, la PA correspond à la mesure prise lorsque le ventricule gauche du cœur se contracte et que le sang est projeté vers les artères. Elle permet alors d'évaluer la quantité de sang pompée par le cœur et l'état général des artères. La PA est composée de deux éléments : la pression artérielle systolique (PAS) qui est la pression maximale lors de la contraction du ventricule gauche, et la pression artérielle diastolique (PAD), mesurée lors du relâchement du ventricule,

lorsque le cœur est au repos. Ces deux mesures permettent de définir la pression artérielle moyenne (PAM) correspondant à un tiers de la PAS plus deux tiers de la PAD.

La pression artérielle est un biomarqueur très intéressant. En effet, sa mesure est facile, non invasive et peu coûteuse. Elle peut être mesurée au quotidien par les patients eux-mêmes ou de façon plus espacée par un professionnel de santé. De plus, elle est facilement contrôlable, que ce soit via des traitements ou des actions de prévention à l'échelle de l'hygiène de vie (nutrition, activité physique, etc). Enfin, elle est directement liée à des événements de santé d'intérêts via l'hypertension artérielle qui est un facteur de risque connu de nombreux événements. Il est donc nécessaire de mesurer la PA afin de diagnostiquer une hypertension et traiter le patient en conséquence.

I.3.1 Hypertension artérielle

I.3.1.1 Définition

L'hypertension artérielle est une condition médicale chronique qui apparaît lorsque la pression dépasse certains seuils. Selon les seuils de l'OMS, une personne souffre d'hypertension artérielle si sa PAS dépasse 140 mmHg ou si sa PAD dépasse 90 mmHg, lors de mesures répétées pendant une consultation médicale et après un repos adéquat.

I.3.1.2 Épidémiologie et facteurs de risques

Selon l'OMS, 33% des adultes âgés entre 30 et 79 ans souffraient d'hypertension artérielle en 2019. Bien que cette prévalence soit restée globalement stable depuis 1990 (32% cette année-là), étant donnée l'augmentation du nombre d'adultes dans le monde, le nombre de cas a doublé entre 1990 et 2019, passant ainsi de 650 millions à 1.3 milliards d'adultes souffrant d'hypertension artérielle (WHO, 2023). Par ailleurs, bien qu'elle soit facile à diagnostiquer et malgré les traitements existants, on estime que seulement 54% des adultes âgés entre 30 et 79 ans et souffrant d'hypertension ont été diagnostiqués, que 42% des 30-79 ans souffrant d'hypertension sont traités et 21% ont une hypertension artérielle contrôlée (WHO, 2023). Une étude réalisée en France en 2015, auprès de 3000 sujets âgés de 18 à 74 ans a permis de trouver une prévalence à 31.4% avec près de la moitié des sujets hypertendus qui l'ignoraient (Vallée et al., 2020). De plus, cette prévalence augmente avec l'âge puisqu'on estime qu'environ 60% des sujets âgés de plus de 60 ans sont hypertendus (Chow et al., 2013). Enfin, parmi les adultes ayant une hypertension artérielle non contrôlée et qui n'ont pas été précédemment diagnostiqués, près de 30% auraient une PAS supérieure à 160 mmHg ou une PAD supérieure à 100 mmHg. Ces valeurs qui dépassent nettement les seuils définis

par l'OMS indiquent un besoin urgent de prise en charge.

Bien que les facteurs génétiques et les antécédents familiaux peuvent contribuer à la prédisposition individuelle face à l'hypertension artérielle, la majorité des facteurs de risques peuvent être contrôlés via une bonne hygiène de vie permettant de lutter contre l'inactivité physique, une alimentation déséquilibrée (consommation élevée de sel par exemple), l'obésité ou encore une consommation excessive d'alcool ou de tabac.

I.3.1.3 L'hypertension artérielle comme facteur de risque

Les risques liés à l'hypertension artérielle sont majeurs, que ce soit pour la mortalité en générale ou pour des maladies cardiovasculaires ou rénales. Souffrir d'hypertension artérielle est un risque pour la santé globale. En effet, plus la PA est élevée, plus le coeur doit pomper fortement. Cet excès de pression va ainsi endommager de nombreux organes, en particulier le cerveau, le coeur ou les reins. Toujours d'après l'OMS, une PAS élevée (>110-115 mmHg) est le principal facteur de risque de mortalité dans le monde, causant plus de décès que tout autre facteur de risque comportemental, environnemental ou métabolique (WHO, 2023). Il a notamment été estimé que si tous les adultes avaient une PAS inférieure à 110-115 mmHg, environ 19% des décès auraient pu être évités en 2019. En particulier, cette année-là, plus de la moitié des décès cardiovasculaires ou par AVC est attribuée à une trop forte valeur de PAS. Bien que l'âge soit un facteur de risque non négligeable pour ces maladies et pour l'hypertension artérielle, notons tout de même que 38%, soit 4 millions, de décès dus à l'hypertension artérielle ont eu lieu chez des sujets de moins de 70 ans.

Le lien entre hypertension artérielle et maladie cardiovasculaire ou cérébrovasculaire est aujourd'hui bien reconnu et a maintes fois été démontré. En particulier, une méta-analyse portant sur l'association entre la pré-hypertension, soit une PAS comprise entre 120 et 139 mmHg et une PAD entre 80 et 89 mmHg, et les accidents cardiovasulaires a mis en évidence une augmentation de 44%, 73% et 79% du risque de décès cardiovasculaires, d'AVC et d'infarctus du myocarde respectivement (Guo et al., 2013). Plus précisément, en ce qui concerne l'AVC, Willey et al. (2014) et Clark et al. (2019) ont montré que le risque d'AVC attribuable à l'hypertension artérielle est de l'ordre de 25 à 30%, et Hankey (2020) suggère qu'il s'agit du premier facteur de risque modifiable de la survenue d'AVC.

Le lien entre hypertension et démence est plus controversé. En effet, certaines études observationnelles suggèrent une association en U, plutôt qu'une association linéaire, entre la PA et l'incidence de démence chez les sujets âgés (Lee et al., 2022; van Dalen et al., 2022).

En particulier, van Dalen et al. (2022) ont mis en évidence un risque de démence plus faible chez les personnes âgées ayant des niveaux de PAS plus élevés et des associations en forme de U chez les personnes de plus de 75 ans. Wang et al. (2018) ont trouvé quant à eux une relation non linéaire entre la PAS et le risque de démence dans une population âgée de 65 ans et plus. Il ont d'abord mis en évidence qu'une PAS comprise entre 110 et 120 mmHg jouait un rôle protecteur chez les personnes âgées de 62 à 82 ans. En parallèle il ont montré qu'une PAS supérieure à 162 mmHg augmenterait significativement le risque de démence chez les personnes âgées de 70 à 86.5 ans. D'autres études suggèrent que seule l'hypertension artérielle à l'âge moyen, défini à partir de 40 ans, est un facteur de risque, et non pas l'hypertension de la personne âgée (Livingston et al., 2020). En particulier, dans la cohorte *Framingham Offspring*, une PAS supérieure à 140 mmHg autour de 55 ans était associé à un risque accru de développer la démence ($HR = 1.6\,;\ IC = [1.1, 2.4]$) sur une période de suivi de 18 ans (McGrath et al., 2017). De plus, dans cette étude le risque augmentait si l'hypertension persistait jusqu'à un âge avancé (autour de 69 ans). Cependant, chez les individus non-hypertendus à l'âge moyen, une baisse rapide de la PAS entre l'âge moyen et l'âge avancé était également associée à une augmentation de plus de 2 fois du risque de démence (McGrath et al., 2017). Par ailleurs, dans la même cohorte, les individus avec des paramètres cardiovasculaires idéaux (non-fumeur actuel, indice de masse corportel entre 18,5 et 25 $\mathrm{kg/m^2}$, activité physique régulière, alimentation saine, PAS inférieure à 120 mmHg et PAD inférieure à 80 mmHg, cholestérol et glycémie normales), autour de 62 ans, avaient un risque plus faible sur 10 ans, de démence toutes causes confondues ($HR = 0.8\,;\ IC = [0.1, 1.0]$), de démence vasculaire ($HR = 0.5\,;\ IC = [0.3, 0.8]$) et de maladie d'Alzheimer ($HR = 0.8\,;\ IC = [0.6, 1.0]$) que des individus du même âge présentant au moins un risque cardiovasculaire (Pase et al., 2016). Dans une cohorte britannique il a été montré qu'une mesure de la PAS de 130 mmHg ou plus à 50 ans était associé à un risque accru de démence, alors que ça n'était pas le cas pour une mesure à 60 ans ou à 70 ans (Abell et al., 2018). Une autre étude de cohorte a rapporté de potentielles explications sur les mécanismes en jeux (Lane et al., 2019). D'après cette étude, l'hypertension à l'âge moyen serait associée à une réduction des volumes cérébraux et une augmentation du volume de certaines lésions de la substance blanche, mais pas à une accumulation d'amyloïde. Enfin, il a été mis en évidence que la prise d'un traitement antihypertenseur diminuerait le risque de survenue de la démence (in't Veld et al., 2001; Forette et al., 2002; Ding et al., 2020).

I.3.2 Variabilité de la pression artérielle comme facteur de risques

Plus récemment, la littérature scientifique a mis en évidence un nouveau facteur de risque cardio et cérébrovasculaire mais aussi de démence et de décès : la variabilité de la pression artérielle. Dans les années 1980, Parati et al. (1998) suspectaient déjà un lien entre cette variabilité et la lésion de certains organes. Cependant, ce n'est que dans les années 2010 que la littérature a commencé à fortement se développer sur cette hypothèse avec les travaux de Rothwell (Rothwell et al., 2010).

I.3.2.1 Lien avec les événements cardio et cérébrovasculaires

Les travaux de Rothwell et al. (2010) sont à l'origine d'un intérêt grandissant pour l'étude de la variabilité de la pression artérielle comme facteur de risque de certains événements majeurs de santé publique. Ainsi, ces auteurs ont pu mettre en évidence une association forte entre la variabilité de la PAS et les AVC ischémiques, indépendamment du niveau courant de PA. Cependant, les résultats de cette analyse ne portant que sur des essais cliniques, il n'est pas possible de transposer ces résultats à la population générale. En 2014, Yu et al. (2014) ont pu répliquer ces résultats dans une cohorte d'environ 120 000 individus hypertendus sans antécédents d'AVC. Ils ont trouvé une augmentation du risque d'AVC (tout type) de 4.2% pour une augmentation d'environ 9 mmHg de la variabilité de la PAS, indépendamment du niveau courant de la PAS. Ils ont trouvé des résultats semblables pour l'effet de la variabilité de la PAD. Par la suite, Mehlum et al. (2018) ont mis en évidence une augmentation du risque d'événements cardiovasculaires pour les patients du plus haut quintile de pression artérielle (HR = 2.1, CI = [1.7,2.4]), ainsi qu'une augmentation du risque de décès de 10% pour une augmentation de 5 mmHg de l'écart-type de la PAS. Enfin, dans une méta-analyse regroupant 28 articles, De Courson (2022) ont retrouvé que 17 articles faisaient état d'une association significative entre au moins un indicateur de variabilité de la pression artérielle et le risque d'AVC, une augmentation de la variabilité de la pression artérielle induisant une augmentation du risque instantané d'AVC.

I.3.2.2 Lien avec la démence

En ce qui concerne la démence, l'intérêt porté à l'étude de la variabilité de la pression artérielle comme facteur de risque est plus récent et cela n'a été examiné que dans quelques études de cohortes prospectives (Ma et al., 2020). Alpérovitch et al. (2014), ont montré qu'une plus grande variabilité de la pression artérielle sur 3 visites étalées sur 4 ans était associée

à un risque plus élevé de démence dans la cohorte française des Trois Citées [1]. Ma et al. (2019) confirment avec l'étude de Rotterdam portant sur une population plus âgée, qu'une augmentation de la variabilité de la PA entre des visites successives est associée à un risque plus élevé de démence à long terme. Cette association est plus forte lorsque la variabilité de la PA est mesurée 15 ans avant le diagnostic de démence. D'autres études ont pu mettre en évidence un lien entre la variabilité de la PA et un déclin cognitif plus élevé (van Middelaar et al., 2018; Rouch et al., 2020).

Cependant, toutes ces études n'utilisent pas toujours la même définition de la variabilité. Ainsi, elles ne sont pas comparables entre elles et il manque aujourd'hui des études sur l'association entre la démence et la variabilité de la pression artérielle à court terme, définie sur des périodes de quelques minutes ou de quelques heures.

I.3.2.3 Hétérogénéité des méthodologies statistiques

Les études citées dans cette section I.3.2 et portant sur l'étude de la variabilité de la pression artérielle comme facteur de risque d'événements de santé sont très hétérogènes en terme de définition et de stratégie d'analyse. Dans cette partie nous décrirons ces différentes approches en discutant des limites de chacune. Nous évoquerons également les différentes stratégies mises en place pour estimer l'effet de la variabilité sur le risque d'événement.

Indicateurs de variabilité

Afin de calculer la variabilité de la PA, divers indicateurs peuvent être utilisés. Le tableau I.1 fait état des principaux indicateurs utilisés et évoqués dans cette thèse.

Dans la revue systématique de la littérature de De Courson (2022), 57% des études se sont intéressées à l'écart-type et 50% à l'écart-type de la PA discrétisé en classes comme définition de la variabilité. Le coefficient de variation et l'écart moyen entre mesures successives (appelé variabilité réelle moyenne dans De Courson (2022)) sont utilisés dans un quart des articles tandis que la racine de l'erreur quadratique moyenne (RMSE) n'est utilisée que dans 11% des cas. La majorité des articles utilise plusieurs indicateurs de variabilités. Par ailleurs, le nombre de mesures de PA utilisées varie grandement entre les différents articles étudiés dans cette revue (entre 3 et 24 mesures) avec une majorité d'articles utilisant au moins 5 mesures. À noter que les indicateurs précédemment listés requièrent au minimum deux mesures pour pouvoir être calculés.

1. Cohorte présentée en section I.4.2

TABLE I.1 – Indicateurs utilisés pour définir la variabilité individuelle de la pression artérielle (PA) (De Courson, 2022), pour n mesures répétées de la pression artérielle au cours du temps.

Mesure	Formule	Précisions
Ecart-type (ET)	$\sqrt{\frac{\sum_{j=1}^{n}(PA_j-\overline{PA})^2}{n-1}}$	- PA_j : la j-ème mesure observée de la PA chez le sujet - $\overline{PA}$: la moyenne empirique des mesures observées de PA - n : le nombre de mesures observées de la PA - reflète la dispersion des valeurs autour de la moyenne empirique global
Coefficient de variation	$\frac{ET}{\overline{PA}} \times 100$	- reflète la dispersion des valeurs autour de la moyenne empirique globale en prenant en compte le niveau moyen de PA. - permet de comparer des variabilités avec des moyennes de PA différentes
Ecart moyen entre mesures successives	$\frac{\sum_{j=1}^{n-1}\lvert PA_{j+1}-PA_j\rvert}{n-1}$	- représente la variabilité moyenne entre deux mesures de PA successives
Racine de l'erreur quadratique moyenne (RMSE)	$\sqrt{\frac{\sum_{j=1}^{n}(\widehat{PA_j}-PA_j)^2}{n}}$	- $\widehat{PA_j}$: la valeur de la j-ème mesure de PA prédite par un modèle de régression linéaire portant sur les n observations de la PA et supposant une tendance linéaire de la PA dans le temps - plus le RMSE est faible, plus la PA suit une tendance linéaire au cours du temps avec peu de variations autour de cette tendance linéaire

Variabilité calculée sur une période initiale

La première stratégie consiste à calculer l'indicateur de variabilité choisi, en général l'écart-type, en utilisant les mesures de PA collectées durant une première période, jusqu'à un temps S fixé. Puis la variabilité ainsi calculée est incluse comme une variable explicative fixe dans le temps dans un modèle de Cox [2] afin d'estimer l'association entre cette variabilité calculée et le risque de développer l'événement d'intérêt après S (Figure I.3 A). Cependant, seuls les sujets encore à risque au temps S sont inclus dans l'estimation du modèle de Cox. Ainsi, si la variabilité de la PA est associée à un risque plus élevé d'événement, on peut supposer que les individus ayant une plus grande variabilité font partie de ceux ayant eu l'événement précocement et sont donc exclus de l'analyse. Cette stratégie entraîne par conséquent une perte d'événements observés, une diminution de la puissance ainsi qu'un biais de sélection.

2. modèle présenté en section II.2.2.

Notons que cette méthode est tout de même utilisée dans la moitié des articles étudiés dans la revue de littérature proposée par De Courson (2022).

Variabilité calculée sur l'ensemble du suivi

L'autre moitié de ces études considère une deuxième stratégie qui consiste à calculer l'indicateur de variabilité sur l'ensemble des mesures de pression artérielle collectées au cours du suivi puis, de façon similaire, d'inclure cette variabilité comme une variable explicative fixe dans un modèle de Cox afin d'estimer l'association entre la variabilité et le risque d'événement sur une même période de temps (Figure I.3 B). Plus de trois quart de ces études incluaient les mesures observées après l'événement. L'inclusion des données collectées après le temps courant ou après le temps d'événement induit un conditionnement sur le futur pouvant mener à un biais (Andersen and Keiding, 2012; de Courson et al., 2018; Torp-Pedersen et al., 2018).

Par ailleurs, dans ces deux stratégies, la variabilité de la pression artérielle est considérée comme une variable fixe dans le temps faisant ainsi l'hypothèse qu'elle est constante au cours du temps, ce qui, d'un point de vue clinique, semble très peu probable.

Variabilité dépendante du temps

Pour éviter ces biais, l'écart-type (ou tout autre indicateur) de la PA, peut être considéré comme une variable dépendante du temps. La méthode naïve consiste à remettre à jour l'écart-type de la pression artérielle à chaque visite en incluant la ou les nouvelle(s) valeur(s) observée(s) dans le calcul. Ainsi la variabilité de la PA est incluse dans le modèle de Cox comme une variable dépendante du temps (Figure I.3 C). Cela permet de ne pas conditionner sur le futur et de ne pas avoir de problème de biais de sélection tout en autorisant la variabilité à évoluer au cours du temps. Cependant, cette approche fait l'hypothèse d'une variabilité constante entre deux temps de mesures et néglige l'erreur de mesure de l'écart-type, ce qui peut devenir un problème non négligeable lorsque le nombre de mesures diffère entre les individus. Par ailleurs, cette stratégie requière l'imputation de l'écart-type à tous les temps d'événement. Enfin, la PA et son écart-type sont des variables endogènes, c'est-à-dire que leur valeur est modifiée par la survenue de l'événement. Or, le modèle de Cox avec variable dépendante du temps nécessite une variable exogène, donc une variable dont la valeur n'est pas modifiée par la survenue des événements (Rizopoulos, 2010). Ces limites peuvent également introduire des biais.

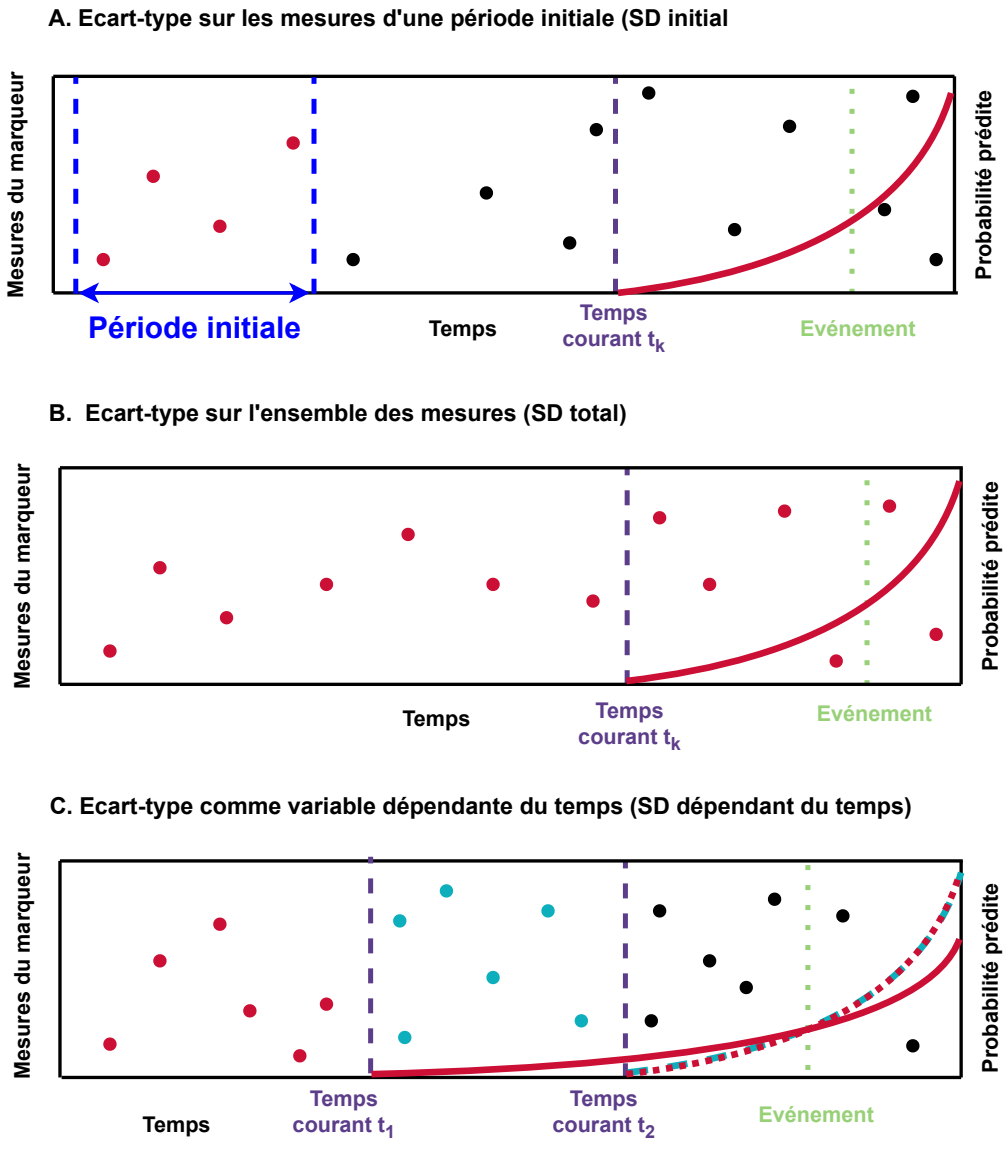

FIGURE I.3 – Les différentes stratégies de prise en compte de la variabilité. La couleur des courbes de probabilité d'événement estimées dépend des observations utilisées pour calculer sa variabilité. Pour les graphiques A et B, les mesures rouges sont utilisées pour calculer la variabilité de la pression artérielle qui est ensuite utilisée comme covariable fixe dans le modèle de Cox. Pour le graphique C, les mesures rouges sont utilisées pour estimer la probabilité d'événement représentée par la courbe rouge. Les mesures en rouge et celles en bleue sont utilisées pour prédire l'événement à partir du temps t_2, ce qui correspond à la courbe bleue et rouge.

L'approche par modèle conjoint à effets aléatoires partagés (détaillée en section II.3) permet de contourner les limites précédemment citées. En effet, ces modèles permettent de prendre en compte la relation entre le temps jusqu'à l'événement et un marqueur longitudinal en respectant les caractéristiques d'une variable dépendante du temps et endogène (Rizopoulos, 2010). Un modèle conjoint combine un modèle linéaire mixte et un modèle à risque proportionnel. Le modèle linéaire mixte permet de modéliser la trajectoire individuelle de l'écart-type de la PA. Ainsi, la valeur courante de l'écart-type est estimée en temps continu et simultanément incorporée comme une variable dépendante du temps dans le modèle à risque proportionnel afin d'estimer l'impact de cette variabilité sur la survenue d'un événement.

Toutes les méthodes présentées jusqu'à présent nécessitent d'avoir au moins deux mesures de la PA pour chaque individu. De plus, la variabilité considérée est une variabilité globale (l'écart-type des mesures observée depuis le début du suivi), ce qui ne donne pas nécessairement une bonne définition de la variabilité de la pression artérielle. En effet, si la pression artérielle diminue ou augmente avec le temps, la variabilité va par nature augmenter. Mais est-ce la variabilité qui augmente ou la moyenne de la pression artérielle qui évolue ? Il semble par conséquent plus intéressant de considérer la variabilité résiduelle, soit la variabilité des mesures autour de la tendance centrale, une trajectoire moyenne.

Modèle conjoint avec variance hétérogène

Gao et al. (2011) et Barrett et al. (2019) ont proposé un modèle conjoint (détaillé en section II.4.2) dans lequel la variance résiduelle du marqueur est spécifique à l'individu, via l'ajout d'un effet aléatoire sur la variance de l'erreur résiduelle. Cet effet aléatoire peut être ajouté comme covariable représentant la variance résiduelle dans le modèle de survie. Cette stratégie fait néanmoins l'hypothèse d'une variabilité fixe dans le temps.

Les méthodes précédemment présentées ont été comparées par de Courson et al. (2021). Les valeurs de variabilité de la pression artérielle diffèrent en fonction de la méthode utilisée (Figure I.4) mais la comparaison a également mis en lumière des résultats contradictoires concernant l'effet de cette variabilité sur le risque d'AVC. En effet, avec l'utilisation de la méthode naïve prenant l'ensemble des mesures de la pression artérielle, une association relativement forte entre la variabilité de la pression artérielle et le risque de récidive d'AVC a été trouvée alors qu'en excluant les mesures postérieures à l'événement cet effet s'inverse. En utilisant les autres méthodes permettant de s'affranchir des limites du modèle précédent,

l'effet s'amenuise voire disparaît.

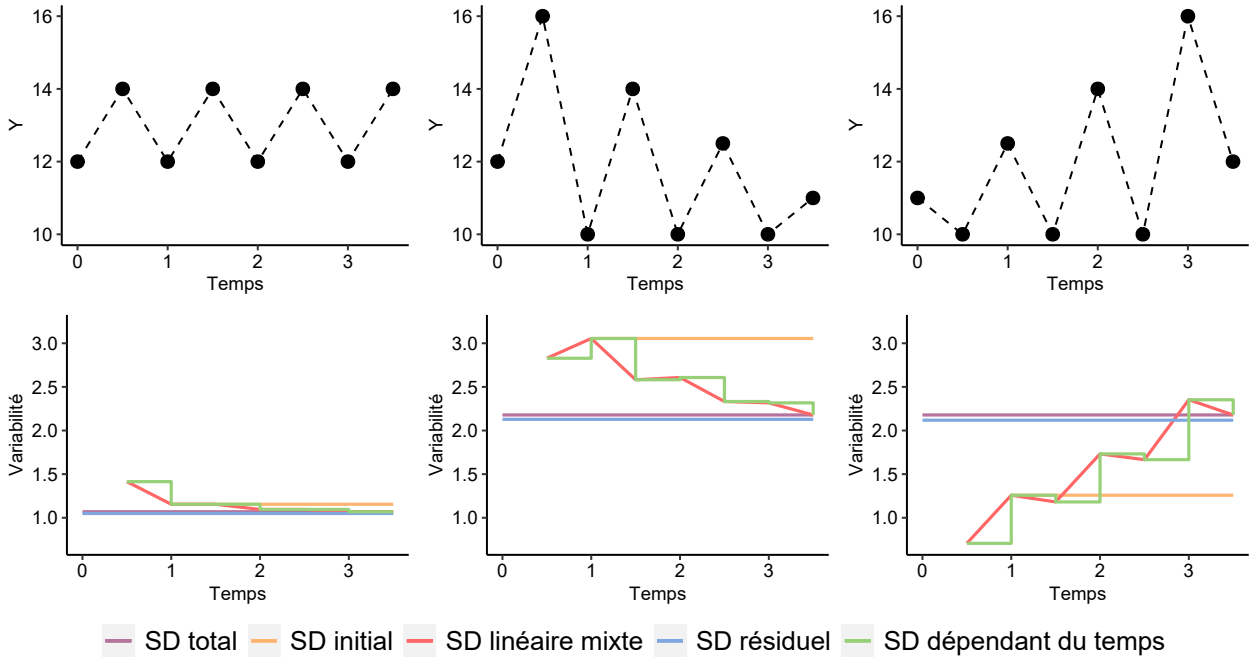

FIGURE I.4 – Représentation de la variabilité de la pression artérielle systolique (en cmHg) au cours du temps en fonction des mesures observées de la pression artérielle de trois patients hypothétiques. Inspiré de de Courson et al. (2021)

Haut : la ligne en pointillés représente la trajectoire de la PA et les ronds les valeurs de PA observées. Bas : variabilité de la PA (VPA) telle que considérées dans les modèles en fonction du temps. Le "SD total" (violet) représente la VPA calculée comme l'écart-type (ET) de toutes les mesures observées. Le "SD initial" (orange) représente la VPA calculée comme l'ET des mesures de PA observées sur la 1ere année (3 mesures). Le "SD linéaire mixte" (rouge) représente la trajectoire observée de VPA calculée à chaque temps de mesure comme l'écart-type (ET) de la mesure actuelle et des mesures passées de PA. Le "SD dépendant du temps" (vert) représente la même trajectoire de VPA mais considérée comme constante entre deux mesures de PA. Le "SD résiduel" (bleu) représente la VPA calculée comme étant la variabilité résiduelle (écarts à la trajectoire moyenne de PA).

I.4 Bases de données

Dans la suite de ce travail, deux bases de données seront utilisées : l'essai clinique PROGRESS (*"perindopril protection against recurrent stroke study"*) sur les maladies cardiovasculaires et cérébrovasculaires (Mac Mahon et al., 2001) ainsi que la cohorte des Trois-Cités, en population générale s'intéressant à la démence (3C Study Group, 2003).

I.4.1 Essai clinique PROGRESS

L'essai clinique PROGRESS est un essai randomisé, contrôlé, en double aveugle, multicentrique et international, ayant pour objectif de déterminer l'effet d'un traitement hypotenseur, le Périndopril, associé ou non à l'Indapamide, pour prévenir la récidive des accidents vasculaires cérébraux (Mac Mahon et al., 2001). Dans cet essai, 7 121 patients ayant des antécédents d'AVC dans les cinq dernières années précédant l'inclusion ont été recrutés entre mai 1995 et novembre 1997. Les critères d'éligibilités n'incluaient pas la PA, bien qu'un traitement spécifique a été recommandé aux patients souffrant d'hypertension non contrôlée avant qu'ils ne participent à l'essai. Les patients ont été suivis dans 172 centres répartis sur 10 pays en Asie, Europe et Océanie.

Avant la randomisation, 7 121 patients ont suivi une période de rodage de quatre semaines pendant laquelle ils ont reçu pendant deux semaines 2 mg de Périndorpil par jour puis 4 mg pendant deux autres semaines. Cela a permis de vérifier la tolérance des patients au traitement. Finalement, 6 105 des 7 121 patients ont été randomisés en double aveugle en suivant un ratio 1 : 1. Ainsi, 3 054 ont été affectés au bras Placebo, arrêtant de prendre du Périndopril, et 3051 au bras Traitement, continuant la prise de 4 mg de Périndopril par jour, combiné ou non à de l'Indapamide.

Le protocole prévoyait cinq visites au cours de la première année puis deux visites annuelles pendant 4.5 ans. À chaque visite, les pressions artérielles (systolique et diastolique) était mesurée deux fois en position assise après cinq minutes de repos.

Dans cette thèse, l'analyse de cet essai clinique a pour objectif d'étudier l'impact de l'évolution de la variance résiduelle individuelle de la PAS sur le risque de récidive de CCVD tout en prenant en compte le risque compétitif de décès.

I.4.2 Cohorte des Trois-Cités

La cohorte des Trois-Cités (3C) est une cohorte prospective française élaborée dans l'objectif d'étudier l'impact des facteurs vasculaires sur le risque de démence et de déclin cognitif (3C Study Group, 2003). Dans cette cohorte, 9 294 personnes ont été recrutées par tirage au sort sur les listes électorales de trois villes françaises : Bordeaux, Dijon et Montpellier. Elles devaient avoir plus de 65 ans et ne pas être institutionnalisées au recrutement qui a eu lieu entre mars 1999 et mars 2001. La cohorte inclue 61% de femmes, 74% des participants ont un niveau d'étude équivalent à l'obtention d'un diplôme supérieur au second degré, 19% sont

porteurs de l'allèle APOE4, bien connu pour être associé à un risque plus important de développer la démence et l'âge moyen à l'entrée dans la cohorte est de 74 ans (écart-type de 6 ans).

Des visites régulières, à 2, 4, 7, 10, 12, 14 et 17 ans de suivi, étaient réalisées à domicile ou dans des centres d'examens afin de faire passer des tests coginitifs aux patients mais aussi récolter des informations de santé. Lors de ces visites, les pressions artérielles (diastoliques et systoliques) étaient mesurées deux ou trois fois, au repos, en position assise. Le dépistage du diagnostic de la démence était réalisé à chaque visite par un neuropsychologue selon les critères du DSM-IV. Les individus suspectés d'être atteints de démence étaient ensuite soumis à un examen clinique auprès d'un neurologue. Pour finir, un panel indépendant d'experts neurologues confirmait éventuellement le diagnostic et l'étiologie de la démence. Les âges de décès ont ensuite été récupérés par information des médecins ou par croisement avec les registres de décès.

L'objectif de l'analyse de 3C dans cette thèse consiste à étudier et comparer l'impact de la variabilité inter-visites et de la variabilité intra-visite de la PAS sur le risque de démence et de décès tout en traitant la censure par intervalle présente dans de telles données.

I.5 Objectifs et challenges méthodologiques

À la vue des méthodes jusqu'ici utilisées, l'objectif de cette thèse était de développer de nouveaux modèles conjoints permettant de modéliser rigoureusement le lien entre la variabilité de la PA et le risque d'événements de santé en traitant certaines difficultés méthodologiques.

Premièrement, le modèle proposé par Gao et al. (2011) et Barrett et al. (2019) (discuté en section I.3.2.3 et détaillé en section II.4.2) repose sur des hypothèses paramétriques trop restrictives. Pour commencer, il ne permet pas d'ajuster le risque sur la valeur courante du marqueur. En effet, seuls les effets aléatoires, constants dans le temps, sont inclus comme covariables dans le modèle à risques proportionnels. Or, la PA étant un facteur de risque connu des événements considérés, il apparaît nécessaire d'ajuster le modèle de survie sur la valeur courante de la PA afin d'évaluer si la variabilité de la PA est un prédicteur indépendant de la PA. D'autre part, les hypothèses concernant le risque de base du modèle à risques proportionnels sont peu flexibles (fonction de type Weibull ou constante par morceaux). Nous proposerons une modélisation permettant d'ajuster sur la valeur courante de la PA et la valeur courante de la pente, tout en offrant la possibilité de définir un risque de base plus

flexible basé sur des B-splines.

Deuxièmement, il est intéressant de considérer une variance résiduelle pouvant dépendre du temps et de diverses covariables, ou encore qu'elle se définisse par différentes sources de variabilité. Dans le modèle conjoint proposé, la variance résiduelle est supposée constante dans le temps, alors qu'il est cliniquement plus probable qu'elle varie. Par ailleurs, si la variabilité de la PA est identifiée comme un facteur de risque, il est alors pertinent d'examiner quelles sont les covariables qui l'influencent afin d'améliorer la prise en charge et le suivi des patients. En outre, la pression artérielle étant en général mesurée plusieurs fois au cours d'une même visite, il apparaît cliniquement pertinent de distinguer et d'étudier l'impact respectif de la variance résiduelle inter-visites (à long terme) et de la variance intra-visite (à court terme) sur le risque d'événements. En effet, s'il advient que la variabilité intra-visite est un aussi bon prédicteur que la variabilité inter-visites, cela pourrait être intéressant car sa mesure est plus aisée.

Troisièmement, il est nécessaire d'étendre la partie survie du modèle conjoint proposé afin de prendre en compte des schémas d'observation complexes : troncature à gauche, risques compétitifs ou semi-compétitifs, et censure par intervalle. En effet, lorsque l'âge est utilisé comme échelle de temps, il arrive bien souvent que les individus entre dans l'étude à des âges différents sous condition de ne pas avoir subi l'événement avant cet âge. Il est alors nécessaire de prendre en compte cette sélection appelée la troncature à gauche, pour ne pas biaiser les estimations. Par ailleurs, l'incidence de la démence et des CCVD augmente fortement avec l'âge. La population la plus à risque de tels événements de santé est donc également plus à risque de décéder. De plus, le décès partage plusieurs facteurs de risques avec la démence et les CCVD. Ainsi, si la corrélation entre le décès et ces événements n'est pas prise en compte, l'estimation de l'effet d'un facteur de risque sur le risque de démence ou de CCVD peut être biaisée. C'est pourquoi il est nécessaire de considérer le décès comme un risque compétitif pour de tels événements : on s'intéresse alors à la survenue du premier événement entre un CCVD et le décès.
Ensuite, le décès peut également être considéré comme un risque semi-compétitif, puisque l'événement d'intérêt ne peut pas survenir après le décès alors que des patients souffrant de démence ou ayant fait un CCVD non mortel peuvent ensuite décéder. Il peut donc être pertinent d'analyser les trois transitions possibles.
Enfin, dans les études de cohortes, l'âge de survenue de la démence n'est pas précisément connue puisque celle-ci n'est diagnostiquée qu'aux temps de visites planifiés. Cela introduit une incertitude sur le temps exact de survenue de la démence, qui se situe entre la dernière

visite où l'individu a été vu sans symptômes et la visite de diagnostic. L'âge de démence est donc censuré par intervalle. De plus, un patient peut développer la démence et décéder entre deux visites, sans que la démence n'ait pu être diagnostiquée. Par conséquence, quand la censure par intervalle n'est par rigoureusement prise en compte, le risque de démence peut être sous-estimé.

I.6 Plan

Le chapitre II présente l'état de l'art des différentes modélisations et approches sur lesquelles s'appuient les développements proposés dans ce travail de thèse. Tout d'abord, nous présentons les modèles linéaires mixtes avec variance résiduelle individuelle. Ensuite, nous décrivons les modèles permettant de traiter les données de survie, et notamment des schémas de données complexes, telles que les risques compétitifs et la censure par intervalle. Enfin nous présentons les modèles conjoints pour données de survie et données longitudinales, puis leurs extensions avec variance résiduelle individuelle.

Le chapitre III présente un modèle conjoint pour étudier l'impact de la variabilité courante de la pression artérielle sur le risque d'événements tout en prenant en compte un risque compétitif. Ce modèle combine un modèle linéaire mixte avec une variance hétérogène et deux modèles à risques proportionnels afin de traiter les risques compétitifs. La variabilité résiduelle hétérogène entre les individus peut dépendre du temps ou de toute autre covariable. Les deux fonctions de risque peuvent, quant à elles, être ajustées sur la variabilité courante du marqueur mais aussi sur la valeur courante ou la pente du marqueur. Des simulations sont réalisées afin de valider la procédure d'estimation du modèle. Le modèle est ensuite appliqué aux données de l'essai clinique PROGRESS afin d'étudier l'impact de la variabilité de la PAS sur le risque de CCVD tout en prenant en compte le risque compétitif de décès. De plus, les capacités prédictives du modèle sont comparées à un modèle conjoint classique.

Le chapitre IV présente un deuxième modèle conjoint incluant une variabilité résiduelle hétérogène permettant cette fois-ci de distinguer les variabilités inter-visites et intra-visite. Le modèle longitudinal est combiné à un modèle *illness-death* pour estimer le risque de démence, tout en considérant le décès comme un événement semi-compétitif et la censure par intervalle. Des simulations sont également réalisées afin de valider la procédure d'estimation. Enfin, le modèle est appliqué à la cohorte 3C afin d'évaluer l'impact des variabilités inter et intra-visites de la PAS sur le risque de démence et de décès, tout en ajustant sur la pente et la valeur courante de la PAS.

Le chapitre V présente le package `R` développé permettant aux utilisateurs d'estimer différents modèles conjoints issus de ce travail. Il peut gérer les différentes modélisations

de la variabilité résiduelle ainsi que un ou deux événements (semi-)compétitifs, la censure par intervalle ou encore l'entrée retardée. Ce package propose également des outils pour l'évaluation graphique de l'ajustement du modèle aux données et la prédiction dynamique des risques d'événements de santé, basés sur les estimations du modèle. Ce chapitre rappelle les notations, la procédure d'estimation et les évaluations des modèles présentés dans les chapitres III et IV. Il a l'avantage de généraliser la présentation en mettant en évidence toutes les variantes des modèles conjoints *location-scale*, définis à la fois par le choix du modèle linéaire mixte *location-scale* et la nature du modèle de survie.

Finalement, le chapitre VI clôture ce manuscrit via une discussion générale résumant les avantages et les limites des méthodes proposées et présentant les perspectives possibles.

Chapitre II

Etat de l'art

Sommaire

Dans l'étude des données de santé, deux intérêts principaux concernent d'une part la survenue d'événements de santé et d'autre part l'évolution de biomarqueurs. Pour cela, que ce soit dans les essais cliniques ou en routine avec les entrepôts de données, les données sont collectées au cours du suivi d'un patient. Les données collectées de manière répétée au cours du suivi sont appelées données longitudinales. Elles peuvent être utilisées pour prédire la survenue d'événements de santé. Ce chapitre présente les modèles proposés dans la littérature pour traiter les problématiques liées aux données longitudinales ainsi qu'aux données de survie. Nous aborderons d'abord le modèle linéaire à effets mixtes et une de ses extensions permettant de prendre en compte une variabilité résiduelle hétérogène entre les individus, puis nous passerons en revue les différentes problématiques liées aux données de survie, comme les risques compétitifs et la censure par intervalle, avant de décrire le modèle conjoint à effets aléatoires partagés. Enfin, nous détaillerons les extensions du modèle conjoint avec variance résiduelle hétérogène.

II.1 Modélisation des données longitudinales

Les données longitudinales, contrairement aux données transversales, sont collectées de manière répétée sur les mêmes individus à différents moments, permettant ainsi de suivre l'évolution au cours du temps. Alors que les données transversales fournissent une vue instantanée à un moment donné, les données longitudinales offrent une perspective dynamique, en mettant en lumière les changements individuels et les trajectoires au sein d'une population. Comparées aux séries temporelles, qui se concentrent généralement sur des mesures répétées d'une seule entité (comme un capteur par exemple) à intervalles réguliers, les données longitudinales concernent plusieurs individus et peuvent inclure des intervalles de temps inégaux entre les mesures pour chaque individu. Cette distinction permet aux données longitudinales de capturer à la fois les variations temporelles et les différences individuelles. Dans les deux bases de données utilisées pour les travaux de cette thèse, la PAS est une variable longitudinale.

Étant donnée la corrélation intra-individuelle due à la répétition des mesures, l'hypothèse d'indépendance entre deux observations d'un même individu n'est pas vérifiée et les modèles de régression classiques ne sont alors pas appropriés. Afin de prendre en compte cette corrélation, Laird et Ware ont développé des modèles à effets mixtes (Laird and Ware, 1982). En fonction de la nature de la variable d'intérêt différents modèles mixtes peuvent être utilisés. En particulier, on utilise un modèle linéaire mixte lorsque la variable d'intérêt est distribuée selon une loi normale et un modèle linéaire généralisé mixte dans le cadre d'une distribution

appartenant à la famille exponentielle.

II.1.1 Modèle linéaire à effets mixtes

II.1.1.1 Définition du modèle

Soit $\mathbb{Y}_N = (Y_1, ..., Y_N)$ un N-échantillon gaussien tel que, pour tout $i \in \{1, ..., N\}$,

$$Y_i = (Y_{i1}, ...Y_{in_i})^\top \in \mathbb{R}^{n_i},$$

où $n_i \in \mathbb{N}^*$ est le nombre de mesures observées pour l'individu i. Ainsi, pour tout $j \in \{1, ..., n_i\}$, Y_{ij} représente la j-ème mesure associée à l'individu i et observée au temps t_{ij}. Cette valeur est une mesure bruitée de la vraie valeur qui n'est pas observée. L'objectif du modèle linéaire à effets mixtes consiste à expliquer Y_{ij} par des facteurs à effets fixes et par des facteurs à effets aléatoires, permettant de prendre en compte la corrélation intra-individuelle et l'hétérogénéité entre les individus.

Pour tout $i \in \{1, ..., N\}$ et pour tout $j \in \{1, ..., n_i\}$, le modèle linéaire à effets mixtes s'écrit :

$$Y_{ij} = \widetilde{Y}_i(t_{ij}) + \epsilon_{ij} = X_{ij}^\top \beta + Z_{ij}^\top b_i + \epsilon_{ij}$$

La valeur observée du marqueur Y_{ij} au temps t_{ij} est définie comme la somme de la vraie valeur de la variable Y pour le sujet i au temps t_{ij}, notée $\widetilde{Y}_i(t_{ij})$, et d'une erreur résiduelle homoscédastique gaussienne ϵ_{ij} telle que $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. On note que $\widetilde{Y}(t)$ est définie en temps continue même si les observations ne sont réalisées qu'en certains temps de mesure. Le vecteur de paramètre $\beta \in \mathbb{R}^p$, $p \in \mathbb{N}^*$ est le vecteur des effets fixes associé au vecteur de covariables X_{ij}, et $b_i \in \mathbb{R}^q$, $q \in \mathbb{N}^*$, est le q-vecteur correspondant aux effets aléatoires et associé au vecteur de covariables Z_{ij}. Les effets fixes permettent de décrire l'évolution moyenne de Y au niveau populationnel. Ils représentent la part expliquée du marqueur par les variables explicatives communes. Les effets aléatoires permettent quand à eux de tenir compte de la corrélation intra-individuelle entre les mesures répétées de l'individu i. Ils sont spécifiques à chaque individu et décrivent l'écart individuel à la trajectoire moyenne de Y. Les b_i sont des vecteurs gaussiens indépendants et identiquement distribués de telle sorte que $b_i \sim \mathcal{N}(0, B)$ avec B la matrice de covariance de dimension $q \times q$. Ils sont supposés indépendants des erreurs de mesures ϵ_{ij}. La forme de B définit la structure de corrélation entre les effets aléatoires. Dans le cas le plus simple, aucune corrélation n'est supposée entre les b_i et B est alors une matrice diagonale de q paramètres. Le plus souvent, la structure de B est indéfinie, donc tous les effets aléatoires sont supposés corrélés, et B est composée de

$q \times (q+1)/2$ paramètres. Soulignons que le vecteur β est commun entre tous les individus tandis que le vecteur b_i varie d'un individu à l'autre.
En général, un intercept est compris dans les p covariables du vecteur X_{ij}. La figure II.1 représente des simulations de différentes trajectoires possibles pour différents individus en fonction du modèle considéré. Dans un cas où ni intercept, ni pente aléatoires ne sont inclus dans le modèle, les droites de régression se superposent. Si un intercept aléatoire est ajouté, alors l'ordonnée à l'origine diffère d'un individu à l'autre. Enfin si une pente aléatoire est ajoutée, le coefficient directeur de la droite varie.

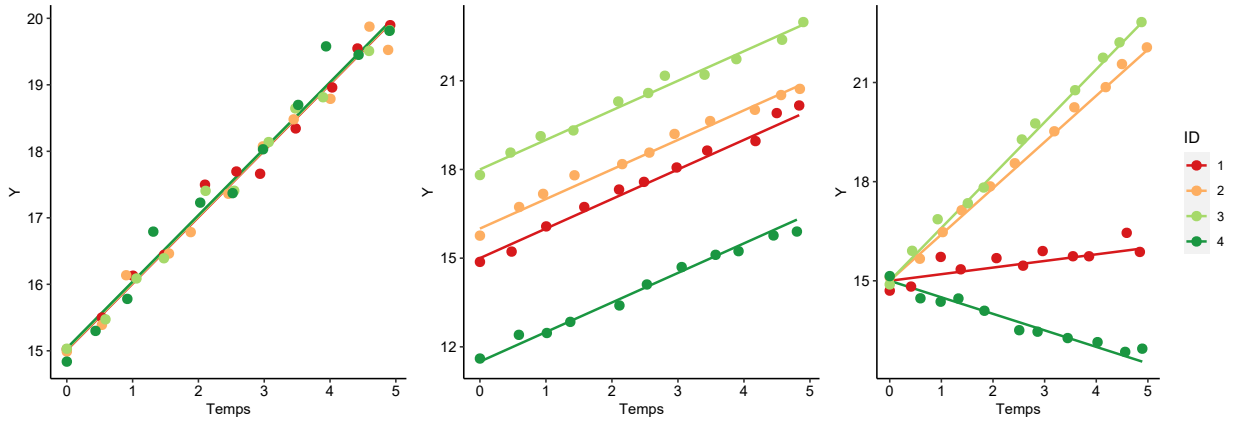

FIGURE II.1 – Illustration des données longitudinales simulées à partir de modèles linéaires à effets mixtes pour 4 individus. Les points correspondent aux réalisations de Y_{ij} à différents temps fixés et les trajectoires au prédicteur linéaire $\widetilde{Y}_i(t)$ défini conditionnellement aux effets aléatoires et en tout temps.

Pour tout $i \in \{1, ..., N\}$, l'espérance et la variance marginales de Y_i sont définies par :

$$\mathbb{E}(Y_i) = X_i\beta \text{ et } \mathbb{V}(Y_i) = V_i = Z_i B Z_i^\top + \sigma^2 I_{n_i} \tag{II.1}$$

avec X_i et Z_i les matrices $n_i \times p$ et $n_i \times q$ dont les lignes sont respectivement $X_{ij}^\top$ et $Z_{ij}^\top$ et I_{n_i} la matrice identité de taille $n_i \times n_i$. L'espérance marginale de Y_i représente la valeur moyenne de Y_i. Elle est commune pour tous les individus qui partagent les mêmes caractéristiques X_i que l'individu i. En outre, on peut définir l'espérance de Y_i conditionnement aux effets aléatoires qui représente l'espérance spécifique à chaque individu et la tendance aléatoire expliquée par les effets aléatoires. Elle est donnée par :

$$\mathbb{E}(Y_i|b_i) = X_i\beta + Z_i b_i.$$

II.1.1.2 Estimation basée sur la vraisemblance

L'approche fréquentiste est très largement majoritaire pour estimer les paramètres du modèle linéaire mixte. Nous décrirons donc seulement l'estimation par le maximum de vraisemblance (ML) et l'estimation par le maximum de vraisemblance restreinte (REML). L'approche bayésienne sera évoquée en section II.3.1.2. pour des modèles plus complexes. Notons $\phi = (\sigma, B)^\top$ l'ensemble des paramètres de variance et de co-variance intervenant dans la définition de V_i (équation II.1) et $\theta = (\beta, \phi)^\top$ le vecteur des paramètres à estimer.

La première approche consiste à maximiser la log-vraisemblance suivante

$$\ell(\theta) = -\frac{1}{2}\sum_{i=1}^{N}\left\{n_i \ln(2\pi) + \ln|V_i(\phi)| + (Y_i - X_i\beta)^\mathsf{T} V_i^{-1}(\phi)(Y_i - X_i\beta)\right\}.$$

Dans le cas où les paramètres ϕ de variance sont connus, il suffit de trouver β qui maximise la vraisemblance précédente. Cela revient à annuler la dérivée première telle que :

$$\frac{\partial \ell(\theta)}{\partial \beta} = \sum_{i=1}^{N} X_i^\mathsf{T} V_i(\phi)^{-1}(Y_i - X_i\beta) = 0.$$

On obtient alors l'expression suivante :

$$\widehat{\beta} = \left(\sum_{i=1}^{N} X_i^\mathsf{T} V_i^{-1}(\phi) X_i\right)^{-1}\left(\sum_{i=1}^{N} X_i^\mathsf{T} V_i^{-1}(\phi) Y_i\right). \tag{II.2}$$

En revanche, dans la majorité des cas ϕ est inconnu. Si une estimation de la matrice V_i est disponible, β peut être estimé par l'expression précédente dans laquelle V_i est remplacé par $\widehat{V}_i$. Pour obtenir une estimation de V_i, la log-vraisemblance $\ell(\theta)$ peut être maximisée pour une valeur donnée de β par des algorithmes d'optimisation numérique (par exemple, Newton-Raphson (Ypma, 1995)). Cela permet d'obtenir une estimation asymptotiquement non biaisée de V_i. Cependant, dans les petites échantillons, l'estimateur de V_i peut être biaisé puisque la perte de degrés de libertés induite par l'estimation des effets fixes n'est pas prise en compte.

Pour résoudre ce problème, il est possible d'utiliser le maximum de vraisemblance restreinte (Harville, 1974). Pour estimer V_i on cherche à éliminer β de la vraisemblance pour qu'elle soit définie qu'en fonction de V_i. Il advient alors qu'il faut maximiser la log-vraisemblance

restreinte suivante :

$$\ell_{REML}(B, \sigma^2) = \ell(\widehat{\beta}, B, \sigma^2) - \frac{1}{2} \log |\sum_{i=1}^{N} X_i^\intercal V_i^{-1} X_i|.$$

avec $\widehat{\beta}$ défini par l'équation II.2. L'estimation de $\widehat{V_i}$ est obtenue en maximisant cette log-vraisemblance modifiée. Dans l'absence de solution analytique il est nécessaire d'utiliser des algorithmes d'optimisation numérique. Les deux algorithmes les plus couramment utilisés sont celui d'*Expectation-Maximization* (EM) (Dempster et al., 1977) et de Newton-Raphson (Ypma, 1995). Les packages les plus utilisés en `R` sont `nlme` (Pinheiro et al., 2024) et `lme4` (Bates et al., 2015).

L'estimation par la méthode REML est à privilégier lorsque le nombre de données est faible ou que le nombre d'effets fixes est grand. Lorsque le nombre d'individus est suffisamment grand, les estimations obtenues par les deux méthodes sont proches. Étant donné que nous travaillons sur de grands échantillons, nous nous focaliserons seulement sur la méthode ML dans la suite de cette thèse.

Par ailleurs, lorsque la variable réponse Y ne suit pas une loi normale, le modèle linéaire à effets mixtes ne peut plus être utilisé et il faut alors recourir au modèle linéaire généralisé à effets mixtes (Molenberghs and Verbeke, 2006). Enfin, lorsque l'on souhaite étudier un lien non linéaire entre la variable Y et les covariables, il est possible d'utiliser des modèles non linéaires à effets mixtes (Lindstrom and Bates, 1990). Dans ce cas, la vraisemblance fait apparaître une intégrale sur les effets aléatoires qui n'a pas de solution analytique et requiert donc une intégration numérique. Cela ne sera pas détaillé ici car ces modèles ne sont pas utilisés dans cette thèse.

II.1.1.3 Prédictions individuelles du marqueur

Pour réaliser des prédictions individuelles de la trajectoire du marqueur ou pour évaluer l'ajustement du modèle aux données, il est nécessaire de prédire la valeur du marqueur Y à différents temps d'observations. La prédiction individuelle pour l'individu i au temps t est obtenue en estimant l'espérance conditionnelle étant donnés les effets aléatoires :

$$\widehat{Y}_i(t) = \widehat{\mathbb{E}}(Y_i(t)|b_i) = X_i(t)\widehat{\beta} + Z_i(t)\widehat{b_i}$$

où les effets aléatoires prédits correspondent aux estimateurs Bayésiens empiriques $\widehat{b_i}$ obtenus conditionnellemnt aux effets fixes par l'espérance de la distribution *a posteriori* des effets

aléatoires sachant les données et les paramètres estimés, $\widehat{\beta}$ et $\widehat{V}_i$:

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

II.1.2 Modèle linéaire mixte avec variance résiduelle hétérogène

Dans le modèle linéaire à effets mixtes décrit précédemment, la variance de l'erreur caractérise la variance intra-sujet et la variance des effets aléatoires la variance inter-sujets. Ces variances sont considérées comme homogènes entre les différents individus. Depuis les années 1980, plusieurs auteurs ont proposé des méthodes permettant de prendre en compte l'hétérogénéité de la variance résiduelle, dans un premier temps dans le cadre de la régression linéaire (Cook and Weisberg, 1983; Aitkin, 1987) puis dans le cadre des modèles linéaires à effets mixtes. Davidian and Giltinan (1993) ont notamment proposé des modèles à effets mixtes non linéaires pour des données longitudinales dans lesquels la variance est liée à la moyenne. Ainsi, la variance est spécifique au sujet et est une fonction du prédicteur linéaire de la moyenne. Foulley et al. (1992), Lin et al. (1997) et Hedeker et al. (2008) ont développé des modèles linéaires à effets mixtes avec une variance résiduelle hétérogène pouvant dépendre de covariables pour des données longitudinales. D'un point de vue général, le modèle s'écrit comme un modèle linéaire mixte classique pour lequel l'erreur résiduelle suit une loi normale centrée avec une variance spécifique au sujet. Les modèles proposés par Foulley et al. (1992) et Lin et al. (1997) supposent une distribution gamma pour caractériser la variance individuelle de l'erreur résiduelle. Cette distribution peut en outre dépendre de covariables mesurées à baseline. Hedeker et al. (2008) assouplissent cette hypothèse de distribution gamma mais ne permettent d'inclure qu'un intercept aléatoire dans le modèle, voir une pente dans une de ses extension. Dans la suite, ce modèle sera nommé "modèle linéaire mixte *location-scale*", pour suivre la terminologie proposée par Hedeker et al. (2008).

II.1.2.1 Définition du modèle linéaire mixte *location-scale*

Cette section se focalise sur le modèle linéaire mixte *location-scale* proposé par Hedeker et al. (2008). Soit Y_{ij} l'observation d'une variable gaussienne pour l'individu $i = 1, ..., N$ au temps t_{ij}, $j = 1, ..., n_i$. Le modèle est formulé par :

$$Y_{ij} = \widetilde{Y}_i(t_{ij}) + \epsilon_{ij} = X_{ij}^\top \beta + b_i + \epsilon_{ij}$$

avec

$$b_i \sim \mathcal{N}(0, \sigma_{b_i}^2) \text{ et } \epsilon_{ij} \sim \mathcal{N}(0, \sigma_{\epsilon_{ij}}^2)$$

où

$$\sigma^2_{b_i} = \exp(U_i^\top \alpha) \text{ et } \sigma^2_{\epsilon_{ij}} = exp(M_{ij}^\top \tau + \omega_i)$$

La fonction exponentielle permet d'assurer la positivité des variances. La variance de b_i dépend des valeurs des covariables indépendantes du temps U_i de l'individu i et peut donc être différente entre les individus. De plus, ce modèle permet de prendre en compte des variables dépendantes du temps, M_{ij}, dans la modélisation de la variance intra-individuelle. Cette variance peut également varier entre les individus, au-delà de la contribution des covariables, grâce à l'ajout d'un effet aléatoire individuel $\omega_i \sim \mathcal{N}(0, \sigma^2_\omega)$. Dans ce cas, la variance $\sigma^2_{\epsilon_{ij}}$ suit une loi log-normale.

Dans ce modèle, b_i est un intercept aléatoire influençant la moyenne (*location*) individuelle et ω_i est un effet aléatoire impactant la variance (*scale*) individuelle, d'où le nom de modèle linaire à effets mixtes *location-scale*. Les deux effets aléatoires sont supposés corrélés tels que $cov(b_i, \omega_i) = \sigma_{b\omega}$.

II.1.2.2 Estimation basée sur la vraisemblance

Pour représenter les effets aléatoires sous leur forme standardisée, les auteurs utilisent la factorisation de Cholesky suivante :

$$\begin{bmatrix} b_i \\ \omega_i \end{bmatrix} = \begin{bmatrix} s_{1i} & 0 \\ s_{2i} & s_{3i} \end{bmatrix} \begin{bmatrix} \theta_{1i} \\ \theta_{2i} \end{bmatrix} = \begin{bmatrix} \sigma_{b_i} & 0 \\ \sigma_{b\omega}/\sigma_{b_i} & \sqrt{\sigma^2_\omega - \sigma^2_{b\omega}/\sigma^2_{b_i}} \end{bmatrix} \begin{bmatrix} \theta_{1i} \\ \theta_{2i} \end{bmatrix},$$

le modèle peut s'écrire alors :

$$Y_{ij} = X_{ij}^\top \beta + s_{1i}\theta_{1i} + \epsilon_{ij}$$

où $s_{1i} = \sigma_{b_i}$ et la variance des erreurs résiduelles ϵ_{ij} s'écrit $\sigma^2_{\epsilon_{ij}} = \exp(M_{ij}^\top \tau + s_{2i}\theta_{1i} + s_{3i}\theta_{2i})$. Les effets aléatoires standardisés θ_{1i} et θ_{2i} suivent alors chacun une loi normale centrée réduite et sont indépendants l'un de l'autre.

En notant Y_i le vecteur contenant les n_i mesures de l'individu i on obtient :

$$Y_i = X_i\beta + \mathbb{1}_{n_i} s_{1i}\theta_{1i} + \exp\frac{1}{2}(M_i\tau + \mathbb{1}_{n_i} s_{2i}\theta_{1i} + \mathbb{1}_{n_i} s_{3i}\theta_{2i})\, e_i$$

avec X_i la $n_i \times p$ matrice de covariables influençant la moyenne de Y_i (*location*), M_i la $n_i \times r$ matrice de covariables influençant la variance intra-individuelle de Y_i (*scale*), 1_{n_i} un vecteur de 1 de taille n_i et ϵ_i l'erreur standardisée, suivant une loi normale centrée réduite. Les Y_i sont

distribués selon des lois normales indépendantes de moyenne $X_i\beta$ et de matrice de covariance $\mathbb{1}_{n_i}\mathbb{1}_{n_i}^\top\sigma_{b_i}^2 + \sigma_{\epsilon_{ij}}I_i$. La densité marginale de Y_i s'écrit alors :

$$h(Y_i) = \int_{\theta_i} f(Y_i|\theta_i)g(\theta_i)d\theta_i$$

avec $f(Y_i|\theta_i)$ la distribution normale des Y_i conditionnelle aux effets aléatoires $\theta_i = (\theta_{1i}, \theta_{2i})^\top$, et $g(\theta)$ la densité normale centrée bivariée. Finalement, le modèle est estimé par maximisation de la log-vraisemblance $log(L) = \sum_{i=1}^{N} log(h(Y_i))$.

Le logiciel `MIXREGLS` proposé par Hedeker and Nordgren (2013) puis remplacé par le logiciel `MixWild` (Dzubur et al., 2020) permettent d'estimer ce type de modèle. Le second logiciel étend le modèle proposé ci-dessus à l'ajout d'une pente aléatoire. Ils sont écrits en `Fortran` et maximisent la vraisemblance en utilisant un algorithme EM et une méthode de Newton-Raphson.

II.2 Modélisation des données de survie

Les données de survie sont utilisées pour analyser le délai avant la survenue d'un événement d'intérêt. L'événement d'intérêt peut être le décès, l'apparition ou la progression d'une maladie. L'objectif consiste à soit, étudier la probabilité de ne pas déclarer un événement jusqu'à un temps t, soit étudier le risque de déclarer l'événement à chaque temps d'intérêt. De plus, il peut être intéressant d'analyser comment certaines variables peuvent influencer cette probabilité.

II.2.1 Notions de base

Définissons T^* une variable aléatoire continue positive ou nulle qui représente la durée de survie avant l'événement d'intérêt. Ce temps peut être défini comme le temps depuis la naissance de l'individu, ou le temps depuis l'entrée dans l'étude de l'individu ou encore le temps depuis une date fixe commune à tous les individus. Par ailleurs, il est possible qu'un individu n'ait pas connu l'événement au moment de sa sortie de l'étude. On dit alors qu'il est censuré à droite. Cela peut arriver si un individu n'a pas subi l'événement avant la fin de l'étude ou s'il la quitte prématurément sans avoir connu l'événement, par exemple s'il choisit de ne plus participer ou s'il est perdu de vue. On dit alors qu'il est perdu de vue. Soit C la variable aléatoire représentant le temps de censure à droite correspondant au délai entre le temps d'inclusion dans l'étude et le moment de censure et T^* le temps de survenue de

l'événement. Soit (T, δ) le couple de variables aléatoires tel que

$$\begin{cases} T = \min(T^*, C) \\ \delta = \mathbb{1}_{T^* \leq C} \end{cases}$$

où T représente la durée réellement observée et δ l'indicateur d'événement. Si l'événement a été réalisé durant la durée d'observation alors $\delta = 1$, sinon $\delta = 0$.

La loi de probabilité de T est caractérisée par cinq fonctions définies sur $\mathbb{R}^+$, toutes liées entre elles. Premièrement, pour $t \in \mathbb{R}^+$, la fonction de survie au temps t, $S(t)$, correspond à la probabilité que l'événement ne se produise pas avant le temps t. Elle est définie pour tout $t \geq 0$ par :

$$S(t) = \begin{cases} 1 \text{ si } t = 0 \\ \mathbb{P}(T > t) \text{ si } t > 0 \end{cases}$$

C'est une fonction décroissante.
Ensuite, la fonction de répartition au temps t, $F(t)$, représente la probabilité d'expérimenter l'événement dans l'intervalle $[0, t]$. Elle est définie pour $t \geq 0$ par :

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

Définissons la fonction de densité $f(t)$ la probabilité de réalisation de l'événement dans l'intervalle $[t, t + \Delta t]$ avec $\Delta t \in \mathbb{R}^+$, pour tout $t \geq 0$:

$$\begin{aligned} f(t) &= \lim_{\Delta t \to 0^+} \frac{\mathbb{P}(t \leq T \leq t + \Delta t)}{\Delta t} \\ &= \lim_{\Delta t \to 0^+} \frac{\mathbb{P}(T \leq t + \Delta t) - \mathbb{P}(T \leq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0^+} \frac{F(t + \Delta t) - F(t)}{\Delta t} \\ &= F'(t) \\ &= -S'(t) \end{aligned}$$

La fonction de risque instantanée au temps t, $\lambda(t)$, représente le risque d'expérimenter l'événement dans l'intervalle $[t, t+\Delta t]$ en sachant que l'individu est encore à risque au temps t. Elle est définie pour tout $t \geq 0$ par :

$$\begin{aligned}
\lambda(t) &= \lim_{\Delta t \to 0^+} \frac{\mathbb{P}(t \leq T \leq t+\Delta t | T \geq t)}{\Delta t} \\
&= \frac{1}{\mathbb{P}(t \leq T)} \lim_{\Delta t \to 0^+} \frac{\mathbb{P}(t \leq T \leq t+\Delta t)}{\Delta t} \\
&= \frac{f(t)}{S(t)} \\
&= -[ln(S(t))]'
\end{aligned}$$

Finalement, la fonction de risque cumulée au temps t, $\Lambda(t)$, représente le cumul des risques instantanés $\lambda(t)$ sur l'intervalle $[0, t]$. Ainsi, pour tout $t \geq 0$:

$$\Lambda(t) = \int_0^t \lambda(u)du = \int_0^t -[ln(S(u))]'du = -ln(S(t)).$$

Les fonctions précédemment définies peuvent être estimées par des méthodes paramétriques ou non paramétriques. Parmi les estimateurs non-paramétriques les plus utilisées on retrouve l'estimateur de Kaplan-Meier (Kaplan and Meier, 1958) afin d'estimer la fonction de survie :

$$\widehat{S}(t) = \prod_{j, T_j^* \leq t} (1 - \frac{d_j}{n_j})$$

et sa variance (la formule de Greenwood) :

$$\widehat{\sigma_t}^2 = [\widehat{S}(t)]^2 \sum_{j, T_j^* \leq t} \frac{d_j}{n_j(n_j - d_j)}$$

avec T_j^* les temps d'événements observés, n_j le nombre d'individus à risque en T_j^* et d_j le nombre d'individus subissant l'événement en T_j^*. Il s'agit d'une fonction décroissante en escalier avec un saut à chaque temps d'événement T_j^*. Elle est donc constante entre deux temps d'événements.

L'estimateur de Nelson-Aalen permet d'obtenir quant à lui une estimation de la fonction de risque cumulée $\Lambda(t)$ (Nelson, 1969; Aalen, 1976). Il est définit par :

$$\widehat{\Lambda}(t) = \sum_{j, T_j^* \leq t} \frac{d_j}{n_j}$$

II.2.2 Modèle à risques proportionnels : le modèle de Cox

Le modèle de Cox (Cox, 1972) a été développé afin d'évaluer l'impact de différentes covariables sur la survie des individus et donc d'étudier les facteurs de risque associés à la survenue d'un événement. Ce modèle repose sur trois hypothèses :

- l'indépendance entre les mesures ;
- la proportionnalité des risques, i.e. l'effet des variables explicatives est supposé constant au cours du temps ;
- la log-linéarité des variables explicatives.

La fonction de risque $\lambda(t)$ est modélisée par :

$$\lambda(t|W) = \lambda_0(t)\exp(W\gamma) \tag{II.3}$$

où $W = (W^{(1)\top}, ..., W^{(p)\top})$ est une matrice contenant les p vecteurs de covariables considérées, et $\gamma \in \mathbb{R}^p$ est le vecteur de dimension p contenant les coefficients de régression associés. La fonction $\lambda_0(t)$ est la fonction de risque de base associée au temps t, c'est-à-dire, le risque instantané de survenue de l'événement pour les individus dont les covariables sont nulles. Si la forme de λ_0 est connue, par exemple si c'est une loi de Weibull, le modèle est paramétrique et est estimé en maximisant la vraisemblance définie par :

$$\mathcal{L}(\gamma; T, \delta, W) = \prod_{i=1}^{N} \Big\{ \lambda_0(T_i)\exp(W_i^\top \gamma) \Big\}^{\delta_i} \exp\left(- \int_0^{T_i} \lambda_0(t)\exp(W_i^\top \gamma) dt \right)$$

Si aucune hypothèse n'est faite sur la forme de cette fonction, alors le modèle est semi-paramétrique. Seuls les coefficients γ sont donc estimés par la méthode de maximisation de la vraisemblance partielle qui s'écrit

$$\mathcal{L}(\gamma; T, \delta, W) = \prod_{i=1}^{k} \frac{\exp(W_{(i)}^\mathsf{T}\gamma)}{\sum_{l:T_l \geq t_{(i)}} \exp(W_{(l)}^\mathsf{T}\gamma)}$$

avec $t_{(1)} < ... < t_{(k)}$ les temps observés ordonnés et $(1), ..., (k)$ les individus correspondant. Cette vraisemblance partielle est valide seulement lorsqu'il n'y a qu'un seul événement observé à chaque temps. Lorsque ce n'est pas le cas et que certains individus subissent l'événement en même temps, plusieurs approximations ont été proposées (Therneau and Grambsch, 2000), dont notamment la plus utilisée, l'approximation de Breslow. Cette maximisation est généralement réalisée en utilisant des méthodes itératives.

L'intérêt du modèle de Cox repose dans sa facilité d'interprétation. En effet, il est facile

de calculer le changement du risque pour l'augmentation d'une unité d'une variable $W^{(l)}$, appelé risque relatif ou *hazard ratio (HR)* :

$$HR = \frac{\lambda(t|W^{(l)} = w + 1)}{\lambda(t|W^{(l)} = w)} = \exp(\gamma_l)$$

Ainsi, une augmentation d'une unité de la variable $W^{(l)}$, multiplie le risque d'événement par $\exp(\gamma_l)$.

Dans certains cas, l'évolution de variables explicatives au cours du suivi peut avoir un impact sur la survenue de l'événement d'intérêt. Il est alors nécessaire d'utiliser une des extensions du modèle classique de Cox, le modèle de Cox à variables dépendantes du temps (Houwelingen and Putter, 2012).

II.2.3 Troncature à gauche

Dans les études de cohortes, les données sont fréquemment tronquées à gauche. La troncature à gauche survient dès que le temps d'étude ne correspond pas au temps depuis l'inclusion et que les sujets sont inclus seulement s'ils n'ont pas fait l'événement avant l'inclusion. C'est par exemple le cas dans la cohorte 3C si l'on étudie l'âge jusqu'à la démence ou au décès puisque les individus étaient inclus à partir de 65 ans. Ainsi, seuls les sujets ayant survécu sans démence au moment de l'inclusion sont inclus dans l'échantillon d'analyse. En présence de troncature à gauche, l'estimation doit prendre en compte le conditionnement dû au fait que les individus n'ont pas développé l'événement jusqu'à leur inclusion. En semi-paramétrique, cela consiste à n'inclure le sujet i dans le sous-ensemble considéré pour calculer le risque au temps $t_{(i)}$ seulement si $t_{(i)} \geq T_{0l}$. Pour tous les modèles paramétriques, il suffit de diviser la contribution individuelle à la vraisemblance par la probabilité de ne pas avoir subi l'événement avant l'entrée dans l'étude :

$$\mathcal{L}(\gamma; T, \delta, W) = \prod_{i=1}^{N} \frac{\{\lambda_0(T_i)\exp(W_i^\top\gamma)\}^{\delta_i}\exp\Big(-\int_0^{T_i}\lambda_0(t)\exp(W_i^\top\gamma)dt\Big)}{\exp\Big(-\int_0^{T_{0i}}\lambda_0(t)\exp(W_i^\top\gamma)dt\Big)}$$

où T_{0i} correspond au temps d'entrée dans l'étude de l'individu i.

II.2.4 Risques compétitifs

Dans l'analyse des données de survie, on considère que chaque sujet est à risque de développer un événement d'intérêt jusqu'à ce qu'il le subisse. Cependant, le sujet peut parfois subir

un autre événement empêchant alors l'observation de l'événement d'intérêt. Ce deuxième événement est alors appelé événement compétitif. C'est notamment le cas du décès dans les bases de données présentées en section I.4, où lorsqu'il survient, le CCVD ou la démence ne peuvent plus être observés. Ainsi, ne pas tenir compte des événements compétitifs peut conduire à des biais dans l'estimation de la survie puisque les individus censurés à cause de la survenue d'un événements compétitifs sont considérés comme encore à risque pour l'événement d'intérêt. De plus, dans le modèle de Cox tel que présenté en équation (II.3), la censure est supposée être indépendante de l'événement d'intérêt. Or, en présence de risques compétitifs cette hypothèse est violée, car les événements compétitifs peuvent modifier la probabilité de l'événement d'intérêt. Cela est donc particulièrement important à prendre en compte, d'autant plus que le risque de ces événements d'intérêt partagent de nombreux facteurs de risques. Un modèle à K risques compétitifs peut être représenté graphiquement avec un état initial correspondant à l'état d'entrée dans l'étude, donc libre de tous les événements considérés, et K risques possibles (Figure II.2). Il existe deux types d'approches pour traiter les risques compétitifs (Putter et al., 2007) : l'approche de Fine and Gray (Fine and Gray, 1999) et les modèles causes-spécifiques.

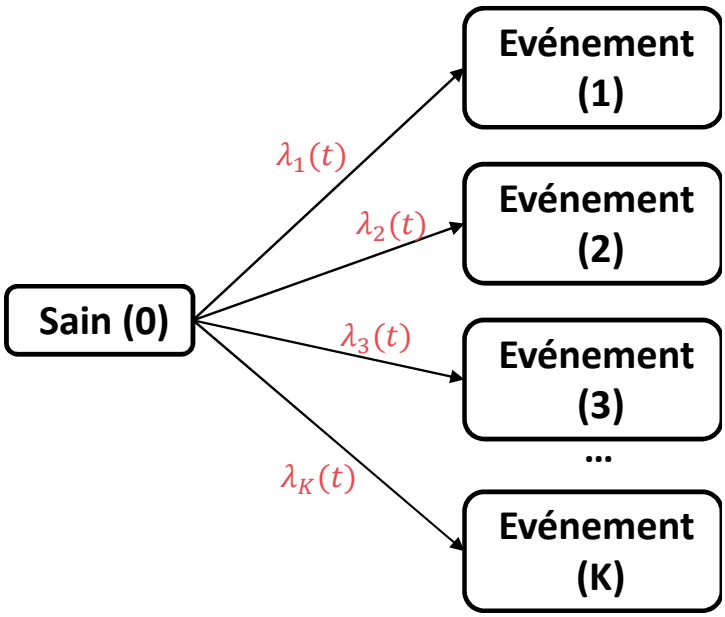

FIGURE II.2 – Modèle à risques compétitifs.

Soient $K \geq 2$ le nombre d'événements compétitifs, $T = \min(T_1^*, ..., T_K^*, C)$ le temps entre l'inclusion et l'occurrence du premier événement, à défaut le temps de censure et $\delta \in \{0, ..., K\}$ l'indicateur d'événement tel que $\delta = k$ si l'individu à subi l'événement compétitif k et $\delta = 0$

si aucun événement d'intérêt n'a été observé. Notons $\lambda_k(t)$ la fonction de risque instantanée au temps t spécifique à l'événement k, $k \in \{1, ..., K\}$. Elle s'écrit

$$\lambda_k(t) = \lim_{\Delta t \to O^+} \frac{\mathbb{P}(t \leq T \leq t + \Delta t, \delta = k | T \geq t)}{\Delta t}$$

Il est aisé d'obtenir ainsi la fonction de survie au temps t, interprétée comme la probabilité de ne subir aucun événement avant t :

$$S(t) = \exp\left(-\sum_{k=1}^{K} \int_0^t \lambda_k(u)du\right) = \exp\left(-\sum_{k=1}^{K} \Lambda_k(t)\right)$$

avec $\Lambda_k(t) = \int_0^t \lambda_k(u)du$ la fonction de risque cumulée cause-spécifique qui peut être approchée par l'estimateur de Aalen-Johansen (Aalen and Johansen, 1978), noté $\widehat{P_{0k}}$, et défini par :

$$\widehat{P_{0k}}(t) = \sum_{t_j \leq t} \widehat{P_{00}}(t_{j-1}) \frac{d_{0kj}}{r_{0j}} \text{ avec } \widehat{P_{00}}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_{0j}}{r_{0j}}\right)$$

où $t_1 < t_2 < ... < t_j < ...$ sont l'ensemble des temps d'événements observés, d_{0kj} est le nombre d'individus ayant fait l'événement k au temps t_j, $d_{0j} = \sum_{k=1}^{K} d_{0kj}$ est le nombre total d'événements survenu au temps t_j et r_{0j} donne le nombre d'individus encore à risques juste avant le temps t_j. Enfin, on peut obtenir la fonction d'incidence cumulée de l'événement k, interprétée comme la probabilité de subir l'événement k avant le temps t :

$$I_k(t) = \mathbb{P}(T \leq t, \delta = k) = \int_0^t \lambda_k(u) S(u) du$$

Afin d'estimer l'effet de covariables dans le cadre des risques compétitifs, Fine and Gray (Fine and Gray, 1999) ont proposé un modèle semi-paramétrique à risques proportionnels basé sur la fonction d'incidence cumulée. Pour cela, une fonction de risque, appelée fonction de risque de sous-répartition, et associée à la fonction d'incidence cumulée précédemment présentée, est définie par :

$$\overline{\lambda}_k(t|X) = -\frac{d \log(1 - I_k(t|X))}{dt} = \overline{\lambda}_{0k}(t) \exp(W_k^\top \gamma_k).$$

L'estimation des paramètres se fait sur le même principe que pour l'estimation d'un modèle de Cox. Cependant, il est important de noter que ce risque de sous-répartition ne correspond pas au risque cause-spécifique. En effet, pour le risque cause-spécifique, l'échantillon d'individus décroît à chaque fois qu'un événement survient. Alors que dans ce modèle, les individus ayant subi un événement l différent de l'événement k sont toujours comptabilisés dans l'effectif à

risque. Par exemple, si le décès fait parti des événements compétitifs considérés, un individu décédé sera toujours considéré à risque pour les autres événements après sa date de décès. De plus, cela engendre une considération de l'effet d'une variable sur l'événement d'intérêt, y compris lorsque cet effet est indirect et est le résultat de l'effet de cette variable sur un événement compétitif. Cela rend donc complexe l'interprétation de la quantité calculée ainsi que des coefficients. Pour ces raisons, nous ne retiendrons pas cette approche pour la suite.

La deuxième approche consiste à définir les transitions de l'état initial vers chaque événement $k \in (1, ..., K)$ par un modèle à risques proportionnels spécifiques :

$$\lambda_{ik}(t) = \lambda_{0k}(t) \exp(W_{ik}^{\top} \gamma_k)$$

avec $\lambda_{0k}(t)$ le risque de base pour la transition vers l'événement k, γ_k le vecteur des paramètres de régression associé au vecteur de covariables W_{ik} et δ_{ik} les indicateurs d'événements. La vraisemblance s'écrit alors :

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} \lambda_{ik}(T_i; \theta)^{\delta_{ik}} S_i(T_i; \theta)$$

avec θ le vecteur des paramètres de régression et $S_i(t)$ la fonction de survie globale définie par $S_i(t) = \exp(-\int_0^t \sum_{l=1}^{K} \lambda_{il}(s)ds)$. L'estimation peut se faire par maximisation de la vraisemblance partielle avec une fonction de risque de base non spécifiée.

II.2.5 Risques semi-compétitifs

Lorsque deux événements sont étudiés de façon compétitive mais dont l'un peut être observé avant ou après l'observation du premier et la transition entre les deux événements nous intéresse, cela s'appelle des risques semi-compétitifs. Ces transitions peuvent être représentées par un modèle multi-état (Figure II.3). Par exemple, dans l'étude de la démence, ou toute autre maladie, chez les personnes âgées, le risque de décéder est également très important et peut survenir avant ou après la maladie. Il est alors utile de modéliser trois transitions :

- de sain à malade,
- de malade à décédé,
- de sain à décédé.

L'estimateur de Aalen-Johansen étendu aux modèles multi-états peut permettre d'estimer ces transitions.

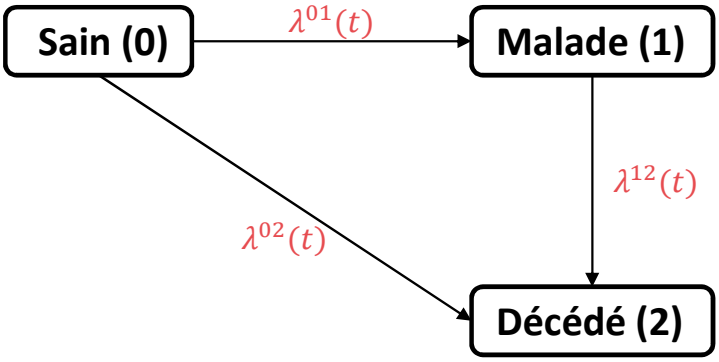

FIGURE II.3 – Modèle Illness-Death.

II.2.6 Censure par intervalle

Dans les sections précédentes, le temps de survenue de l'événement était supposé connu ou censuré à droite. Or, tous les temps d'événements ne peuvent pas être enregistrés au moment exact où ils apparaissent. C'est notamment le cas dans l'étude de la démence où le moment de sa survenue n'est pas connue précisément. En effet, le diagnostic de la démence n'est réalisé qu'aux temps de visites. Ainsi, en notant L_i la dernière visite avant la survenue de la démence, et R_i la première visite après son occurrence, le temps de survenue de la démence est compris dans l'intervalle $]L_i, R_i]$. On dit alors que l'événement est censuré par intervalle. Cela conduit donc à une incertitude sur le temps d'événement qu'il est nécessaire de prendre en compte lors de l'estimation. Des méthodes naïves consistent à imputer le temps de survenue de la maladie par le milieu de l'intervalle de censure (Law and Brookmeyer, 1992; Helmer et al., 2001) ou par le temps de diagnostic (Hazzouri et al., 2011). Cependant, ces approximations peuvent mener à des estimateurs biaisés (Freitag et al., 2006; Odell et al., 1992) et des écart-types potentiellement sous-estimés (Kim, 2003). Finkelstein (1986) a proposé une extension du modèle de Cox à risques proportionnels en considérant le temps de façon discrète. La contribution à la vraisemblance est alors définie par :

$$\mathcal{L}_i = \mathbb{P}(T_i > L_i|W_i) - \mathbb{P}(T_i > R_i|W_i)$$

Elle est maximisée par l'algorithme EM. Ce modèle a par la suite été étendu par Alioum and Commenges (1996) pour prendre en compte des données tronquées à gauche et un temps continu. La vraisemblance est alors calculée conditionnellement au temps d'entrée dans l'étude. Joly et al. (1998) ont proposé de modéliser la fonction de risque instantanée de façon paramétrique ou semi-paramétrique. Dans le cadre d'une estimation semi-paramétriques, des

M-splines sont utilisées et la vraisemblance est pénalisée. L'idée consiste à utiliser un grand nombre de noeuds pour avoir une fonction lisse et s'approchant d'une méthode non paramétrique.

Cependant, en présence d'un risque semi-compétitif tel que le décès, la censure par intervalle est davantage problématique puisque le décès peut survenir avant le diagnostic de la démence. Dans ce cas, R_i n'est pas défini. Ainsi, le statut de démence du sujet est inconnu au moment du décès. Dans le cadre de modèle cause-spécifique pour risques compétitifs, le temps de survenue de la démence est censuré au temps de décès ce qui peut biaiser les estimations.

Afin d'étudier les facteurs de risque de démence en tenant compte du risque semi-compétitif de décès et de la censure par intervalle de la démence, Joly et al. (2002) ont proposé un modèle multi-état comprenant trois états : Sain, Malade et Décès (Figure II.3). Avec L_i le temps de dernière visite sans démence, R_i le temps de visite de diagnostic (non défini si pas de diagnostic) et T_i le temps de décès ou de dernière information sur le statut vital (peut être postérieur au temps de la dernière visite), différentes trajectoires entre les trois états sont alors possibles (Figure II.4) :

- Cas 1 et 2 : l'individu est sain et vivant jusqu'en L_i, il est diagnostiqué dément en R_i, il reste vivant jusqu'en T_i et décède (cas 1) ou est censuré (cas 2) en T_i ;
- Cas 3 et 4 : l'individu est sain et vivant jusqu'en T_i, ($T_i = L_i$) et décède (cas 3) ou est censuré (cas 4) en T_i ;
- Cas 5 et 6 : l'individu est sain à la fin du suivi $L_i = R_i < T_i$, reste vivant jusqu'en T_i et décède en T_i (cas 5) ou est censuré (cas 6) en T_i ;
- Cas 7 et 8 : l'individu développe et est diagnostiqué dément en $L_i = R_i$ (temps exact de survenue de la démence), reste vivant jusqu'en T_i et décède et décède en T_i (cas 7) ou est censuré (cas 8) en T_i. Ce dernier cas n'existe pas en pratique dans l'étude de la démence mais pourrait exister dans le cas d'étude portant sur d'autres maladies.

Dans ce modèle l'intensité de transition d'un état k vers un état l est modélisée par un modèle à risques proportionnels :

$$\lambda_i^{kl}(t) = \lambda_0^{kl}(t) \exp(W_i^{kl\top} \gamma^{kl}) \text{ avec } (k,l) \in \{(0,1),(0,2),(1,2)\}.$$

La vraisemblance du modèle est adaptée pour prendre en compte la censure par intervalle en considérant la possibilité d'une transition vers la démence non-observée pour tous les sujets vus sans démence à leur dernière visite L_i mais avec une information sur leur statut vital

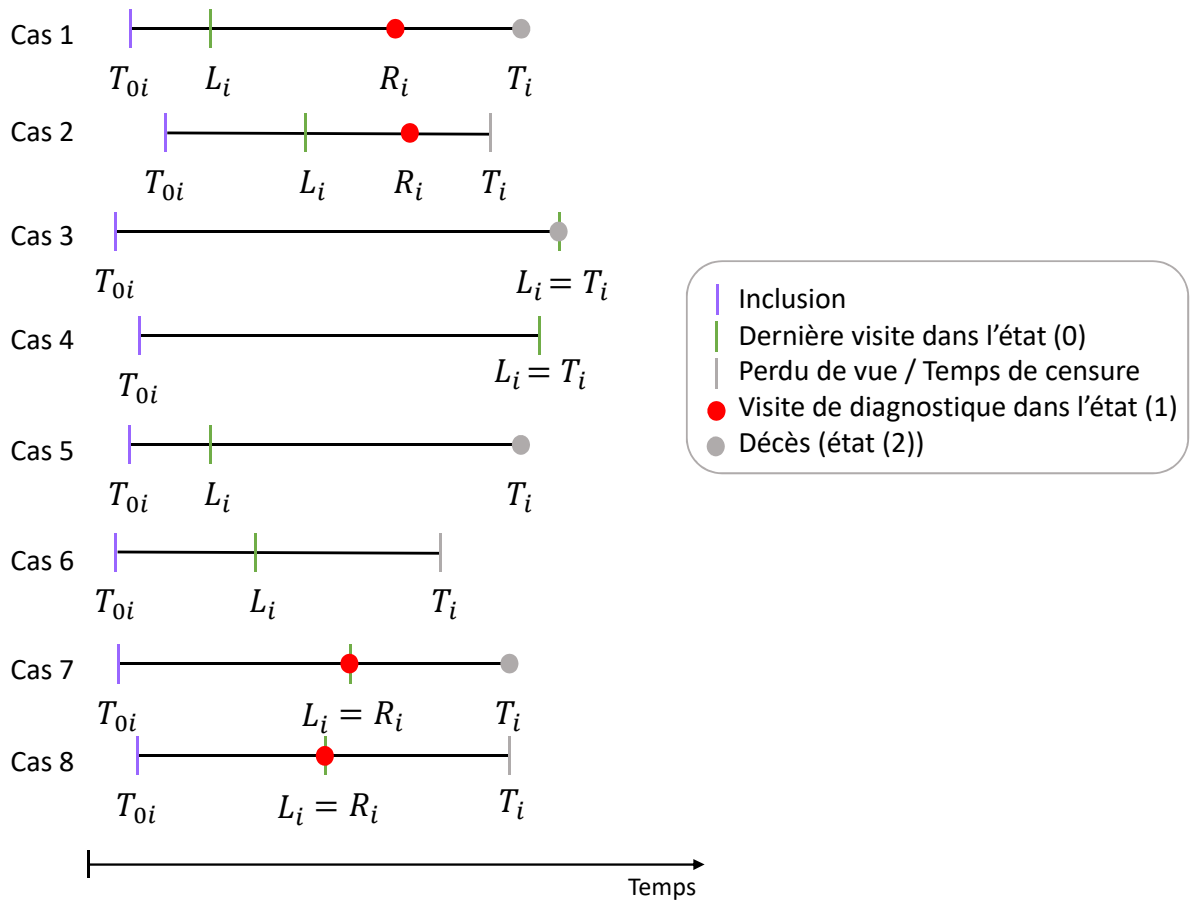

FIGURE II.4 – Les différents schémas possibles entre démence et décès.

collecté ultérieurement. La contribution individuelle à la vraisemblance s'écrit alors :

$$\begin{aligned}\mathcal{L}_i &= \delta_i^{Dem}\lambda_i^{12}(T_i)^{\delta_i^{Death}}\int_{L_i}^{R_i} p_{00i}(0,u)\lambda_i^{01}(u)p_{11i}(u,T_i)du \\ &\quad +(1-\delta_i^{Dem})\left(p_{00i}(0,T_i)\lambda_i^{02}(T_i)^{\delta_i^{Death}} + \lambda_i^{12}(T_i)^{\delta_i^{Death}}\int_{L_i}^{R_i} p_{00i}(0,u)\lambda_i^{01}(u)p_{11i}(u,T_i)du\right)\end{aligned}$$

avec $\delta_i^{Dem} = 1$ si l'individu a été diagnostiqué dément et 0 sinon, $\delta_i^{Death} = 1$ si le sujet est décédé avant la fin du suivi et 0 sinon, $p_{00i}(s,t)$ la probabilité que l'individu soit resté vivant sans démence entre les temps s et $s+t$ et $p_{11i}(s,t)$ la probabilité que l'individu soit resté vivant et dément entre les temps s et $s+t$.

Les intensités de transitions peuvent être modélisées soit par une fonction Weibull soit avec des fonctions M-splines pour plus de flexibilité. La vraisemblance est ensuite maximisée par l'algorithme de Marquardt-Levenberg (Levenberg, 1944; Marquardt, 1963). Si les intensités

de transitions sont modélisées par des M-splines, la vraisemblance est pénalisée par la somme des normes au carré des secondes dérivées des intensités (Joly et al., 2002).

A travers une étude de simulations sur des données de survie censurées par intervalle et avec un risque semi-compétitif de décès, Leffondré et al. (2013) ont montré que les estimations du modèle de Cox en imputant le temps de survenue de la maladie par le milieu de l'intervalle sont biaisées, et ce d'autant plus lorsque les variables sont liées à l'événement d'intérêt. Le modèle illness-death tenant compte de la censure par intervalle proposé par Joly et al. (2002) permet de corriger ces biais.

Dans l'étude de la survenue de la démence et du décès, deux hypothèses peuvent être faites sur la transition de la démence au décès : l'hypothèse markovienne ou semi-markovienne. Sous l'hypothèse de Markov, on suppose que l'intensité de transition entre la démence et le décès dépend de l'âge t et est modélisée par :

$$\lambda_i^{12}(t) = \lambda_0^{12}(t) \exp(W^{12\top}\gamma^{12}).$$

En revanche, sous l'hypothèse de Semi-Markov, l'intensité de transition vers le décès, parmi les individus déments, est supposé dépendre du temps passé en démence plutôt que de l'âge, définissant alors la fonction de transition entre démence et décès par :

$$\lambda_i^{12}(t, T_i^{dem}) = \lambda_i^{12}(t - T_i^{dem}) = \lambda_0^{12}(t - T_i^{dem}) \exp(W^{12\top}\gamma^{12})$$

avec T_i^{dem} l'âge à l'apparition de la démence. Rouanet et al. (2016) ont comparé les deux hypothèses et mis en avant que l'hypothèse Markovienne permettait d'obtenir de meilleurs résultats dans le cadre de l'étude de la démence.

II.3 Modélisation conjointe des données longitudinales et des données de survie

Lorsque les données de survie sont collectées en même temps que les données longitudinales il est possible d'étudier le lien entre la survenue d'un événement clinique et l'évolution d'un biomarqueur au cours du suivi. Plusieurs approches ont été développées pour permettre l'analyse de ces données. Une approche naïve consiste à inclure cette variable dépendante du temps dans un modèle à risques proportionnels tel qu'un modèle de Cox. Or, cette méthode possède plusieurs limites induisant une estimation biaisée des paramètres (Prentice, 1982).

En effet, elle ne tient pas compte de l'erreur de mesure sur la variable longitudinale. De plus, l'estimation du modèle à risques proportionnels avec variable dépendante du temps par maximisation de la vraisemblance partielle de Cox, nécessite qu'elles soit mesurée à tous les temps d'événements. Comme cela n'est pratiquement jamais le cas étant donné que la variable longitudinale n'est mesurée qu'à des temps de visites, une imputation de toutes les valeurs au temps d'événements où l'individu est à risque est nécessaire. Usuellement, la dernière valeur observée avant le temps d'événement est imputée (méthode LOCF [1]). Cette imputation et la non prise en compte de l'erreur de mesure induisent des biais. Enfin, la variable longitudinale est généralement une variable endogène. Cela signifie que sa valeur peut être modifiée par la survenue d'un événement. Or, le modèle de Cox avec variable dépendante du temps nécessite une variable exogène, c'est-à-dire, une variable donc la valeur n'est pas modifiée par la survenue des événements (Rizopoulos, 2010).
Les modèles conjoints ont été développés pour étudier la relation entre le risque d'un événement au cours du temps et l'évolution d'une variable longitudinale. Un deuxième objectif des modèles conjoints est le développement de modèles de prédiction du risque basés sur des variables longitudinales en s'affranchissant des biais décrits ci-dessus. La figure II.5 présente la structure générale d'un modèle conjoint. Le principe consiste à décrire le changement au cours du temps des mesures répétées d'un marqueur en présence de censure informative et d'étudier l'association entre un processus longitudinal et un processus de survie afin de prédire le temps d'événements étant donnée la trajectoire du marqueur.

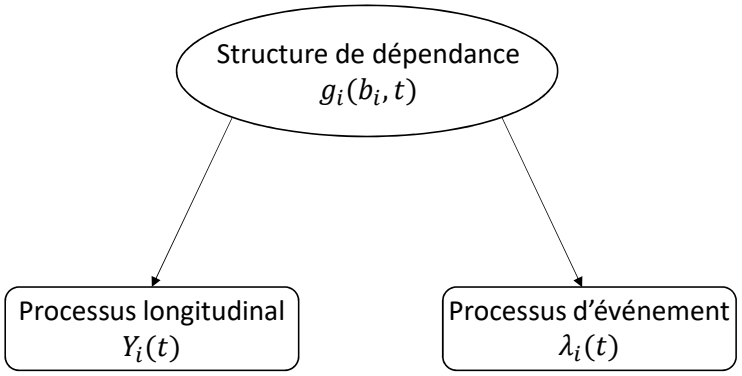

FIGURE II.5 – Schéma de la structure d'un modèle conjoint à effets aléatoires partagés.

Le premier modèle conjoint développé est le modèle conjoint à effets aléatoires partagés (Wulfsohn and Tsiatis, 1997; Tsiatis and Davidian, 2004). L'association entre le processus

1. *Last Observation Carried Forward*

longitudinal et le processus d'événements est représentée par une fonction des effets aléatoires individuels. Le second modèle est le modèle conjoint à classes latentes (Proust-Lima et al., 2009) dans lequel la population est supposée hétérogène et divisée en plusieurs classes latentes (non observées). Ces classes sont caractérisées par une trajectoire moyenne du processus longitudinal et un risque d'événement propres à chacune. Dans la suite du manuscrit, nous nous intéressons exclusivement aux modèles conjoints à effets aléatoires partagés.

II.3.1 Modèle conjoint à effets aléatoires partagés

II.3.1.1 Spécification du modèle

Le modèle conjoint à effets aléatoires partagés est constitué de deux sous-modèles : un modèle linéaire à effets mixtes pour les mesures répétées de la variable longitudinale et un modèle de survie pour le risque individuel de l'événement d'intérêt. Ces deux sous-modèles sont liés par une structure de dépendance qui est une fonction des effets aléatoires. Elle est introduite comme covariable dans le modèle de survie et permet par exemple d'étudier l'impact d'une ou plusieurs caractéristiques de la trajectoire du marqueur longitudinal (comme la valeur courante ou la pente) sur le risque d'événement. On suppose par ailleurs l'indépendance entre le marqueur et le temps d'événement, conditionnellement aux effets aléatoires.

En conservant les notations définies précédemment, le modèle conjoint à effets aléatoires partagés se définit formellement, pour tout $i \in \{1, ..., N\}$, pour tout $j \in \{1, ..., n_i\}$, par :

$$\begin{cases} Y_{ij} = X_{ij}^\top \beta + Z_{ij}^\top b_i + \epsilon_{ij} \\ \lambda_i(t) = \lambda_0(t) \exp\left(W_i^\top \gamma + g_i\left(b_i, t\right)^\top \alpha\right) \end{cases} \tag{II.4}$$

avec

- Y_{ij} la variable aléatoire associée à la j-ème mesure de l'individu i,
- $\beta \in \mathbb{R}^p$, $p \in \mathbb{N}^*$, le vecteur des effets fixes associé au vecteur de covariables X_{ij},
- $b_i \in \mathbb{R}^q$, $q \in \mathbb{N}^*$, tel que $b_i \sim \mathcal{N}(0, B)$, le vecteur des effets aléatoires associé au vecteur de covariable Z_{ij},
- $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, l'erreur résiduelle pour l'individu i à la mesure j, telle que pour tout $i \in \{1, ..., N\}$, pour tout $j \neq j'$, $j, j' \in \{1, ..., n_i\}$, $\epsilon_{ij} \perp \epsilon_{ij'}$ et $\epsilon_{ij} \perp b_i$,
- $\lambda_0(t)$ la fonction de risque de base associée au temps t, pouvant être paramétrique ou non,
- $\gamma \in \mathbb{R}^r$, $r \in \mathbb{N}$, le vecteur contenant les coefficients de régression du modèle de survie associé au vecteur de covariables indépendantes du temps, W_i,

- α, le vecteur des coefficients de régression liés à la fonction $g_i(b_i, t)$ qui mesure l'association entre le risque et l'évolution de Y.

La fonction $g_i(b_i, t)$ peut être définie de différentes manières en fonction du contexte et de l'objectif de l'analyse. Quelques unes des formes les plus rencontrées sont :

- Les effets aléatoires : $g_i(b_i, t) = b_i$,
- La valeur courante : $g_i(b_i, t) = \widetilde{y}_i(t) = X_{ij}^\top(t)\beta + Z_{ij}^\top(t)b_i$,
- La pente courante : $g_i(b_i, t) = \widetilde{y}_i'(t) = \frac{\partial \widetilde{y}_i(t)}{\partial t}$,
- La valeur courante et la pente : $g_i(b_i, t) = (\widetilde{y}_i(t), \widetilde{y}_i'(t))$.

II.3.1.2 Estimation

Deux approches principales ont été développées pour estimer un modèle conjoint : l'estimation en deux étapes (ou *two-stages)* et l'estimation conjointe.

Estimation *Two-Stages*

Dans l'approche en deux étapes, une première étape consiste à estimer les paramètres du modèle linéaire mixte, le plus souvent à l'aide du maximum de vraisemblance (section II.1.1.2). Puis une deuxième étape permet d'estimer le modèle de Cox en incluant comme covariable la fonction $g_i(\hat{b}_i, t)$ avec $\hat{b}_i$ les effets aléatoires prédits calculés selon la méthode développée en section II.1.1.3. On distingue deux méthodes selon la quantité d'information utilisée : la méthode séquentielle (Tsiatis et al., 1995) et la méthode de régression calibration (Ye et al., 2008).
La méthode séquentielle consiste à estimer un modèle linéaire mixte pour chaque temps d'événement t en incluant les données collectées jusqu'en t pour les individus encore à risque en ce temps. La fonction $g_i(b_i, t)$ est alors estimée pour chaque temps t en utilisant l'espérance des b_i conditionnellement aux données collectées jusqu'en t. La valeur de cette fonction est donc très variable et dépendante du nombre de mesures répétées jusqu'en t ainsi que de la taille d'échantillon. En outre, cette méthode est particulièrement lourde numériquement, ce qui la rend très peu utilisée.
La méthode de régression calibration estime indépendamment les deux sous-modèles en incluant toute l'information disponible. Ainsi, le modèle mixte est estimé sous l'hypothèse de données manquantes aléatoires. Dans le cas où la fonction g_i est associée au risque de subir l'événement, les données manquantes sont probablement informatives, ce qui génère un biais lors de l'estimation du risque (Albert and Shih, 2010).

Estimation conjointe

La deuxième approche consiste à utiliser une estimation conjointe des deux sous-modèles permettant alors d'estimer simultanément l'ensemble des paramètres du modèle, soit par une inférence fréquentiste, soit par une inférence bayésienne, toutes deux basées sur la vraisemblance du modèle. On se place dans le cas d'un modèle paramétrique dans lequel la forme de la fonction du risque de base est connue. Notons $\theta = (\beta, \gamma, \alpha, B, \sigma, \lambda_0)$ l'ensemble des paramètres à estimer. Sous l'hypothèse d'indépendance des données longitudinales et des données de survie conditionnellement aux effets aléatoires, la vraisemblance s'écrit :

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

En supposant le marqueur longitudinal normalement distribué, la contribution individuelle à la vraisemblance pour cette partie est donnée par :

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

où $||.||$ représente la norme euclidienne et

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

la distribution des effets aléatoires individuels supposés gaussiens. La contribution individuelle à la vraisemblance de la partie survie est

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

avec $\lambda(.)$ la fonction de risque instantané.
La log-vraisemblance s'écrit alors :

$$\ell(\theta|Y, T, \delta) = \sum_{i=1}^{N} \log\left(\int_{\mathbb{R}^q} f_Y(Y_i|b_i, \theta) f_E(T_i, \delta_i|b_i, \theta) f_b(b_i, \theta) db_i\right).$$

Inférence Bayésienne
L'inférence bayésienne permet d'estimer la distribution *a posteriori* des paramètres, notée $\pi(\theta|Y, T, \delta)$. À une constante près, cette distribution se déduit par le produit entre la vraisemblance du modèle $\mathcal{L}(\theta|Y, T, \delta)$ et la loi *a priori* des paramètres, $\pi(\theta)$, puisque d'après la

formule de Bayes il advient que :

$$\pi(\theta|Y,T,\delta) \propto \pi(\theta) \times \mathcal{L}(\theta|Y,T,\delta).$$

Cependant, la loi *a posteriori* n'a généralement pas d'expression analytique et diverses méthodes d'échantillonnage (e.g. algorithmes MCMC) sont utilisées pour approcher la loi cible. Ainsi, en simulant un grand échantillon de valeurs issues de la loi *a posteriori*, la distribution empirique (i.e. d'échantillonnage) fournit une bonne approximation de la loi *a posteriori* des paramètres et permet l'inférence statistique.

Inférence fréquentiste
L'inférence fréquentiste repose sur la maximisation de la log-vraisemblance. Pour cela, il est d'abord nécessaire de calculer une approximation des différentes intégrales n'ayant pas de solution analytique. En particulier, l'intégrale de la fonction de risque peut-être approchée par la quadrature de Gauss-Kronrod (Rizopoulos, 2016). L'intégrale sur les effets aléatoires peut être approchée par différentes méthodes d'approximation, comme par exemple, la quadrature de Gauss-Hermite (Williamson et al., 2008), l'intégration de Monte-Carlo (Hickey et al., 2018) ou de Quasi-Monte-Carlo (QMC) (Philipson et al., 2020), ou encore par approximation de Laplace lorsque le nombre d'effets aléatoires est important (Rizopoulos, 2016). Puis divers algorithmes itératifs d'optimisation numérique peuvent être utilisés tels que l'algorithme EM ou encore les algorithmes de types Newton (Newton-Raphson ou Quasi-Newton ou Marquardt-Levenberg).

II.3.1.3 Évaluation de l'ajustement aux données

Pour évaluer l'adéquation du modèle aux données, des examens visuels peuvent être réalisés. Pour le modèle linéaire à effets mixtes, il paraît intéressant de comparer graphiquement la moyenne des valeurs observées du marqueur Y à chaque temps de mesure, et son intervalle de confiance, à la trajectoire moyenne des prédictions conditionnelles $\widehat{\mathbb{E}}(Y_i(t)|X_i(t),b_i)$ calculées pour tous les individus observés aux temps de mesure considérés. Les effets aléatoires sont alors préalablement estimés par l'estimateur empirique Bayésien défini comme l'espérance *a posteriori* de b_i sachant les données et les paramètres estimés :

$$\mathbb{E}(b_i|Y_i,T_i,\delta_i;\widehat{\theta}) = \int b \frac{L(Y_i,T_i,\delta_i|b;\widehat{\theta})}{L(Y_i,T_i,\delta_i;\widehat{\theta})} f_{b_i}(b;\widehat{\theta})db$$

que l'on peut approximer par le mode de $\mathcal{L}(Y_i,T_i,\delta_i|b;\widehat{\theta})f_{b_i}(b;\widehat{\theta})$.

Pour le modèle de survie, la courbe de survie prédite moyenne peut être comparée à l'estimateur non-paramétrique de Kaplan-Meier. Dans le cas de risques compétitifs, il suffit de comparer les courbes de risques cumulés prédites pour chaque risque à l'estimateur non paramétrique de Nelson-Aalen associé (Nelson, 1969; Aalen, 1976).

II.3.1.4 Extensions

Le modèle conjoint présenté précédemment peut être étendu de diverses manières. D'une part pour prendre en compte des schémas de données de survie plus complexes, tels qu'explicités dans la section II.2. Et d'autre part afin de traiter des marqueurs longitudinaux ne suivant pas une loi normale.

Troncature à gauche

De la même façon que défini en section II.2, il est nécessaire de prendre en compte la troncature à gauche lorsque le temps d'inclusion dans l'étude diffère entre les individus. Pour cela, il suffit de diviser la contribution individuelle à la vraisemblance du modèle par la probabilité de ne pas avoir subi l'événement avant l'entrée dans l'étude :

$$\ell(\theta|Y,T,\delta) = \sum_{i=1}^{N} \log\left(\frac{\int_{\mathbb{R}^q} f_Y(Y_i|b_i,\theta) f_E(T_i,\delta_i|b_i,\theta) f_b(b_i,\theta) db_i}{\int \exp(-\Lambda_i(T_{0i}|b_i;\theta)) f_b(b_i;\theta) db_i}\right).$$

où T_{0i} correspond au temps d'entrée dans l'étude pour l'individu i et $\Lambda_i(t|b_i;\theta) = \int_0^t \lambda_i(u|b_i;\theta)du$.

Risques compétitifs

L'extension des modèles conjoints à effets aléatoires partagés à K risques compétitifs a été proposée par Elashoff et al. (2008) en définissant pour chaque risque $k \in \{1,\ldots,K\}$ une fonction pouvant dépendre des caractéristiques du modèle longitudinal :

$$\lambda_{ik}(t) = \lambda_{0k}(t)\exp(W_{ik}^{\top}\gamma_k + g_{ik}(b_i,t)^{\top}\alpha_k)$$

avec $g_k(b_i,t)$ une fonction des effets aléatoires et du temps et $\lambda_{0k}(t)$ la fonction de risque de base pour l'événement k. Conditionnellement aux effets aléatoires b_i, le marqueur et les K événements compétitifs sont supposés indépendants. La contribution individuelle à la vraisemblance est alors définie par :

$$\mathcal{L}_i(\theta;Y_i,T_i,\delta_i) = \int f_Y(Y_i|b_i;\theta)\exp\left(-\sum_{k=1}^{K}\Lambda_{ik}(T_i|b_i;\theta)\right)\prod_{k=1}^{K}\lambda_{ik}(T_i|b_i;\theta)^{\mathbb{1}_{\delta_i=k}} f_b(b_i;\theta)db_i,$$

avec $\Lambda_{ik}(T_i|b_i;\theta)$ la fonction de risque cumulée pour l'événement k donnée par :

$$\Lambda_{ik}(t|b_i;\theta) = \int_0^t \lambda_{ik}(u|b_i;\theta)du \tag{II.5}$$

L'estimation du modèle se fait similairement à l'estimation du modèle conjoint classique, par inférence bayésienne ou fréquentiste. En général, et dans la suite de ce manuscrit, on considérera seulement deux risques compétitifs ($K = 2$).

Censure par intervalle

Dans le cadre des modèles conjoints à effets aléatoires partagés, Gueorguieva et al. (2012) ont proposé un modèle traitant les risques compétitifs où tous les temps d'événements sont censurés par intervalle. Cependant, le type d'événement était connu pour tous les individus ayant expérimenté un événement puisque ce modèle considérait des risques compétitifs (Figure II.2) et non des risques semi-compétitifs (Figure II.3). Rouanet et al. (2016) ont développé un modèle conjoint à classes latentes avec traitement de la censure par intervalle et un modèle multi-état. Leur comparaison avec la stratégie consistant à imputer par le milieu de l'intervalle a mis en évidence des estimateurs biaisés. Cependant, la censure par intervalle et les risques compétitifs sont plus faciles à traiter dans le cas des modèles conjoints à classes latentes car le calcul de la vraisemblance ne nécessite pas d'intégrer sur les effets aléatoires. Il n'existe à ce jour aucun travaux et package permettant d'estimer un modèle conjoint à effets aléatoires partagés *illness-death* traitant la censure par intervalle.

Autres extensions

Enfin, de nombreuses autres extensions, non considérées dans la suite de ce manuscrit, ont été proposées dans la littérature. En effet, en fonction de la nature et du nombre de marqueurs considérés, le modèle linéaire à effets mixtes peut être remplacé par un modèle linéaire généralisé à effets mixtes ou par un modèle à effets mixtes multivarié. Par exemple, (Andrinopoulou et al., 2014) ont proposé un modèle avec deux marqueurs longitudinaux et des risques compétitifs. Ils peuvent également être étendues aux événements récurrents permettant d'étudier la survenue successive d'événements du même type (Rondeau et al., 2007). Les modèles conjoints à classes latentes ont également été étendus aux événements récurrents (Han et al., 2007), aux multi-marqueurs et risques compétitifs (Proust-Lima et al., 2016), ainsi qu'au traitement de la censure par intervalle dans le cas de risques semi-compétitifs (Rouanet et al., 2016). Cependant, toutes ces extensions présentent des contraintes computationnelles. Par exemple, pour les modèles multi-marqueurs, plus le nombre d'effets aléatoires augmente, plus le calcul de l'intégrale est rendu complexe.

II.3.1.5 Packages

De nombreux packages ont été développés en `R` pour permettre l'estimation de modèles conjoints, chacun offrant des fonctionnalités spécifiques et des méthodes d'approximation variées. Le package `JM` (Rizopoulos, 2010) a été l'un des premiers packages `R` permettant l'estimation de modèles conjoints à effets aléatoire partagés. Il permet d'estimer un modèle conjoint avec un seul risque et seulement la valeur courante comme fonction d'association entre les deux sous-modèles. Plusieurs fonctions de risque de base peuvent être envisagées, comme un modèle de Weibull ou des splines pénalisées. L'intégrale sur les effets aléatoires est approchée par l'intégration de Gauss-Hermite sauf dans le cas où des splines sont considérées pour le risque de base. Dans ce cas, l'approximation de Laplace peut être choisie. L'inférence repose sur la maximisation de la log-vraisemblance en commençant avec un algorithme EM pour un nombre fixé d'itérations et si la convergence n'est pas atteinte, l'algorithme de Quasi-Newton est utilisé jusqu'à la convergence. Le package `JMbayes` (Rizopoulos, 2016) peut être vu comme une extension de JM intégrant des méthodes bayésiennes en utilisant des techniques de Monte Carlo par Chaînes de Markov (MCMC) pour estimer les distributions *a posteriori* des paramètres du modèle. Il permet d'estimer des modèles conjoints dans lesquels le marqueur longitudinal est estimé par un modèle linéaire généralisé à effets mixtes. Le risque de base est modélisé par des B-splines pénalisées. La troncature à gauche et des variables exogènes et dépendantes du temps peuvent être prises en compte pour la survie. Enfin, diverses associations sont possibles entre le marqueur et la survie telles que la valeur courante, la pente et l'intégrale de la valeur courante. L'intégrale de la fonction de survie est calculée soit par Gauss-Kronrod soit par Gauss-Legendre. Le package `JMbayes2` est une évolution du package `JMbayes`, offrant une gamme plus large de sous-modèles longitudinaux et de survie. Il permet entre autres d'estimer des modèles conjoints multi-marqueurs et de prendre en compte les risques compétitifs. Le package `joineR` (Philipson et al., 2018) repose sur un algorithme EM et une intégration par Gauss-Hermite. Le modèle conjoint estimé par ce package est un modèle semi-paramétrique, dans lequel le risque de base n'est pas spécifié. Il permet de prendre en compte les risques compétitifs, et une seconde version de ce package, `joineRML` (Hickey et al., 2018) permet de traiter le cas des modèles multi-marqueurs, cette fois-ci avec une intégration par Monte-Carlo. Le package `rstanarm` (Brilleman et al., 2018) propose également une approche bayésienne en utilisant des techniques avancées d'échantillonnage bayésien comme le No-U-Turn Sampler (NUTS). Le package `INLAjoint` (Rustand et al., 2023) introduit quant à lui une approche bayésienne alternative en utilisant INLA (*Integrated Nested Laplace Approximation*) offrant ainsi une estimation rapide. Il permet de prendre en compte plusieurs marqueurs avec différentes distributions et de traiter les risques compétitifs et les modèles multi-états. Enfin, le package `lcmm` (Proust-Lima et al., 2017)

permet l'estimation de modèles conjoints à classes latentes, avec risques compétitifs, par maximisation de la vraisemblance en utilisant l'algorithme de Marquardt-Levenberg (Levenberg, 1944; Marquardt, 1963).

II.3.2 Prédictions individuelles

L'approche médicale actuelle porte de plus en plus d'intérêt à la médecine personnalisée à travers l'étude du pronostic individuel du patient. Pour cela, les prédictions dynamiques du risque individuel ont été développées afin de mettre à jour, au fil du temps, le risque pour un patient de déclarer un événement en fonction de l'évolution de ses facteurs de risques et marqueurs biologiques ou cliniques. Les modèles conjoints permettent notamment de calculer ces prédictions en prenant en compte l'ensemble de l'historique médical disponible du patient.

II.3.2.1 Prédiction des effets aléatoires individuels

Dans un objectif de prédictions individuelles, il est nécessaire de prédire la variable réponse $Y_i(t)$ pour tout temps t pour chaque individu i. Pour ce faire il s'agit premièrement de calculer les effets aléatoires prédits en prenant les estimateurs Bayésiens empiriques, ce qui correspond au mode de la distribution conditionnelle *a posteriori* des effets aléatoires sachant les données et les paramètres estimés. Cela revient à maximiser le produit suivant $f(Y_i, T_i, \delta_i; \widehat{\theta}) f_b(\widetilde{b_i}; \widehat{\theta})$ dans le cadre d'un modèle conjoint. La valeur prédite du marqueur au temps t correspond finalement à l'espérance conditionnelle étant donnés les effets aléatoires prédits, définis par :

$$\widehat{Y_i}(t) = \widehat{\mathbb{E}}(Y_i(t)|\widetilde{b_i}) = X_i(t)\widehat{\beta} + Z_i(t)\widetilde{b_i}.$$

II.3.2.2 Prédictions individuelles dynamiques

Une prédiction individuelle dynamique représente la probabilité, pour un individu i, de subir un événement sur une fenêtre de temps $]s, s+t]$, $s \in \mathbb{R}^{+*}$, sachant qu'il est toujours à risque au temps s et en tenant compte de toute l'information collectée jusqu'au temps s. Le temps s est appelé temps *landmark* et le temps t correspond à l'horizon de prédiction. Dans le cadre des modèles conjoints à effets aléatoires partagés, cette probabilité se définit par (Rizopoulos, 2011) :

$$\pi_i(s,t) = \mathbb{P}_{\hat{\theta}}(s < T_i \leq s+t, \delta_i = 1 | T_i > s, \mathcal{Y}_i(s), W_i)$$

avec $\mathbb{P}_{\hat{\theta}}$ la probabilité paramétrée par le vecteur des estimateurs obtenus avec le modèle conjoint, $\mathcal{Y}_i(s)$ toute l'information disponible au temps s pour l'individu i : $\mathcal{Y}_i(s) = \{Y_{ij} : 0 \leq t_{ij} \leq s, j = 1, ..., n_i\}$, T_i le temps d'événement de l'individu i, δ_i l'indicateur d'événement pour l'individu i, $\delta_i = 1$ si c'est l'événement d'intérêt et W_i l'ensemble des covariables de l'individu i.

Dans le cadre des modèles conjoints à effets aléatoires partagés avec K risques compétitifs, cette prédiction pour l'événement $k \in \{1, ..., K\}$ se calcule par :

$$\begin{aligned}\pi_{ik}(s, t; \widehat{\theta}) &= P(s < T_i < s + t, \delta_i = k | T_i > s, \mathcal{Y}_i(s), \widehat{\theta}) \\ &= \frac{\int \left[\int_s^{s+t} \exp\left(-\sum_{c=1}^K \Lambda_{ic}(u|b_i, \widehat{\theta})\right) \lambda_{ik}(u|b_i, \widehat{\theta}) du \right] f(\mathcal{Y}_i(s)|b_i, \widehat{\theta}) f(b_i|\widehat{\theta}) db_i}{\int \exp\left(-\sum_{c=1}^K \Lambda_{ic}(s|b_i, \widehat{\theta})\right) f(\mathcal{Y}_i(s)|b_i, \widehat{\theta}) f(b_i|\widehat{\theta}) db_i}\end{aligned} \quad \text{(II.6)}$$

II.3.2.3 Évaluation des capacités prédictives

L'évaluation des capacités prédictives est construite sur deux phases. Dans un premier temps, appelée phase d'apprentissage, les paramètres θ du modèle sont estimés, puis dans un second temps, appelée phase de validation, les prédictions individuelles sont calculées et les performances prédictives sont évaluées. Afin d'éviter un effet de sur-apprentissage, il est nécessaire que les deux étapes soient réalisées sur deux échantillons de données indépendants. Pour cela, la validation peut se faire de manière externe ou de manière interne. Puis, les performances prédictives sont évaluées à partir de divers outils tels que l'aire sous la courbe *Receiver Operating Characteristiv (ROC)* (AUC) ou le Brier Score, sur les prédictions obtenues dans l'échantillon de validation.

Validation

La validation externe consiste à évaluer l'outil de prédiction sur un nouvel échantillon provenant de la même population d'intérêt. Par exemple, dans le cas d'essais cliniques ou de cohortes multicentriques il est possible de faire l'apprentissage sur un centre et la prédiction sur un autre centre, sous l'hypothèse qu'ils soient comparables. Cette méthode est la meilleure méthode permettant d'évaluer les performances des prédictions obtenues, mais il n'est pas toujours évident d'avoir à disposition des données de validations.
Lorsque cette méthode n'est pas faisable, il faut utiliser la validation interne. Il existe différentes approches de validation interne mais la plus utilisée est la validation croisée en K-blocs (en général $K = 5$ ou $K = 10$). Le principe de cette approche consiste à diviser aléatoirement

les données en K blocs, puis pour chaque bloc k il suffit de :

1. entraîner le modèle sur les données issues des $(K-1)$ blocs, en excluant le bloc k ;
2. calculer les prédictions sur les données du $k-$ème bloc sachant l'estimation à l'étape précédente.

Finalement, les prédictions individuelles obtenues sont aggrégées pour avoir une prédiction pour chaque individu.

Critères d'évaluation

L'évaluation des performances prédictives consiste à quantifier les capacités du modèle à prédire correctement la survenue de l'événement et à ordonner les risques. Afin de quantifier ce niveau de performance pour ensuite comparer et sélectionner la meilleure méthode de prédictions, les deux critères les plus couramment utilisés sont la calibration et la discrimination. La calibration permet d'évaluer la différence entre le risque prédit par le modèle et le risque effectivement observé sur les données de validation. La discrimination correspond à la capacité du modèle à différencier les individus faisant l'événement de ceux ne le faisant pas.
De nombreux outils ont été proposés pour évaluer ces critères, les deux les plus classiquement utilisés étant l'aire sous la courbe ROC et le Brier Score. Ces critères étant appropriés pour l'évaluation pronostique (Gerds et al., 2008), leur version dynamique est présentée ci-dessous.

Brier Score
Le Brier Score (Brier, 1950; Gerds and Schumacher, 2006) est défini comme étant l'erreur quadratique moyenne entre la prédiction $\pi_k(s,t)$ et le statut de l'individu, $D_k(s,t)$, par rapport à l'événement d'intérêt k au temps d'horizon t chez les sujets encore à risques de l'événement au temps s :

$$BS_k(s,t) = \mathbb{E}\left[\left(D_k(s,t) - \pi_k(s,t)\right)^2 | T > s\right].$$

Ce critère permet d'évaluer à la fois la calibration et la discrimination du modèle. Il est défini entre 0 et 1, tel que plus la valeur est faible, meilleure est la performance prédictive. Une valeur de 0.25 correspond à des probabilités individuelles complètement aléatoires.

En présence de données censurées, le statut de l'individu $D_k(s,t)$ n'est pas toujours observé. Il convient alors de définir $\widetilde{D}_k(s,t) = \mathbb{1}_{(s<\widetilde{T}+t,\delta=k)}$, tel que si l'événement d'intérêt est observé entre s et $s+t$, $\widetilde{D}_k(s,t) = 1$, sinon $\widetilde{D}_k(s,t) = 0$. Dans ce cas, le Brier Score est estimé en utilisant la méthode IPCW (*Inverse Probability of Censoring Weighting*) (Gerds

and Schumacher, 2006; Blanche et al., 2015) :

$$\widehat{BS}_k(s,t) = \frac{1}{\sum_{i=1}^{N} \mathbb{1}_{\widetilde{T}_i > s}} \sum_{i=1}^{N} \widehat{W}_i(s,t)(\widetilde{D}_{ik}(s,t) - \widetilde{\pi}_{ik}(s,t))^2$$

avec $\widehat{W}_i(s,t)$ représentant les poids qui permettent de prendre en compte la censure des données :

$$\widehat{W}_i(s,t) = \frac{\mathbb{1}_{(\widetilde{T}_i > s+t)}}{\widehat{G}(s+t)/\widehat{G}(s)} + \frac{\mathbb{1}^{\delta_{ik}}_{(s<\widetilde{T}_i \leq s+t)}}{\widehat{G}(\widetilde{T}_i)/\widehat{G}(s)}$$

où $\widehat{G}$ est l'estimateur de Kaplan-Meier de la fonction de survie de la censure.

AUC

L'AUC est probablement l'outil le plus couramment utilisé pour mesurer les capacités de discrimination d'un modèle. Elle correspond à la probabilité de concordance entre l'outil de prédiction et les observations. Elle est comprise entre 0 et 1, et contrairement au Brier Score, plus l'AUC est proche de 1, meilleure est la capacité de discrimination du modèle. Le seuil de 0.5 correspond à des prédictions totalement aléatoires.

On définit l'AUC pour l'événement k, entre un temps landmark s et un temps d'horizon t, par (Zheng and Heagerty, 2007; Blanche et al., 2013) :

$$AUC_k(s,t) = P\left(\pi_{ki}(s,t) > \pi_{kj}(s,t) | D_{ki}(s,t) = 1, D_{ki}(s,t) = 0, T_i > s, T_j > s\right).$$

Elle représente donc la probabilité, pour deux individus indépendants, i et j, que la prédiction dynamique de l'individu i, c'est-à-dire, la probabilité d'expérimenter l'événement d'intérêt ($\delta_k = 1$) entre les temps s et $s+t$, est supérieure à celle de l'individu j, sachant que l'individu i déclare l'événement d'intérêt entre les temps s et $s+t$ sans déclarer d'autre événement avant le temps s, et que l'individu j n'a pas présenté d'événement avant le temps s et ne présente pas l'événement d'intérêt entre s et $s+t$.

De même que pour le Brier Score, dans le cas de données censurées, l'AUC est estimée par la méthode IPCW (Blanche et al., 2015) :

$$\widehat{AUC}_k(s,t) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \widetilde{D}_{ik}(s,t)(1-\widetilde{D}_{jk}(s,t))\widehat{W}_i(s,t)\widehat{W}_j(s,t)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{1}_{(\hat{\pi}_{ik}(s,t)>\hat{\pi}_{jk}(s,t))}\widetilde{D}_{ik}(s,t)(1-\widetilde{D}_{jk}(s,t))\widehat{W}_i(s,t)\widehat{W}_j(s,t)}.$$

II.4 Modélisation conjointe avec variabilité hétérogène

Dans les modèles conjoints présentés précédemment, l'erreur résiduelle est commune entre tous les individus. Comme vu en section II.1.2, des modèles *location-scale* linéaires à effets mixtes ont été développés afin de prendre en compte l'hétérogénéité de la variance résiduelle. Après ces travaux, de nouveaux travaux (Gao et al., 2011; Elliott et al., 2012; Barrett et al., 2019) se sont intéressés aux modèles conjoints avec variance résiduelle hétérogène en ajustant le sous-modèle de survie sur cette variance.

II.4.1 Variable binaire ajustée sur la variabilité résiduelle hétérogène

Les modèles conjoints présentés en section II.3 prennent en compte des données de survie. Or, il est également possible de considérer un modèle conjoint permettant de modéliser l'effet d'un marqueur longitudinal sur une variable réponse binaire. Dans cette logique, Elliott et al. (2012) ont proposé un tel modèle avec une variance résiduelle hétérogène. Soit Y et W deux variables aléatoires correspondant respectivement au marqueur longitudinal et à la variable réponse binaire. Le modèle s'écrit :

$$\begin{cases} Y_{it}|\beta_i, \sigma_i^2 \sim \mathcal{N}(f(\beta_i, t), \sigma_i^2) \\ W_i|\beta_i, \sigma_i^2, Z_i, \gamma \sim \mathcal{B}er(\pi_i), \text{ avec } logit(\pi_i) = \log(\frac{\pi_i}{1-\pi_i}) = g(\gamma, \beta_i, \sigma_i^2, Z_i) \end{cases}$$

où $\beta_i \sim \mathcal{N}(\beta, \Sigma)$ et $\log(\sigma_i^2) \sim \mathcal{N}(\sigma, \Psi^2)$. Ce modèle fait l'hypothèse que le marqueur longitudinal est distribué selon une loi normale de moyenne $f(\beta_i, t)$ qui peut être une fonction linéaire ou non du temps, et de variance résiduelle spécifique à l'individu i, notée σ_i^2. Pour la variable binaire, la probabilité d'observer l'événement est ajustée sur la fonction $g(.)$ permettant une association linéaire ou non avec les effets aléatoires du marqueur (de la moyenne et de la variance), via un lien logit.

L'estimation est réalisée par inférence bayésienne avec une approche MCMC combinant un échantillonneur de Gibbs et un algorithme de Metropolis-Hastings, à l'aide du logiciel WinBUGS (Cowles, 2004).
Bien que permettant de prédire la probabilité de survenue d'un événement, ce modèle ne tient pas compte des temps de survie bien souvent disponibles dans les études.

II.4.2 Risque d'événement ajusté sur la variabilité résiduelle hétérogène

Gao et al. (2011), puis Barrett et al. (2019) ont proposé un modèle conjoint incluant un effet aléatoire spécifique au sujet sur la variance de l'erreur de mesure. Le modèle mixte se définit pour $i \in \{1, ..., n\}$, et $j \in \{1, ..., n_i\}$ par :

$$Y_{ij} = X_{ij}^\top \beta + Z_{ij}^\top b_i + \epsilon_{ij}$$

avec

$$\epsilon_{ij} \sim N\left(0, \sigma_i^2\right) \quad \text{et} \quad \begin{pmatrix} b_i \\ \log \sigma_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \mu_\sigma \end{pmatrix}, \begin{pmatrix} B & \Sigma_{b\sigma} \\ \Sigma_{b\sigma}^\top & \tau_\sigma^2 \end{pmatrix}\right).$$

Gao et al. (2011) ne considèrent des effets aléatoires que sur l'intercept et la pente.
Concernant le modèle de survie, ils définissent la fonction de risque comme :

$$\lambda_i(t) = \lambda_0(t) \exp\left(W_i^\top \gamma + \alpha_b^\top b_i + \alpha_\sigma \log(\sigma_i^2)\right)$$

Alors que Barrett et al. (2019) considère :

$$\lambda_i(t) = \lambda_0(t) \exp\left(W_i^\top \gamma + \alpha_b^\top b_i + \alpha_\sigma \sigma_i\right)$$

où α_b et α_σ traduisent respectivement le lien entre les effets aléatoires b_i et l'effet aléatoire σ_i et la fonction de risque, et $\lambda_0(t)$ est la fonction de risque de base, définie comme une fonction de Weibull (Gao et al., 2011) ou une fonction constante par morceaux (Barrett et al., 2019).

La méthode d'estimation proposée dans ces deux travaux repose sur l'inférence bayésienne via l'utilisation de l'échantilleur WinBUGS ou JAGS afin d'obtenir les échantillons *a posteriori* des paramètres.

Bien que considérant une variance résiduelle individuelle, ce modèle présente de nombreuses restrictions. D'une part, la fonction de risque de base est trop contrainte et n'apporte pas suffisamment de flexibilité contrairement à l'utilisation de splines. D'autre part, ce modèle évalue l'impact des effets aléatoires sur le risque d'événement, or, d'autres fonctions complexes des effets aléatoires, telles que la valeur courante ou la pente courante, peuvent être plus intéressantes à considérer en pratique. En effet, les effets aléatoires sont constants dans le temps et leur interprétation clinique n'est pas toujours évidente. Par ailleurs, la variance est considérée constante dans le temps, ce qui n'est cliniquement pas évident. Enfin, ce modèle ne tient pas compte des risques compétitifs, limitant ainsi son utilisation.

II.4.3 Association avec une variabilité non résiduelle du marqueur

Wang et al. (2024) ont récemment proposé un modèle conjoint ajusté sur la variabilité globale du marqueur. Ils proposent une mesure permettant de caractériser la variabilité biologique inhérente d'un biomarqueur dont la trajectoire sous-jacente possède une forme lisse et non linéaire. Dans ce modèle, le risque dépend d'un indicateur de fluctuation de la trajectoire individuelle, la racine carrée de l'intégrale du carré de la dérivée seconde, qui dépend lui-même des effets aléatoires du marqueur.
En supposant que la trajectoire des mesures longitudinales soit une fonction qui dépend linéairement des covariables de base et non paramétriquement du temps, Wang et al. (2024) proposent le modèle semi-paramétrique suivant :

$$\widetilde{Y}_i(t) = X_i^\top \eta + m_i(t)$$

et

$$Y_{ij} = \widetilde{Y}_i(t_{ij}) + Z_{ij}^\top \xi + \epsilon_{ij}$$

où η et ξ sont les vecteurs des coefficients associés aux covariables X_i et Z_{ij} respectivement, $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ l'erreur résiduelle et $m_i(t)$ est une fonction du temps modélisée par des splines avec des effets aléatoires de telle sorte que :

$$m_i(t) = \sum_{k=1}^{q} (\beta_k + b_{ik}) B_k(t)$$

avec $\{B_k(.)\}_{k=1,...,q}$ une base de B-splines cubiques du temps, de dimension q, $\beta = (\beta_1, ..., \beta_q)^\top$ le vecteur des effets fixes et $b_i = (b_{i1}, ..., b_{iq})^\top$ le vecteur des effets aléatoires, supposés indépendants et identiquement distribués selon $\mathcal{N}(0, D)$ et indépendants des ϵ_{ij}. Ainsi, l'ensemble $\{m_i(t)\}_{i=1,...,N}$ peut être vu comme une collection des trajectoires aléatoires variant autour d'une évolution moyenne commune $\sum_{k=1}^{q} \beta_k B_k(t)$ avec des déviations aléatoires spécifiées par les effets aléatoires individuels.

Ensuite, les auteurs utilisent la norme deux de la dérivée seconde de la trajectoire individuelle du marqueur, soit l'intégrale de la dérivée seconde au carré de la trajectoire individuelle entre le temps initial t_0 et le temps courant t, soit $\int_{t_0}^{t} \{m_i''(s)\}^2 ds$, pour représenter la variabilité cumulative de la trajectoire du biomarqueur $m_i(.)$ de l'individu i entre t_0 et t. Ainsi, la racine carrée de cette quantité est ajoutée comme prédicteur dans la fonction de survie telle

que :

$$\lambda_i(t) = \lambda_0(t) \exp \left\{ W_i^\top \gamma + \alpha_1 (X_i^\top \eta + m_i(t)) + \alpha_2 \left(\int_{t_0}^{t} \{m_i''(s)\}^2 ds \right)^{1/2} \right\}$$

avec $\lambda_0(.)$ la fonction de risque de base non paramétrique, γ un vecteur de coefficient de régression des covariables à baseline W_i, α_1 et α_2 des paramètres caractérisant les effets de la valeur courante du biomarqueur et de la racine carrée de la variabilité cumulative.

L'estimation se fait par un algorithme EM cherchant à maximiser la fonction de vraisemblance.

Ce modèle ne répond cependant pas à notre problématique car il ne s'intéresse pas à la variabilité résiduelle du marqueur. En outre, il ne permet pas de traiter les risques compétitifs ni la censure par intervalle.

Chapitre III

Modèle conjoint avec variance résiduelle dépendante du temps et risques compétitifs

Sommaire

RÉSUMÉ

Sous la forme d'un article scientifique international, ce chapitre introduit un modèle conjoint permettant d'étudier l'impact de la variabilité résiduelle de la pression artérielle sur les risques de CCVD et de décès. En effet, comme explicité dans le chapitre I, leur prévention est un enjeu majeur de santé publique et malgré l'intérêt croissant pour l'étude de la variabilité de la PAS comme facteur de risque, les études existantes souffrent encore de faiblesses méthodologiques. Ce travail a conduit à un article scientifique en cours de révision dans une revue internationale. Ce chapitre correspond à la version courante de l'article qui est disponible librement en tant que pre-print (Courcoul et al., 2024b) [1]. Ainsi, la suite de ce chapitre sera rédigée en anglais mais un résumé en français est donné ci-après.

Après une introduction motivant le développement d'un tel modèle, nous proposons un nouveau modèle conjoint *location-scale* pour les mesures répétées d'un marqueur et la survenue d'événements compétitifs. Ce modèle conjoint combine un modèle linéaire mixte comprenant une variance résiduelle spécifique au sujet définie suivant un prédicteur linéaire incluant des effets fixes (associés au temps et autres variables explicatives) et des effets aléatoires individuels, et un modèle à risques proportionnels cause-spécifique pour les événements compétitifs. Le risque d'événements peut dépendre simultanément de la valeur courante de la variance, ainsi que de la valeur courante et de la pente de la trajectoire du marqueur. Le modèle est estimé en maximisant la fonction de vraisemblance à l'aide de l'algorithme de Marquardt-Levenberg et d'une intégration numérique par approximation de Quasi-Monte-Carlo. La procédure d'estimation est validée par une étude de simulation. Ce modèle est ensuite appliqué sur les données de l'essai PROGRESS pour étudier l'association entre la variabilité de la PAS et le risque de CCVD et de décès pour d'autres causes. La modélisation permet, d'une part de mettre en évidence une différence de variabilité entre les individus. En effet, la variabilité est significativement plus faible pour les individus traités, et son évolution au cours du temps diffère d'un individu à l'autre (variance des effets aléatoires individuels non nulle). Et d'autre part, nous constatons que la variabilité individuelle de la PAS est associée au risque de CCVD et de décès. Enfin, ces résultats ont été confirmés dans une comparaison de modèles où le modèle avec variance résiduelle homogène est moins bon au sens de l'AIC, ce qui est validé par les ajustements visuels aux données ou encore par les prédictions dynamiques.

1. *A location-scale joint model for studying the link between the time-dependent subject-specific variability of blood pressure and competing events*, L. Courcoul, C. Tzourio, M. Woodward, A. Barbieri et H. Jacqmin-Gadda, arXiv :2306.16785

III.1 Introduction

Cardiovascular diseases, such as ischaemic heart disease, and cerebrovascular events are two leading causes of death. Moreover these diseases lead often to acquired physical disability or to dementia. Medical care and disability management following this type of diseases generate significant societal, human, and financial distress (De Pouvourville, 2016). Given the frequency of cardio and cerebrovascular diseases (CVD) and its dramatic consequences at the individual and societal level, the identification of modifiable risk factors is essential to implement prevention programs. Hypertension (high values of blood pressure) is a well-known risk factor for these diseases. More recently, some studies have suggested that the visit-to-visit variability of blood pressure could be associated with an increased risk of stroke and cardiovascular events independently of the level of blood pressure (Pringle et al., 2003; Rothwell et al., 2010; Shimbo et al., 2012). These studies have used the individual empirical standard deviation, or some other measures of variation (e.g. the coefficient of variation) or extreme value (e.g. the maximum), of blood pressure as an explanatory variable in a Cox model for the event risk. However, they were exposed to methodological issues. A first strategy consists of calculating the empirical standard deviation of blood pressure on all available measurements (Mehlum et al., 2018). This strategy induces conditioning on the future, likely leading to bias because measurements after the current time (and sometimes after the event time) are used to predict the event at the current time (Andersen and Keiding, 2012; de Courson et al., 2021). A second strategy consists of computing the standard deviation of blood pressure on the measurements collected over an initial period of the study, keeping in the sample only the individuals who did not have the event before the end of this period in order to predict the risk beyond this period (de Courson et al., 2021). This could induce selection bias and certainly creates loss of power. To avoid these issues, the standard deviation of blood pressure can be considered as a time-dependent variable and calculated using only measurements before the event. Nevertheless, this approach neglects the measurement error of the standard deviation, which is a serious issue when the number of measurements differs between individuals, and requires imputation of the standard deviation at all event times. These limitations may introduce bias (Prentice, 1982).

Recently, de Courson et al. (2021) compared these different approaches and obtained contradictory results depending on the estimator used for the variability. Furthermore, they pointed out that blood pressure and its standard deviation are endogenous variables for which the Cox model is not suitable (Commenges and Jacqmin-Gadda, 2015). Thus, they also considered a joint model approach combining a mixed model to fit individual trajectories of standard deviation of blood pressure measurements and a proportional hazard model to assess

the impact of the current variability on the event risk. In the joint model, the current value of standard deviation of blood pressure is included as a time-dependent explanatory variable in the time-to-event model. This allows the evaluation of the impact of the longitudinal data on the event risk without bias, contrary to the two stage estimation (Rizopoulos, 2012; Tsiatis and Davidian, 2004; Henderson et al., 2000). However, the interpretation of the association between the event risk and the variability computed as the standard deviation of measurements observed since the beginning is difficult. Indeed, this global variability encompass the individual time trend of the blood pressure while the clinical question is the following: is irregularity among blood pressure measures a risk factor independently of current values and possibly time-trend of blood pressure?

Location-scale mixed models have been introduced to investigate the heterogeneity of intra-subject variability for longitudinal data by introducing random effects in the residual variance modelling (Hedeker and Nordgren, 2013). For studying the association between the variability of a biomarker and a clinical event, Gao et al. (2011) and Barrett et al. (2019) have proposed a joint model combining a mixed model including a subject-specific random effect for the residual variance and a proportional hazard model for the event risk. However, the considered dependence structure is quite restrictive since, in their models, the event risk depends only on the random effects and not on time-dependent characteristics of the marker trajectory, such as the current value or the current slope. In addition, none of them consider for time-dependent subject-specific variability of the maker and they do not handle competing events. However, it is essential to account for competing death from other causes because mortality and cardiovascular risk increase with age and may be both associated with blood pressure.

The objective of our work was, therefore, to propose a new location-scale joint model accounting for both time-dependent individual variability of a marker and competing events. To do this, we extended the model proposed by Gao et al. (2011) and Barrett et al. (2019) to include a time-dependent variability, competing events, a more flexible dependence structure between the event and the marker trajectory, and more flexible baseline risk functions. In contrast to the previous works we propose a frequentist estimation approach which is implemented in the `R`-package `FlexVarJM`.

This paper is organized as follows. Section 2 describes the model and the estimation procedure using a robust algorithm for maximizing the likelihood. Section 3 presents a simulation study to assess the estimation procedure performance. In section 4, the model is applied to the data from the Perindopril Protection Against Stroke Study (PROGRESS) clinical trial, a blood-pressure lowering trial for the secondary prevention of stroke (Mac Mahon et al., 2001). Finally, Section 5 concludes this work with some elements of discussion.

III.2 Method

Let us consider a sample of N individuals. For each individual $i \in \{1, ..., N\}$, we consider the n_i-vector of repeated measures $Y_i = (Y_{i1}, ..., Y_{in_i})^\top$ with Y_{ij} the value of the longitudinal outcome of individual i at time t_{ij} ($j = 1, \ldots, n_i$). Assuming two competing events, we denote $T_i = \min(T_{i1}^*, T_{i2}^*, C_i)$ the observed time with T_{ik}^* the real time for the event k ($k = 1, 2$) and C_i the censoring time for the ith individual. Censoring event and real time are supposed to be independent. We then denote $\delta_i \in \{0, 1, 2\}$ the individual event indicator such as $\delta_i = k$ if the competing event $k \in \{1, 2\}$ occurs and $\delta_i = 0$ otherwise.

III.2.1 Joint model with time-dependent individual variability

We propose joint modelling for a longitudinal outcome and competing events using a shared random-effect approach. Joint models allow simultaneous analysis of longitudinal data and clinical events. They combine a mixed model for repeated measures of exposure and a time-to-event model. Functions of the random effects from the mixed model are included as explanatory variables in the time-to-event model to account for the association between the two outcomes. The longitudinal submodel is defined by a linear mixed-effect model with heterogeneous variance:

$$\begin{cases} Y_{ij} = Y_i(t_{ij}) = \widetilde{Y}_i(t_{ij}) + \epsilon_{ij} = X_{ij}^\top \beta + Z_{ij}^\top b_i + \epsilon_i(t_{ij}), \\ \epsilon_{ij}(t_{ij}) \sim \mathcal{N}(0, \sigma_i^2(t_{ij})) \quad \text{with} \quad \log(\sigma_i(t_{ij})) = O_{ij}^\top \mu + M_{ij}^\top \tau_i \end{cases} \tag{III.1}$$

with X_{ij}, O_{ij}, Z_{ij} and M_{ij} four vectors of explanatory variables for subject i at visit j, respectively associated with the fixed-effect vectors β and μ, and the subject-specific random-effect vector b_i and τ_i, such as

$$\begin{pmatrix} b_i \\ \tau_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_b & \Sigma_{\tau b} \\ \Sigma_{\tau b}^\top & \Sigma_\tau \end{pmatrix} \right)$$

The risk function for the event $k \in \{1, 2\}$ is defined by:

$$\lambda_{ik}(t) = \lambda_{0k}(t) \exp\left(W_i^\top \gamma_k + \alpha_{1k} \tilde{y}_i(t) + \alpha_{2k} \tilde{y}_i'(t) + \alpha_{\sigma k} \sigma_i(t) \right), \tag{III.2}$$

with $\lambda_{0k}(t)$ the baseline risk function, W_i a vector of baseline covariates associated with the regression coefficient γ_k, and α_{1k}, α_{2k} and $\alpha_{\sigma k}$ the regression coefficients associated with the current value $\tilde{y}_i(t)$, the current slope $\tilde{y}_i'(t)$ and the current variability $\sigma_i(t)$ of the marker, respectively. Different parametric forms for the baseline risk function can be considered,

such as exponential, Weibull, or, for more flexibility, a B-splines base with Q knots defined by:

$$\log(\lambda_{0k}(t)) = \exp\left(\sum_{q=1}^{Q+4} \eta_{qk} B_q(t, \nu_k)\right),$$

where $B_q(t, \nu_k)$ is the q-th basis function of B-splines with the knot vector ν_k and η_{qk} is the associated parameter to be estimated.

III.2.2 Estimation procedure

Let θ be the set of parameters to be estimated including parameters of the Cholesky decomposition of the covariance matrix of the random effects, β, μ, $\alpha^\top = (\alpha_{11}, \alpha_{21}, \alpha_{\sigma 1}, \alpha_{12}, \alpha_{22}, \alpha_{\sigma 2})$, $\gamma^\top = (\gamma_1, \gamma_2)$ and the parameters of the two baseline risk functions. Considering the frequentist approach, the parameters are estimated by maximizing the likelihood function. The contribution of individual i to the marginal likelihood is defined by:

$$\begin{aligned}
\mathcal{L}_i(\theta; Y_i, T_i, \delta_i) &= \int p(Y_i, T_i, \delta_i | b_i, \tau_i; \theta) f(b_i, \tau_i; \theta) db_i d\tau_i \\
&= \int f(Y_i | b_i, \tau_i; \theta) \exp\left(-\sum_{k=1}^{2} \Lambda_{ik}(T_i | b_i, \tau_i; \theta)\right) \\
&\quad \times \prod_{k=1}^{2} \lambda_{ik}(T_i | b_i, \tau_i; \theta)^{\mathbb{1}_{\delta_i = k}} f(b_i, \tau_i; \theta) db_i d\tau_i,
\end{aligned}$$

with $f(b_i, \tau_i; \theta)$ a multivariate Gaussian density and $f(Y_i | b_i, \tau_i; \theta) = \prod_{j=1}^{n_i} f(Y_{ij} | b_i, \tau_i; \theta)$ where $f(Y_{ij} | b_i, \tau_i; \theta)$ is a univariate Gaussian density; $\Lambda_{ik}(T_i | b_i, \tau_i; \theta)$ is the cumulative risk function for event k ($k \in \{1, 2\}$) given by:

$$\Lambda_{ik}(t | b_i, \tau_i; \theta) = \int_0^t \lambda_{ik}(u | b_i, \tau_i; \theta) du \tag{III.3}$$

In cohort studies, data are frequently left-truncated. Left-truncation arises as soon as the time scale is not the time since inclusion and the subjects are enrolled only if they are free of the event at inclusion (Betensky and Mandel, 2015). This is the case in most studies where the time-scale is age. To deal with left-truncation (also called delayed entry), the individual contribution to the likelihood must be divided by the probability to be free of any event at entry time T_{0i}:

$$\mathcal{L}_i^{DE}(\theta; Y_i, T_i, \delta_i) = \frac{\mathcal{L}_i(\theta; Y_i, T_i, \delta_i)}{\int \exp(-\Lambda_{i1}(T_{0i} | b_i, \tau_i; \theta) - \Lambda_{i2}(T_{0i} | b_i, \tau_i; \theta)) f(b_i, \tau_i; \theta) db_i d\tau_i}$$

Because the integral on the random effects does not have an analytical solution, the integral is computed by a Quasi Monte Carlo (QMC) approximation (Pan and Thompson, 2007), using deterministic quasi-random sequences. The approximation of the integral is defined by:

$$\mathcal{L}_i(\theta; Y_i, T_i, \delta_i) \simeq \frac{1}{S} \sum_{s=1}^{S} p(Y_i, T_i, \delta_i | b_i^s, \tau_i^s; \theta)$$

where $(b_i^1, ..., b_i^S)$ and $(\tau_i^1, ..., \tau_i^S)$ are draws of a S-sample in the sobol sequel for the distribution $f(b_i, \tau_i; \theta)$. Moreover, to approximate the cumulative risk function given in equation (III.3), we use the Gauss-Kronrod quadrature approximation with 15 points (Gonnet, 2012).

Parameter estimation is obtained by maximizing the log-likelihood function:

$$\ell(\theta; Y_i, T_i, \delta_i) = \log \left(\prod_{i=1}^{N} \mathcal{L}_i(\theta; Y_i, T_i, \delta_i) \right).$$

The maximization is performed using the `marqLevAlg` R-package based on the Marquardt-Levenberg algorithm (Philipps et al., 2021). The latter is a robust variant of the Newton-Raphson algorithm (Levenberg, 1944; Marquardt, 1963) which iteratively updates the parameters θ to be estimated until convergence with the following formula at iteration $l+1$:

$$\theta^{(l+1)} = \theta^{(l)} - \psi_l (\tilde{H}(\theta^{(l)}))^{-1} \nabla(\ell(\theta^{(l)}))$$

where $\theta^{(l)}$ is the set of parameters at iteration l, $\nabla\left(\ell\left(\theta^{(l)}\right)\right)$ the gradient of the log-likelihood at iteration l and $\widetilde{H}\left(\theta^{(l)}\right)$ the inflated Hessian matrix where the diagonal terms of the Hessian matrix $H\left(\theta^{(l)}\right)$ are replaced by :

$$\widetilde{H}\left(\theta^{(l)}\right)_{ii} = H\left(\theta^{(l)}\right)_{ii} + \phi_l \left[(1 - \rho_l) \left| H\left(\theta^{(l)}\right)_{ii} \right| + \rho_l \operatorname{tr}\left(\theta^{(l)}\right) \right].$$

The scalars ψ_l, ϕ_l and ρ_l are internally determined at each iteration l to ensure that $\widetilde{H}\left(\theta^{(l)}\right)$ be definite-positive, $\widetilde{H}\left(\theta^{(l)}\right)$ approaches $H\left(\theta^{(l)}\right)$ when θ approaches $\widehat{\theta}$ and insure improvement of the likelihood at each iteration. Stringent convergence criteria are used, relying on parameter and function stability, and the relative distance to the maximum computed from the first and second derivatives of the log-likelihood which must not exceed a threshold ε_d: $\frac{\nabla\left(\ell\left(\theta^{(l)}\right)\right)\left(H\left(\theta^{(l)}\right)\right)^{-1}\nabla\left(\ell\left(\theta^{(l)}\right)\right)}{m} < \varepsilon_d$, with m the number of parameters. This algorithm was previously compared to other algorithms (EM, BFGS and L-BFGS-B) and the results showed that this algorithm was the most reliable (Philipps et al., 2021). Regarding the variances of the estimates, they are estimated by computing the inverse of the Hessian matrix computed by finite differences. The variances of the estimated parameters from the covari-

ance matrix of the random effects are computed using the Delta-Method (Meyer et al., 2013).

In the estimation procedure, the choice of the number of QMC draws is an important factor in practice. Indeed, increasing the number of QMC draws ensures good precision of parameter variances for statistical inference but at the cost of a considerable increase in computation time. To limit computation time and insure precise estimates of parameters and their standard error, we propose a two-step strategy for using the algorithm. We first applied the Marquardt-algorithm with a small number $S1$ of QMC draws (e.g. $S1 = 500$) until convergence is achieved. The first step provides a good parameter estimation in a reasonable time, but does not guarantee the accuracy of their estimated variance. In the second step, to improve the accuracy of the computation of the standard error of the estimates, a few additional iterations are performed with a higher number $S2 > S1$ of QMC draws (e.g. $S2 = 5000$) until the Hessian matrix is invertible. Note that for model selection based on likelihood or information criteria, using the results obtained in the first step is sufficient, given the good estimation of model parameters. However, for statistical inference on the final model, we recommend to increase the number of QMCs (step 2).

III.2.3 Individual Predictions

We implemented the computation of individual probability of having event k between time s and $s + t$ given that the subject i did not experience any event before time s, all marker measures collected until time s, $\mathcal{Y}_i(s)$, and the set of estimated parameters. The prediction is defined for subject i by:

$$\begin{aligned}\pi_i(s,t;\widehat{\theta}) &= P(s < T_i < s+t, \delta_i = k | T_i > s, \mathcal{Y}_i(s), \widehat{\theta}) \\ &= \frac{\int \left[\int_s^{s+t} \exp\left(-\sum_{c=1}^{2} \Lambda_{ic}(u|b_i,\tau_i,\widehat{\theta})\right)\lambda_{ik}(u|b_i,\tau_i,\widehat{\theta})du\right] f(\mathcal{Y}_i(s)|b_i,\tau_i,\widehat{\theta}) f(b_i,\tau_i|\widehat{\theta}) db_i d\tau_i}{\int \exp\left(-\sum_{c=1}^{2} \Lambda_{ic}(s|b_i,\tau_i,\widehat{\theta})\right) f(\mathcal{Y}_i(s)|b_i,\tau_i,\widehat{\theta}) f(b_i,\tau_i|\widehat{\theta}) db_i d\tau_i}\end{aligned} \tag{III.4}$$

As previously, the integral over the random effect is computed by QMC approximation and the integral over time with the Gauss-Kronrod quadrature.

The corresponding 95% confidence interval of predictions is obtained by the following Monte Carlo algorithm. For L large enough and $l = 1, ..., L$ ($L = 1000$ for instance):

— Generate $\widetilde{\theta}^{(l)} \sim \mathcal{N}(\widehat{\theta}, V(\widehat{\theta}))$ where $V(\widehat{\theta})$ is given by the inverse of the Hessian matrix at $\widehat{\theta}$;

— Compute $\widetilde{\pi}_i^{(l)}(s, t; \widetilde{\theta}^{(l)})$ from equation (III.4);

— Compute the 95% confidence interval from the 2.5th and 97.5th percentiles of the L-sample of $\widetilde{\pi}_i^{(l)}(s, t; \widetilde{\theta}^{(l)})$.

III.2.4 Software

The `R`-package `FlexVarJM` has been developed for the estimation of the model, the prediction of the subject-specific random effects, and the computation of the individual predicted probabilities of events. The package allows estimation of a model with an unconstrained time-trend for the marker trajectory, one or two events with exponential, Weibull or B-splines baseline risk functions, and a flexible dependence structure between the events and the marker (possibly including the current value, the current slope and the subject-specific time-dependent variability). The development version of `FlexVarJM` is available on Github at the following link: `https://github.com/LeonieCourcoul/FlexVarJM` and the fixed version can be installed from CRAN (Courcoul et al., 2023).

III.3 Simulations

In order to evaluate the performance of the estimation procedure, we performed a simulation study using a design similar to the application data.

III.3.1 Design of simulations

Visit times were generated using a uniform distribution centered around each specified time, with a variation of approximately one month in either direction. For each visit time, one measurement of the marker was generated, using a linear mixed-effects model with fixed and random intercept and slope, and heterogeneous variance:

$$\begin{cases} Y(t_{ij}) = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) \times t_{ij} + \epsilon_i(t_{ij}) \\ \epsilon_i(t_{ij}) \sim \mathcal{N}(0, \sigma_i^2(t_{ij})) \quad \text{with} \quad \log(\sigma_i(t_{ij})) = \mu_0 + \tau_{0i} + (\mu_1 + \tau_{1i}) \times t_{ij} \end{cases} \tag{III.5}$$

with $b_i = (b_{0i}, b_{1i})^\top$ and $\tau_i = (\tau_{0i}, \tau_{1i})^\top$. Competing event times T_{ik}^* $(k = 1, 2)$ were generated using the Brent's univariate root-finding method (Brent, 1973) according to the following proportional hazards models:

$$\lambda_{ik}(t) = \lambda_{0k}(t)\exp(\alpha_{1k}\tilde{y}_i(t) + \alpha_{2k}\tilde{y}'_i(t) + \alpha_{\sigma k}\sigma_i(t)) \tag{III.6}$$

with $\lambda_{0k}(t) = \kappa_k t^{\kappa_k - 1} e^{\zeta_{0k}}$ being a Weibull function. Individuals were censored at C_i the last visit observed in the dataset. Finally, the observed time was defined by $T_i = \min(T^*_{i1}, T^*_{i2}, C_i)$. Measures of the marker Y posterior to T_i were removed from the datasets.

The aim of our simulation study was to evaluate the convergence properties of the proposed estimation procedure for the location-scale joint model. Specifically, we aimed to assess how the procedure performs under different conditions related to the number of repeated measurements and the correlation structure between the random effects. Five distinct scenarios were considered to address various practical situations and to investigate the robustness and accuracy of the estimation procedure:

— Scenario A: a maximum of 7 times of measurement, at 0-year, 0.5 year and then one per year until 5 years; with random effects b_i independent of τ_i.

— Scenario B: a maximum of 13 times of measurements, at 0 year, every 3 months the first year and then twice per year until 5 years (mimicking PROGRESS); with random effects b_i independent of τ_i.

— Scenario C: same time points as in Scenario A; with correlated random effects.

— Scenario D: same time points as in Scenario B; with correlated random effects.

— Scenario E (Misspecified model): marker generated with a quadratic trend but estimated with a linear trend; one event; same time points as in Scenario B; with random effects b_i independent of τ_i. This scenario was performed to assess the effect of a misspecification of the blood pressure time-trend on the estimation of the variance effect on the event risk.

For each scenario, 300 datasets of 500 and 1000 subjects were generated. Parameter values for data generation are indicated in the tables of results. The models were estimated with the two-step estimation procedure presented in Section III.2.2, given $S1 = 500$ and $S2 = 5000$ draws for the QMC integration approximation.

III.3.2 Results

Tables III.1, III.2 and III.3 report the mean estimates, the empirical and mean asymptotic standard error of the estimated parameters and the coverage rate of their 95% confidence intervals for scenario A, B and C on 500 individuals. Results for scenario D and for larger

sample size are in the Supporting Information (section III.6, Tables III.5 to III.9). The estimation procedure provided satisfactory results for the four scenarios of simulations. Indeed, the bias was minimal, the mean asymptotic and the empirical standard errors were close, and the coverage rates of the 95% confidence interval were close to the nominal value. We can note that the bias is minimal from the first step but the second step helps to reduce the difference between the mean asymptotic and the empirical standard deviations and thus improve the coverage rates. This simulation study also illustrates the impact of the choice of $S1$ and $S2$ on the computation time: for scenario A, the medians of computation time are around 13 minutes (25 iterations in median) and 9 minutes (1 iteration in median) respectively for step 1 and 2, on 10 cores. Finally, Scenario E was performed to evaluate the impact of a misspecified marker trajectory (quadratic versus linear time trend). As expected, the estimates of the fixed effects in the mixed model are biased, but the estimates of the model for the residual variance and of the time-to-event model are robust (Table III.10 in Supporting Information, section III.6).

III.4 Application

III.4.1 PROGRESS clinical trial

To illustrate the proposed model, we estimated the proposed model on the data from the PROGRESS clinical trial (Mac Mahon et al., 2001) designed to evaluate a blood-pressure lowering treatment in secondary prevention. PROGRESS is a multicentre, double-blind randomized clinical trial including patients with a history of stroke or transient ischaemic attack within 5 years before inclusion. Patients were recruited between May 1995 and November 1997. The follow-up comprised five visits in the first year, then two visits each years until the end of the study or the occurrence of a major CVD (stroke, myocardial infarction and cerebral hemorrhage) or death. At each visit, blood pressure was measured twice and we analysed the mean of the two measurements at each time. Prior to randomization, eligible patients were subjected to a 4-week run-in phase to test their tolerance to the treatment. At randomization, patients assigned to the control group stopped the treatment. In order to avoid an effect of the change of therapy at randomization, we removed the blood pressure measure at randomization. The model was estimated only on Non Asian subjects due to differences between Asians and non Asians with regard to CVD risk and risk factors and because the treatment was not exactly the same. Finally, the current study was conducted over 3710 Non Asian patients, 1856 for the controlled group and 1854 for the treatment

Table III.1 – Simulation results for scenario A with 500 subjects (7 measures, b_i and τ_i independent).*

Parameter			Step 1				Step 2			
		True value	**Mean**	**Empirical SE**	**Mean asymptotic SE**	**Coverage rate (%)**	**Mean**	**Empirical SE**	**Mean asymptotic SE**	**Coverage rate (%)**
Longitudinal submodel										
Intercept	β_0	142	142.1	0.736	0.717	94.3	142.1	0.730	0.728	95.0
Slope	β_1	3	2.943	0.288	0.253	90.0	2.945	0.280	0.271	92.6
Variability	μ_0	2.4	2.396	0.027	0.026	94.7	2.395	0.027	0.027	95.3
	μ_1	0.05	0.050	0.017	0.015	93.0	0.051	0.016	0.016	94.3
Survival submodel 1										
Current variance	$\alpha_{\sigma 1}$	0.07	0.064	0.041	0.039	94.7	0.064	0.041	0.039	95.3
Current value	α_{11}	0.02	0.020	0.008	0.007	95.7	0.020	0.008	0.007	95.3
Current slope	α_{21}	0.01	0.008	0.072	0.066	94.3	0.007	0.071	0.067	96.0
Weibull	$\sqrt{\kappa_1}$	1.1	1.099	0.056	0.059	97.7	1.098	0.056	0.059	98.0
	ζ_{01}	-7	-6.884	1.257	1.199	94.3	-6.885	1.257	1.204	94.7
Survival submodel 2										
Current variance	$\alpha_{\sigma 2}$	0.15	0.169	0.091	0.046	92.3	1.678	0.091	0.054	96.3
Current value	α_{12}	-0.01	-0.012	0.014	0.009	96.3	-0.012	0.015	0.010	96.7
Current slope	α_{22}	-0.14	-0.171	0.173	0.086	92.6	-0.167	0.170	0.095	94.7
Weibull	$\sqrt{\kappa_2}$	1.3	1.310	0.102	0.079	97.0	1.311	0.105	0.084	98.0
	ζ_{02}	-4	-4.075	1.389	1.366	96.0	-4.075	1.388	1.401	95.7

SE: Standard Error; Coverage rate: coverage rate of the 95% confidence interval.
* Results for 299 replicates with complete convergence over 300.

Table III.2 – Simulation results for scenario B with 500 subjects (13 measures, b_i and τ_i independent).*

Parameter			Step 1				Step 2			
		True value	**Mean**	**Empirical SE**	**Mean asymptotic SE**	**Coverage rate (%)**	**Mean**	**Empirical SE**	**Mean asymptotic SE**	**Coverage rate (%)**
Longitudinal submodel										
Intercept	β_0	142	142.0	0.779	0.655	90.7	142.0	0.767	0.721	92.7
Slope	β_1	3	2.996	0.252	0.201	90.0	3.000	0.248	0.235	93.3
Variability	μ_0	2.4	2.402	0.019	0.018	92.0	2.401	0.019	0.018	92.7
	μ_1	0.05	0.047	0.012	0.011	92.0	0.049	0.012	0.012	94.0
Survival submodel 1										
Current variance	$\alpha_{\sigma 1}$	0.07	0.063	0.030	0.028	92.7	0.065	0.029	0.028	93.7
Current value	α_{11}	0.02	0.020	0.006	0.007	96.3	0.020	0.006	0.007	96.3
Current slope	α_{21}	0.01	0.007	0.057	0.053	96.3	0.007	0.055	0.055	96.0
Weibull	$\sqrt{\kappa_1}$	1.1	1.106	0.055	0.054	94.0	1.105	0.055	0.055	95.0
	ζ_{01}	-7	-6.912	0.992	1.042	95.7	-6.914	0.992	1.050	95.7
Survival submodel 2										
Current variance	$\alpha_{\sigma 2}$	0.15	0.158	0.034	0.030	92.7	0.159	0.031	0.032	97.3
Current value	α_{12}	-0.01	-0.010	0.008	0.008	94.7	-0.010	0.008	0.008	95.3
Current slope	α_{22}	-0.14	-0.146	0.064	0.061	94.0	-0.145	0.063	0.064	95.0
Weibull	$\sqrt{\kappa_2}$	1.3	1.299	0.067	0.067	95.7	1.299	0.066	0.068	96.3
	ζ_{02}	-4	-4.104	1.173	1.111	93.7	-4.107	1.173	1.139	93.7

SE: Standard Error; Coverage rate: coverage rate of the 95% confidence interval.
* Results for 300 replicates with complete convergence over 300.

Table III.3 – Simulation results for scenario C with 500 subjects (7 measures, b_i and τ_i correlated).*

Parameter			Step 1				Step 2			
		True value	Mean	Empirical SE	Mean asymptotic SE	Coverage rate (%)	Mean	Empirical SE	Mean asymptotic SE	Coverage rate (%)
Longitudinal submodel										
Intercept	β_0	142	141.9	0.833	0.742	92.3	141.9	0.820	0.756	93.3
Slope	β_1	3	3.019	0.320	0.282	90.6	3.023	0.314	0.290	92.0
Variability	μ_0	2.4	2.401	0.035	0.033	93.6	2.399	0.035	0.033	94.3
	μ_1	0.05	0.050	0.016	0.015	92.6	0.050	0.016	0.016	93.7
Survival submodel 1										
Current variance	$\alpha_{\sigma 1}$	0.07	0.065	0.051	0.046	93.3	0.067	0.050	0.047	94.7
Current value	α_{11}	0.02	0.021	0.012	0.011	93.3	0.021	0.012	0.011	93.3
Current slope	α_{21}	0.01	-0.004	0.077	0.076	93.6	-0.051	0.817	0.470	94.0
Weibull	$\sqrt{\kappa_1}$	1.1	1.094	0.052	0.054	96.3	1.095	0.052	0.055	96.7
	ζ_{01}	-7	-7.133	1.437	1.296	94.0	-6.958	3.204	2.612	94.7
Survival submodel 2										
Current variance	$\alpha_{\sigma 2}$	0.15	0.166	0.079	0.051	91.3	0.165	0.077	0.057	95.0
Current value	α_{12}	-0.01	-0.016	0.018	0.013	92.0	-0.012	0.017	0.014	93.7
Current slope	α_{22}	-0.14	-0.154	0.113	0.088	95.6	-0.178	0.486	0.253	95.7
Weibull	$\sqrt{\kappa_2}$	1.3	1.314	0.078	0.069	94.3	1.314	0.074	0.072	94.7
	ζ_{02}	-4	-4.084	1.758	1.524	94.3	-3.983	2.469	2.109	95.0

SE: Standard Error; Coverage rate: coverage rate of the 95% confidence interval.
* Results for 298 replicates with complete convergence over 300 for step 1
and for 300 replicates with complete convergence over 300 for step 2.

group, and included 672 CVD and 150 deaths without CVD, resulting in a 78% censoring rate. There are 2525 (68%) men and 1185 (32%) women. The average age at entry in the study is 67 years old (sd = 9.8) with a minimum at 26 and a maximum at 91 years old.

III.4.2 Specification of the model

This study aimed to evaluate the impact of the blood pressure variability on the risk of CVD and death from other causes. To do so, we estimated the proposed joint model (Model CVCS+V, for current value, current slope and variance) with heterogeneous time-dependent variance defined by (III.1) and (III.2) using the time since the first considered blood pressure measurement. The trajectory of blood pressure was described over time by a linear mixed effect model. The individual time trend of the marker and the variance were modelled by a linear trend. The baseline hazard functions of both events were defined by B-splines with three interior knots placed at the quantiles of the observed events. According to the AIC, the model with three knots was better than models with 1 or 5 knots for each baseline hazard function (respectively 298946.2, 298948.8 and 298951). The model allowed the risk of each event to depend on the time-dependent intra-subject variability, the individual current value and the current slope. The longitudinal submodel and the variance submodel were adjusted for treatment group and survival submodels were adjusted for treatment group, age at baseline and sex (male versus female):

$$\begin{cases} y(t_{ij}) = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) \times t_{ij} + \beta_2 \times trt_i + \epsilon_i(t_{ij}) \\ \epsilon_i(t_{ij}) \sim \mathcal{N}(0, \sigma_i^2(t_{ij})) \quad \text{with} \quad \log(\sigma_i(t_{ij})) = \mu_0 + \tau_{0i} + (\mu_1 + \tau_{1i}) \times t_{ij} + \mu_2 \times trt_i \\ \lambda_{ik}(t) = \lambda_{0k}(t) \exp(\gamma_{0k} trt_i + \gamma_{1k} male_i + \gamma_{2k} age_i + \alpha_{1k} \tilde{y}_i(t) + \alpha_{2k} \tilde{y}_i'(t) + \alpha_{\sigma k} \sigma_i(t)) \end{cases}$$

The estimation was performed with $S1 = 500$ and $S2 = 10000$ draws of QMC to ensure a greater accuracy.

This model was compared to two classical joint model without heterogeneous variance , i.e. $\sigma_i^2(t_{ij}) = \sigma^2$ for all $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$. The first one allowed the risk of each event to depend only on the individual current value (Model CV) and the second one on both the individual current value and current slope (Model CVCS).

III.4.3 Results

The AIC from the complete model (298946.2) was clearly better than the AIC from the two joint models with a constant residual variance and a dependence either on the current value only (302062.4) or on both the current value and the current slope (302011.6), showing the importance of taking into account a time-dependent subject-specific variance.

Table III.4 provides estimates from the complete joint model and Table III.7 presents the covariance matrix of the random effects and their standard errors computed through the Delta-Method. Blood pressure was lower for individuals from the treatment group ($\widehat{\beta_2} = -8.03$, $p - value < 0.001$). The variance of the residual error was heterogeneous between the subjects ($\widehat{Var}(\tau_{0i}) = 0.13$, $sd = 7e-3$) and was lower for treated patients ($\widehat{\mu_2} = -0.030$, $p - value = 0.028$). The risk of CVD events increased with age ($\widehat{HR} = 1.04$ for one year, $\widehat{IC} = [1.03; 1.05]$) and was higher for men ($\widehat{HR} = 1.34$, $\widehat{IC} = [1.13; 1.60]$). Adjusting for age, sex and treatment group, the risk of CVD increased with the current blood pressure variance ($\widehat{HR} = 1.07$, $\widehat{IC} = [1.03; 1.10]$): but more surprisingly, the risk decreased with the current slope of blood pressure ($\widehat{HR} = 0.86$, $\widehat{IC} = [0.82; 0.90]$). A possible explanation for this unexpected result could be that subjects with decreasing blood pressure are those with the highest pre-treatment level of blood pressure. It's important to emphasize that the negative association with the slope is not due to the adjustment for blood pressure variance, as the same result is observed in the CVCS model (see Table III.11 in Supporting Information, section III.6). Adjusting for the variance and the slope of blood pressure, the risk of CVD did not depend on the current blood pressure value ($\widehat{HR} = 0.99$, $\widehat{IC} = [0.986; 1.002]$). The risk of death from other causes was higher for older individuals ($\widehat{HR} = 1.05$, $\widehat{IC} = [1.03; 1.07]$ for one year) and for men ($\widehat{HR} = 1.69$, $\widehat{IC} = [1.16; 2.48]$) but it was not associated with the treatment group. More importantly, the death risk significantly increased with the current variance of blood pressure ($\widehat{HR} = 1.13$, $\widehat{IC} = [1.05; 1.21]$) and decreased when the current value ($\widehat{HR} = 0.90$, $\widehat{IC} = [0.816; 0.993]$ for 5 mmHg) and the current slope of blood pressure increased ($\widehat{HR} = 0.89$, $\widehat{IC} = [0.798; 0.998]$). As for CVD, these unexpected results regarding the current value and the current slope are also observed without adjustment on the current blood pressure variance (CVCS model in Table III.11 of Supporting Information, section III.6). However, these results must be interpreted cautiously considering that the study population consists only of subjects who have survived a stroke and are included in a clinical trial, thus benefiting from close monitoring.

Table III.4 – Parameter estimates of the joint model on the Progress clinical trial data (CVCS+V model).

Parameter	**Estimate**	**Standard error**	**p-value**
Survival submodel for CVD			
BP current variance	0.064	0.017	< 0.001
BP current value	-0.006	0.004	0.160
BP current slope	-0.152	0.022	< 0.001
treatment group	-0.153	0.085	0.073
male	0.296	0.089	< 0.001
age	0.038	0.005	< 0.001
Survival submodel for Death			
BP current variance	0.120	0.035	< 0.001
BP current value	-0.021	0.010	0.030
BP current slope	-0.114	0.057	0.045
treatment group	-0.117	0.171	0.493
male	0.527	0.194	0.006
age	0.051	0.010	< 0.001
Longitudinal submodel			
Blood Pressure Mean			
intercept	142.5	0.330	< 0.001
time	-0.104	0.072	0.150
treatment group	-8.029	0.441	< 0.001
Blood Pressure Residual Variance			
intercept	2.341	0.012	< 0.001
time	0.007	0.004	0.086
treatment group	-0.030	0.014	0.028

BP: Blood Pressure

$$\widehat{\Sigma} = \begin{bmatrix} \widehat{Var}(b_{0i}) & & & \\ \widehat{Cov}(b_{0i}, b_{1i}) & \widehat{Var}(b_{1i}) & & \\ \widehat{Cov}(b_{0i}, \tau_{0i}) & \widehat{Cov}(b_{1i}, \tau_{0i}) & \widehat{Var}(\tau_{0i}) & \\ \widehat{Cov}(b_{0i}, \tau_{1i}) & \widehat{Cov}(b_{1i}, \tau_{1i}) & \widehat{Cov}(\tau_{0i}, \tau_{1i}) & \widehat{Var}(\tau_{1i}) \end{bmatrix} = \begin{bmatrix} 212.3_{(5.6)} & & & \\ -19.0_{(1.2)} & 8.6_{(0.4)} & & \\ 2.2_{(0.15)} & -0.29_{(0.04)} & 0.13_{(7e-3)} & \\ -0.18_{(0.06)} & 0.16_{(0.02)} & -0.02_{(3e-3)} & 0.012_{(1e-3)} \end{bmatrix} \quad \text{(III.7)}$$

III.4.4 Goodness-of-fit assessment

To assess the fit of the time-to-event submodels, we computed for each event, the predicted cumulative hazard function at each event time by plugging the empirical Bayes estimates of the random effects in the formula for the risk function. Then we compared the mean of this predicted cumulative hazard function with its Nelson-Aalen estimator for the whole sample (Figure III.4 of the Supporting Information, section III.6) and stratified according to sex and randomization group (Figure III.5). These figures show that the joint model adequately fitted both risks and that the proportional risk assumption was valid for each categorical variable.

To highlight the impact of adding a subject-specific and time-dependent residual variance in the mixed model, we computed the individual predictions of the marker over time for some selected subjects. The predicted value of blood pressure corresponds to the conditional expectation given the random effects, defined by $\hat{\mathbb{E}}(Y_i(t)|\widetilde{b}_i, \widetilde{\tau}_i)$ and the prediction interval around this predicted values is given by $\hat{\mathbb{E}}(Y_i(t)|\widetilde{b}_i, \widetilde{\tau}_i) \pm 1.96\sqrt{\hat{\mathbb{V}}(Y_i(t)|\widetilde{b}_i, \widetilde{\tau}_i)}$. For each subject the empirical Bayes estimates of the random effects, denoted by $(\widetilde{b}_i, \widetilde{\tau}_i) = \text{argmax} f(b_i, \tau_i|Y_i, T_i, \delta_i)$, corresponds to the mode of their estimated conditional posterior given the data. They are computed by maximising $f(Y_i, T_i, \delta_i)f(b_i, \tau_i)$ with the Marquardt-Levenberg algorithm.

For some selected subjects still at risk at 3 years, Figure III.1 presents the predicted values and their confidence intervals from the models with and without subject-specific residual variance (CVCS+V and CVCS). It shows that assuming a time-dependent and subject-specific residual variability allows a better fit of the uncertainty around the individual prediction.

III.4.5 Predictions

We compared the predictive abilities of models with and without time-dependent individual variability using AUC using a 5-fold cross-validation. The individual predictions of having CVD (or death) between 3 and 5 years for subjects free of any event at 3 years were computed using equation (III.4). The AUC was computed using the `timeROC` package (Blanche et al., 2015). The results are slightly better for the model with heterogeneous variability. We obtained respectively 0.609 (0.067) and 0.576 (0.067) for the risk of CVD, and 0.637 (0.078) and 0.616 (0.079) for the risk of death.

To illustrate the effect of taking into account the current value of individual variance, we

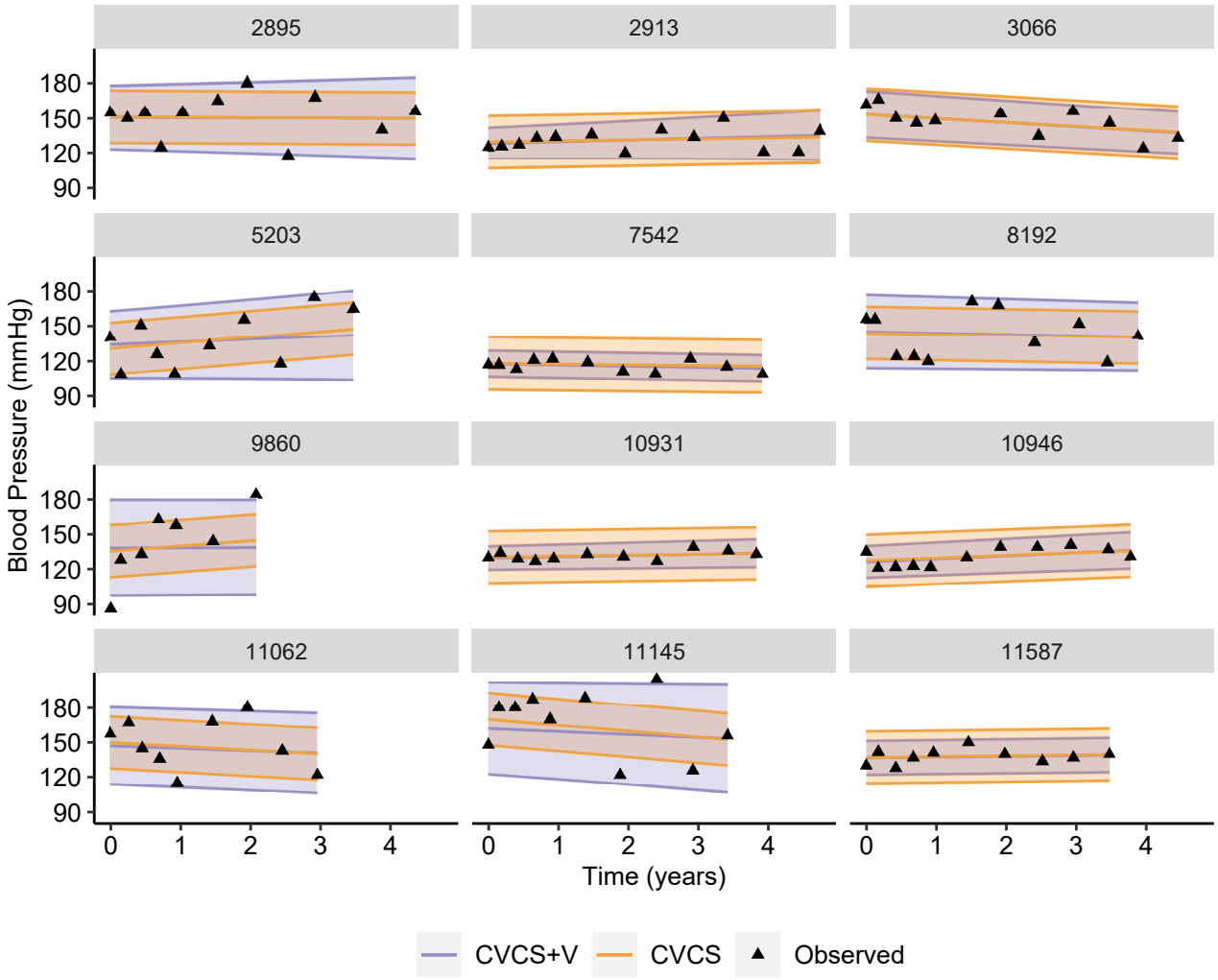

Figure III.1 – Prediction over time of the individual blood pressure and its prediction interval at 95% for 12 subjects. Model CVCS+V assumed a time-dependent subject-specific variability and the model CVCS assumes a homogeneous and constant variability. The black triangles are the observed measurements.

also computed the predicted risk of the events between 3 and 5 years for different subjects from both models, with and without time-dependent individual variability. We used the subjects selected for Figure III.1 and present their predictions obtained via the cross-validation procedure. Figures III.2 and III.3 shows that, for both the risk of CVD and the risk of death, the predicted probability is higher with the complete model when the individual experienced the event between 3 and 5 years than with the model without the heterogeneous variability. Conversely, the predicted risk is smaller with the complete model when the individual do not experience the corresponding event.

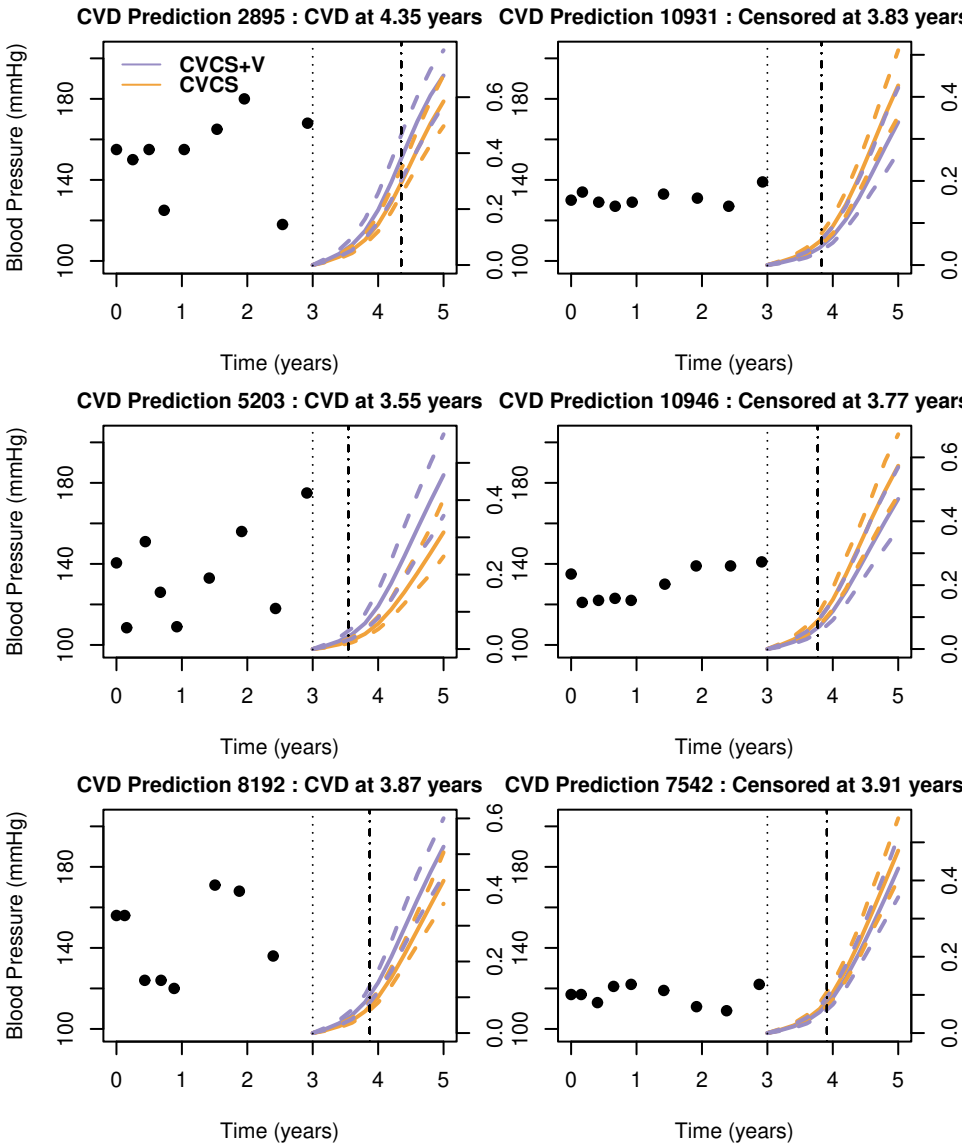

Figure III.2 – Prediction of the risk of CVD between 3 and 5 years (with its 95% confidence interval indicated by dashed lines), for six patients at risk at 3 years, for Model CVCS+V (purple) and Model CVCS (orange). The dashed lines represent the observed time.

III.5 Discussion

In this work, we have proposed a new joint model with a subject-specific time-dependent variance that extends the models proposed by Gao et al. (2011) and Barrett et al. (2019). Indeed, this new model allows time and covariate dependent individual variance and a flexible dependence structure between the competing events and the longitudinal marker. In particular, the risk of events may depend on both the current value and the current slope of the marker, in addition to the subject-specific time-dependent standard deviation of the residual

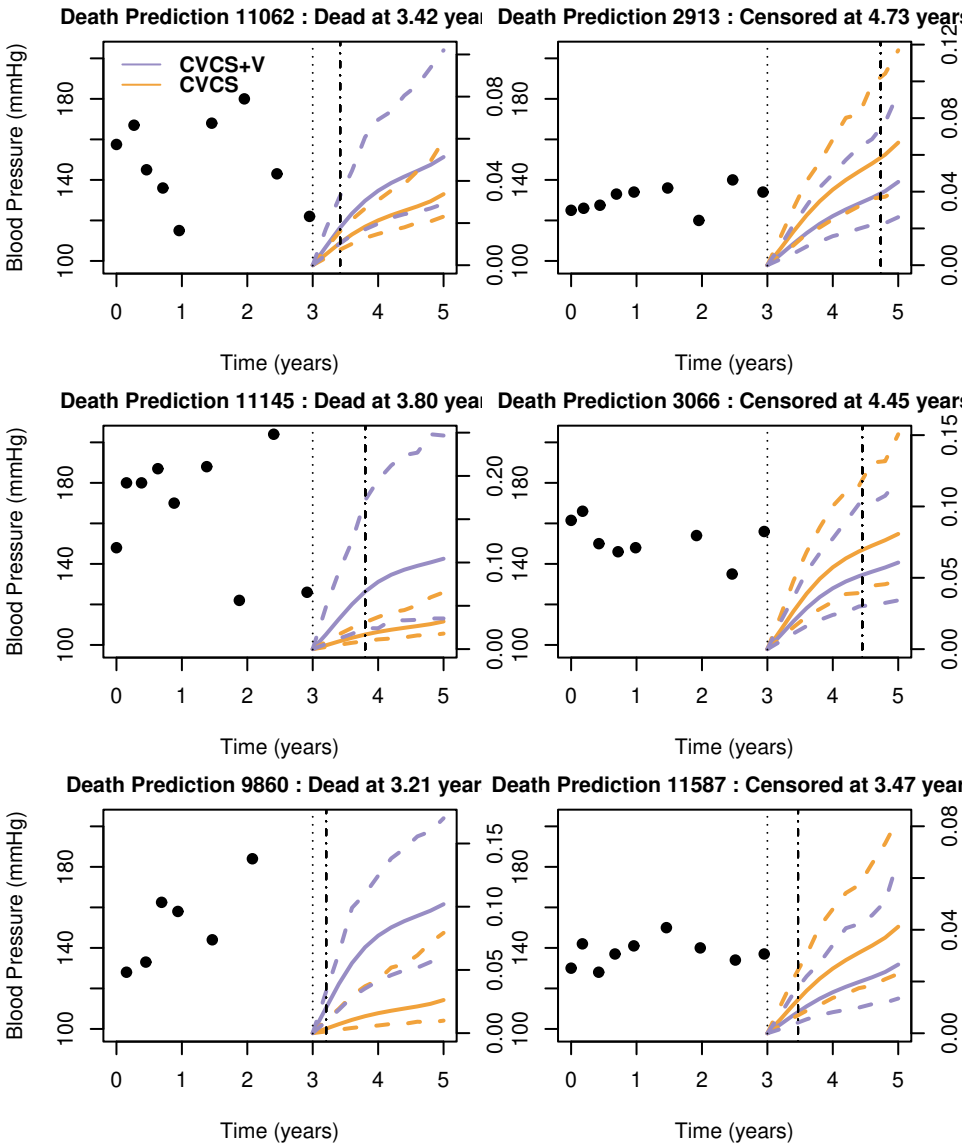

Figure III.3 – Prediction of the risk of Death between 3 and 5 years (with its 95% confidence interval indicated by dashed lines), for six patients at risk at 3 years, for Model CVCS+V (purple) and Model CVCS (orange). The dashed lines represent the observed time.

error. This is an important asset of the model given that, in most health research contexts, it is more sensible to assume that the event risk depends on the time-dependent current value or slope of the marker instead of only time-independent random effects. Moreover, accounting for competing events may be important in many clinical applications. Simulation study allows us to demonstrate the good performance of the estimation procedure and to study the impact of the choice of $S1$ and $S2$. The model converged without bias and with good coverage rates, whatever the number of individual and the number of visits. Moreover, the estimates of the time-to-event sub-model are quite robust to a misspecification of the marker

trajectory. In addition, we provided an R-package that allows frequentist estimation with a robust estimation algorithm which had shown very good behaviour in our simulations and in a previous work with different models (Philipps et al., 2021).

The analysis of the PROGRESS trial has shown that a high variability of blood pressure is associated with a high risk of CVD and death from other causes. Moreover, the individual residual variability depends on treatment group. These results were obtained on a population with an history of stroke. It would be very interesting to replicate this kind of analysis on representative samples of the general population to evaluate if blood pressure variability could be a target of primary prevention for CVD, death or stroke.

The proposed model is flexible enough to handle time-varying and subject-specific residual variance but it may be applied assuming a time-fixed subject-specific residual variance only. We have not considered more flexible time-trend for the variance in the application because, in most applications, the number of repeated measurements is not large enough for estimating such a flexible model. However, the use of more flexible time-trend is possible in the package.

In this work, we have supposed that the visit times were not informative and that missing measurements before the event were missing at random. In the PROGRESS clinical study, these hypotheses are quite plausible since visits were planed following a pre-specified protocol and the rate of missed visits before the event was low (less than 3%). For application to observational studies, it could be useful to extend this approach to consider an informative observation process. However, such a model would require three submodels: a mixed model for the evolution of the marker, a submodel for repeated events to describe the visit process and a model for the competing events of interest. This model would rely on non-verifiable parametric assumptions and its estimation process would be much more cumbersome.

The proposed approach addressed both right censoring and left-truncation, the two most common observation schemes for time-to-event data. Considering interval censoring and semi-competing events could represent a valuable enhancement. This extension would be useful when the exact time of onset of the main event is unknown (dementia for instance) and the competing event may arise after the main event (death). However, this would necessitate modeling the three transition intensities and the interval censoring would significantly complicate the computation of the likelihood.

Such joint models with dependence on the heterogeneous variance (that can be viewed as an extension of the location-scale mixed model (Hedeker and Nordgren, 2013)) are of great interest to investigate the association between the variability of markers or risk factors and the risk of health events in various fields of medical research, possibly allowing to improve the prediction ability for the event. For instance, hypotheses have emerged about the link

between emotional instability and the risk of psychiatric events, or the variability of glycemia and the prognosis of diabetes. Thanks to wearable devices, recent medical research studies often include frequent repeated measures of exposures or biomarkers, allowing the investigation of hypotheses regarding the variability.

III.6 Supporting Information

Table III.5 – Simulation results for scenario A with 1000 subjects (7 measures, b_i and τ_i independent).*

Parameter			Step 1				Step 2			
		True value	Mean	Empirical SE	Mean asymptotic SE	Coverage rate (%)	Mean	Empirical SE	Mean asymptotic SE	Coverage rate (%)
Longitudinal submodel										
Intercept	β_0	142	142.0	0.533	0.509	94.97	142.0	0.530	0.515	95.33
Slope	β_1	3	2.987	0.189	0.181	94.63	2.985	0.184	0.194	96.33
Variability	μ_0	2.4	2.399	0.019	0.019	94.30	2.398	0.019	0.019	94.00
	μ_1	0.05	0.050	0.012	0.011	91.95	0.050	0.012	0.012	94.67
Survival submodel 1										
Current variance	$\alpha_{\sigma 1}$	0.07	0.066	0.028	0.026	94.63	0.067	0.027	0.026	94.67
Current value	α_{11}	0.02	0.020	0.005	0.005	94.63	0.020	0.006	0.005	94.33
Current slope	α_{21}	0.01	0.012	0.045	0.046	95.64	0.011	0.045	0.047	95.00
Weibull	$\sqrt{\kappa_1}$	1.1	1.094	0.042	0.041	93.96	1.094	0.042	0.041	94.33
	ζ_{01}	-7	-6.991	0.837	0.820	95.30	-6.990	0.838	0.825	95.00
Survival submodel 2										
Current variance	$\alpha_{\sigma 2}$	0.15	0.155	0.034	0.029	90.94	0.155	0.033	0.032	94.33
Current value	α_{12}	-0.01	-0.010	0.006	0.006	94.30	-0.010	0.006	0.006	94.33
Current slope	α_{22}	-0.14	-0.143	0.060	0.054	95.97	-0.143	0.058	0.057	96.67
Weibull	$\sqrt{\kappa_2}$	1.3	1.297	0.054	0.053	94.97	1.297	0.054	0.055	96.00
	ζ_{02}	-4	-4.035	0.919	0.897	96.31	-4.035	0.915	0.916	97.67

SE: Standard Error; Coverage rate: coverage rate of the 95% confidence interval.
* Results for 300 replicates with complete convergence over 300.

Table III.6 – Simulation results for scenario B with 1000 subjects (13 measures, b_i and τ_i independent).*

Parameter			Step 1				Step 2			
		True value	Mean	Empirical SE	Mean asymptotic SE	Coverage rate (%)	Mean	Empirical SE	Mean asymptotic SE	Coverage rate (%)
Longitudinal submodel										
Intercept	β_0	142	142.0	0.544	0.476	91.30	142.0	0.539	0.500	93.33
Slope	β_1	3	2.993	0.184	0.149	89.97	2.992	0.174	0.162	93.33
Variability	μ_0	2.4	2.401	0.012	0.013	94.98	2.400	0.012	0.013	95.33
	μ_1	0.05	0.048	0.008	0.008	92.31	0.049	0.008	0.008	93.33
Survival submodel 1										
Current variance	$\alpha_{\sigma 1}$	0.07	0.063	0.020	0.019	92.31	0.064	0.020	0.019	93.67
Current value	α_{11}	0.02	0.020	0.005	0.005	96.32	0.020	0.005	0.005	96.33
Current slope	α_{21}	0.01	0.010	0.041	0.038	91.97	0.009	0.041	0.039	93.67
Weibull	$\sqrt{\kappa_1}$	1.1	1.098	0.036	0.038	95.99	1.097	0.036	0.038	95.67
	ζ_{01}	-7	-6.934	0.703	0.722	96.32	-6.934	0.702	0.724	96.00
Survival submodel 2										
Current variance	$\alpha_{\sigma 2}$	0.15	0.154	0.022	0.020	93.31	0.154	0.021	0.021	95.33
Current value	α_{12}	-0.01	-0.010	0.005	0.005	95.65	-0.010	0.005	0.005	96.00
Current slope	α_{22}	-0.14	-0.149	0.050	0.042	92.98	-0.148	0.044	0.043	97.00
Weibull	$\sqrt{\kappa_2}$	1.3	1.304	0.048	0.047	95.65	1.303	0.048	0.047	95.67
	ζ_{02}	-4	-4.123	0.765	0.776	93.65	-4.124	0.763	0.786	95.00

SE: Standard Error; Coverage rate: coverage rate of the 95% confidence interval.
* Results for 300 replicates with complete convergence over 300.

Table III.7 – Simulation results for scenario C with 1000 subjects (7 measures, b_i and τ_i correlated).*

Parameter			Step 1				Step 2			
		True value	Mean	Empirical SE	Mean asymptotic SE	Coverage rate (%)	Mean	Empirical SE	Mean asymptotic SE	Coverage rate (%)
Longitudinal submodel										
Intercept	β_0	142	141.9	0.564	0.529	93.31	141.9	0.554	0.535	94.00
Slope	β_1	3	3.006	0.215	0.202	93.98	3.002	0.206	0.210	95.67
Variability	μ_0	2.4	2.402	0.024	0.023	93.65	2.400	0.024	0.023	95.33
	μ_1	0.05	0.049	0.012	0.011	93.98	0.050	0.011	0.011	95.33
Survival submodel 1										
Current variance	$\alpha_{\sigma 1}$	0.07	0.066	0.034	0.032	94.98	0.067	0.033	0.031	94.00
Current value	α_{11}	0.02	0.021	0.008	0.007	92.64	0.021	0.008	0.007	92.67
Current slope	α_{21}	0.01	-0.004	0.055	0.051	93.65	-0.005	0.053	0.054	93.67
Weibull	$\sqrt{\kappa_1}$	1.1	1.092	0.037	0.037	94.98	1.092	0.037	0.039	95.00
	ζ_{01}	-7	-7.123	0.960	0.883	94.31	-7.122	0.959	0.902	94.33
Survival submodel 2										
Current variance	$\alpha_{\sigma 2}$	0.15	0.154	0.034	0.034	94.98	0.154	0.034	0.034	96.67
Current value	α_{12}	-0.01	-0.010	0.009	0.009	95.99	-0.010	0.009	0.009	96.67
Current slope	α_{22}	-0.14	-0.143	0.061	0.056	94.31	-0.141	0.061	0.057	94.67
Weibull	$\sqrt{\kappa_2}$	1.3	1.302	0.047	0.045	94.31	1.303	0.046	0.046	95.67
	ζ_{02}	-4	-4.107	1.003	1.021	96.66	-4.098	1.005	1.052	96.33

SE: Standard Error; Coverage rate: coverage rate of the 95% confidence interval.
* Results for 299 replicates with complete convergence over 300 for step 1 and 300 replicates over 300 for step 2.

Table III.8 – Simulation results for scenario D with 500 subjects (13 measures, b_i and τ_i correlated).*

Parameter			Step 1				Step 2			
		True value	Mean	Empirical SE	Mean asymptotic SE	Coverage rate (%)	Mean	Empirical SE	Mean asymptotic SE	Coverage rate (%)
Longitudinal submodel										
Intercept	β_0	142	141.9	0.825	0.746	91.33	141.9	0.812	0.756	91.64
Slope	β_1	3	2.980	0.327	0.282	89.67	2.993	0.311	0.295	93.65
Variability	μ_0	2.4	2.400	0.034	0.033	93.67	2.398	0.034	0.033	94.65
	μ_1	0.05	0.049	0.016	0.015	93.33	0.049	0.016	0.016	93.98
Survival submodel 1										
Current variance	$\alpha_{\sigma 1}$	0.07	0.069	0.050	0.046	94.33	0.070	0.049	0.046	94.65
Current value	α_{11}	0.02	0.021	0.011	0.011	92.67	0.021	0.011	0.011	93.31
Current slope	α_{21}	0.01	0.013	0.086	0.076	94.00	0.011	0.082	0.076	95.32
Weibull	$\sqrt{\kappa_1}$	1.1	1.102	0.054	0.054	95.0	1.102	0.054	0.053	94.65
	ζ_{01}	-7	-7.170	1.389	1.289	93.33	-7.178	1.379	1.292	93.98
Survival submodel 2										
Current variance	$\alpha_{\sigma 2}$	0.15	0.184	0.196	0.057	93.0	0.176	0.167	0.063	95.99
Current value	α_{12}	-0.01	-0.017	0.049	0.015	89.67	-0.015	0.041	0.016	93.31
Current slope	α_{22}	-0.14	-0.170	0.238	0.093	96.67	-0.167	0.330	0.104	97.32
Weibull	$\sqrt{\kappa_2}$	1.3	1.326	0.157	0.072	93.00	1.324	0.164	0.075	93.98
	ζ_{02}	-4	-3.697	3.413	1.643	91.00	-3.817	2.702	1.718	92.46

SE: Standard Error; Coverage rate: coverage rate of the 95% confidence interval.
* Results for 300 replicates with complete convergence over 300.

Table III.9 – Simulation results for scenario D with 1000 subjects (13 measures, b_i and τ_i correlated).*

Parameter			Step 1				Step 2			
		True value	Mean	Empirical SE	Mean asymptotic SE	Coverage rate (%)	Mean	Empirical SE	Mean asymptotic SE	Coverage rate (%)
Longitudinal submodel										
Intercept	β_0	142	141.9	0.575	0.532	91.30	141.9	0.557	0.540	93.67
Slope	β_1	3	3.003	0.209	0.202	93.31	3.007	0.197	0.209	95.67
Variability	μ_0	2.4	2.405	0.025	0.023	93.65	2.402	0.025	0.023	94.33
	μ_1	0.05	0.049	0.012	0.011	93.31	0.049	0.011	0.011	94.33
Survival submodel 1										
Current variance	$\alpha_{\sigma 1}$	0.07	0.063	0.035	0.031	91.30	0.065	0.034	0.031	94.00
Current value	α_{11}	0.02	0.021	0.008	0.007	93.65	0.021	0.008	0.007	94.00
Current slope	α_{21}	0.01	0.014	0.053	0.051	94.31	0.015	0.051	0.050	94.67
Weibull	$\sqrt{\kappa_1}$	1.1	1.099	0.035	0.037	95.32	1.099	0.034	0.037	96.33
	ζ_{01}	-7	-7.087	0.850	0.860	96.66	-7.087	0.853	0.867	96.33
Survival submodel 2										
Current variance	$\alpha_{\sigma 2}$	0.15	0.159	0.042	0.035	92.64	0.159	0.040	0.035	95.00
Current value	α_{12}	-0.01	-0.012	0.011	0.009	94.65	-0.012	0.010	0.009	94.67
Current slope	α_{22}	-0.14	-0.151	0.074	0.060	92.98	-0.150	0.063	0.058	95.67
Weibull	$\sqrt{\kappa_2}$	1.3	1.313	0.046	0.046	94.98	1.313	0.047	0.046	95.00
	ζ_{02}	-4	-3.938	1.160	1.041	94.31	-3.937	1.156	1.054	94.67

SE: Standard Error; Coverage rate: coverage rate of the 95% confidence interval.
* Results for 300 replicates with complete convergence over 300.

Table III.10 – Simulation results for scenario E with 500 subjects (Misspecified model).*

Parameter			Step 1				Step 2			
		True Value	Mean	Empirical SE	Mean asymptotic SE	Coverage rate (%)	Mean	Empirical SE	Mean asymptotic SE	Coverage rate (%)
Longitudinal submodel										
Intercept	β_0	142	140.8	0.808	0.667	52.67	140.8	0.791	0.701	56.67
Slope	β_1	0.7	3.129	0.208	0.190	0	3.135	0.202	0.192	0
Variability	μ_0	2.4	2.406	0.024	0.023	94.00	2.403	0.023	0.024	95.67
	μ_1	0.05	0.052	0.011	0.011	92.00	0.052	0.011	0.011	93.00
Survival submodel 1										
Current variance	α_σ	0	-0.001	0.023	0.022	96.00	0.001	0.022	0.022	95.00
Current value	α_1	0.03	0.030	0.005	0.005	95.67	0.030	0.005	0.005	95.67
Weibull	$\sqrt{\kappa}$	1.1	1.108	0.041	0.040	94.00	1.108	0.041	0.040	94.00
	ζ_0	-7	-6.965	0.788	0.789	94.33	-6.966	0.786	0.787	94.33

SE: Standard Error; Coverage rate: coverage rate of the 95% confidence interval.
* Results for 300 replicates with complete convergence over 300.

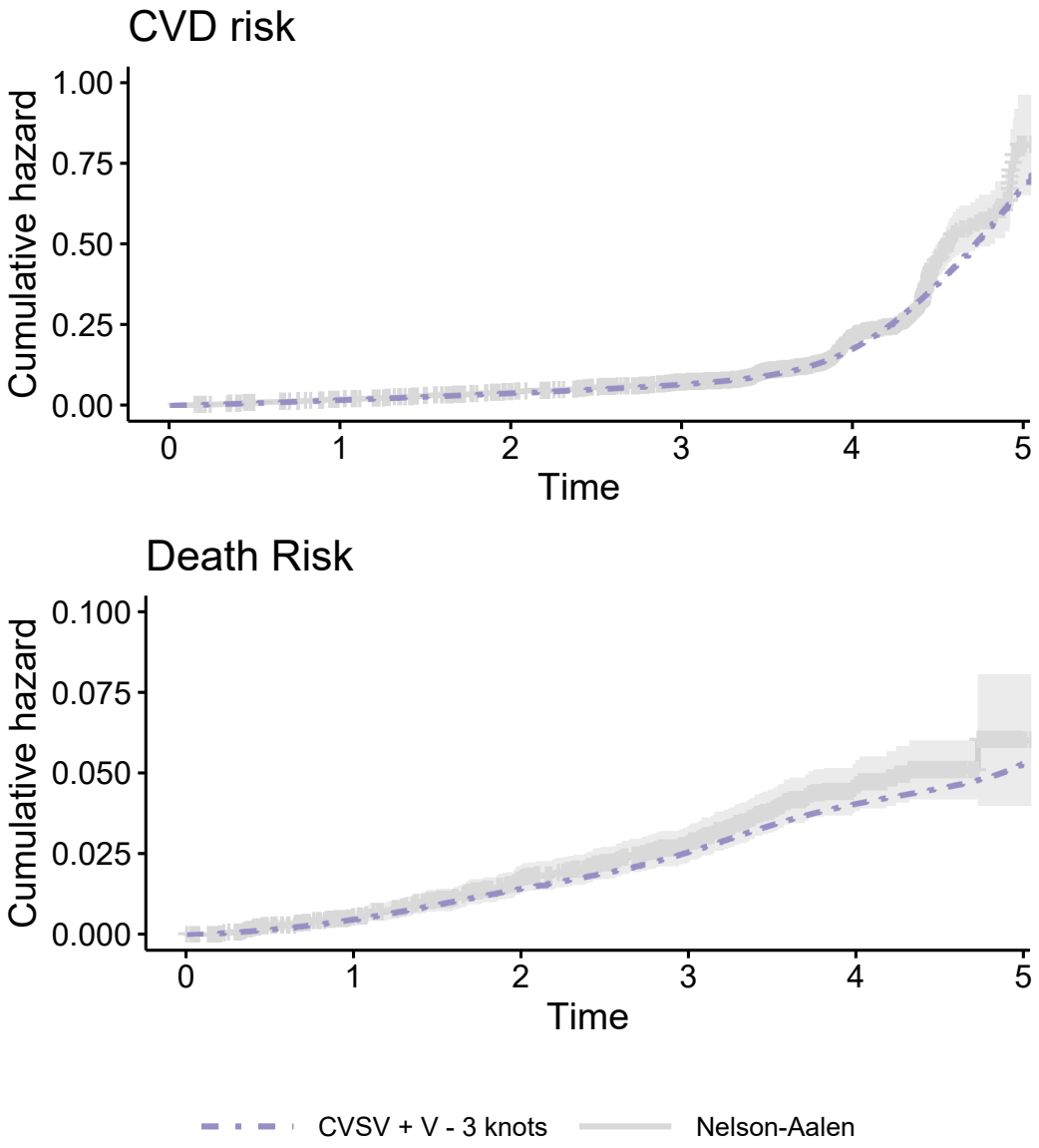

Figure III.4 – Survival submodel for CVD (top) and death (bottom) fit assessment: comparison between predicted cumulative hazard function (in purple) and Nelson Aalen estimator (in grey).

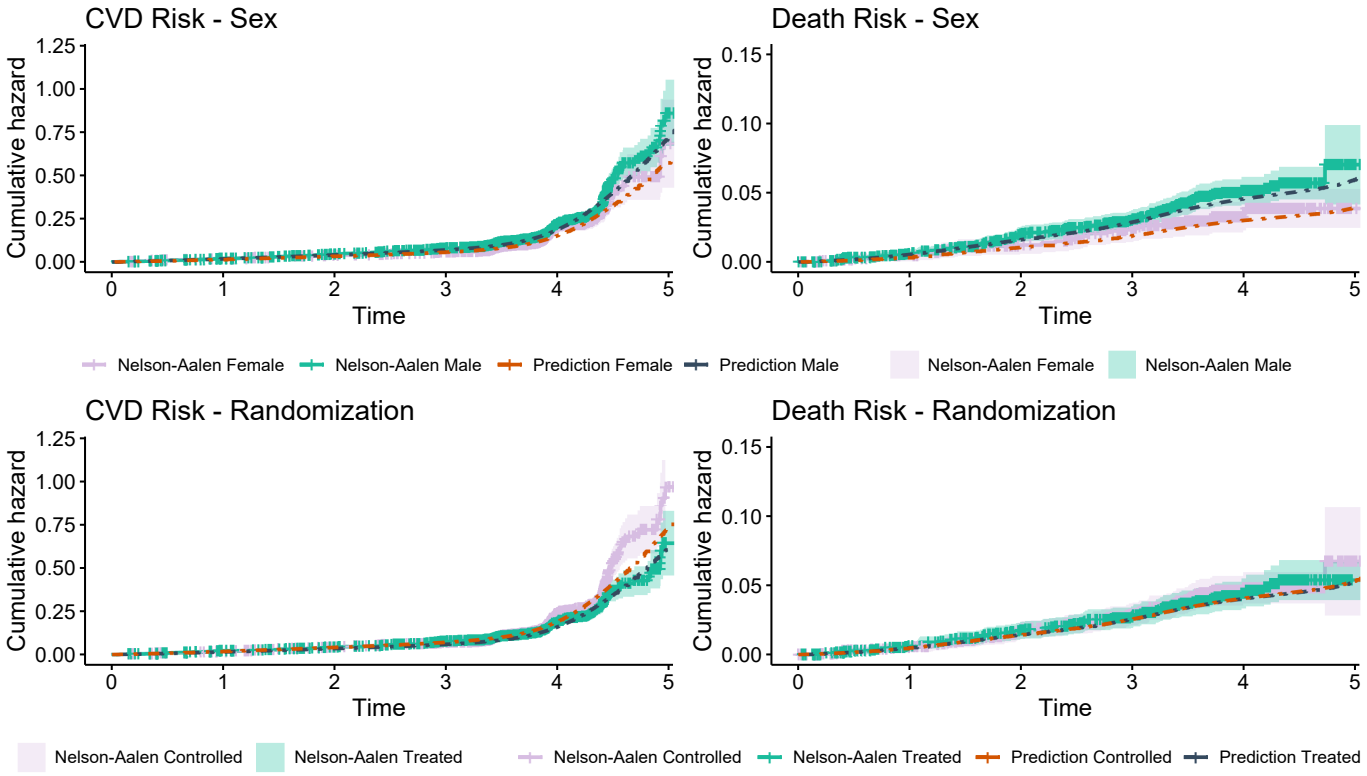

Figure III.5 – Survival submodel for CVD (left) and death (right) fit assessment for Sex (top) and Randomization group (bottom): comparison between predicted cumulative hazard function and Nelson Aalen estimator.

Table III.11 – Parameter estimates of the CVCS joint model on the Progress clinical trial data.

Parameter	**Estimate**	**Standard error**	**p-value**
Survival submodel for CVD			
BP current value	0.003	0.003	0.318
BP current slope	-0.122	0.017	< 0.001
treatment group	-0.093	0.081	0.253
male	0.277	0.087	0.002
age	0.041	0.005	< 0.001
Survival submodel for Death			
BP current value	-0.005	0.007	0.487
BP current slope	-0.100	0.045	0.038
treatment group	-0.008	0.067	0.907
male	0.493	0.194	0.011
age	0.057	0.010	< 0.001
Longitudinal submodel			
Blood Pressure Mean			
intercept	142.7	0.345	< 0.001
time	-0.146	0.073	0.046
treatment group	-7.942	0.457	< 0.001

BP: Blood Pressure

Chapitre IV

Modèle conjoint avec variances résiduelles inter-visites et intra-visite pour données censurées par intervalle

Sommaire

RÉSUMÉ

Ce chapitre propose un nouveau modèle conjoint *location-scale* qui permet, d'une part de distinguer les variabilités résiduelles intra et inter-visites, et d'autre part, de traiter les risques semi-compétitifs censurés par intervalle. En effet, dans le cas de l'étude de la démence et du décès, la date de survenue de la démence est inconnue. Cela peut induire des biais si cette censure n'est pas bien prise en compte (section II.2.6). De plus, la PAS étant mesurée deux ou trois fois à chaque visite, il paraît pertinent de distinguer la variabilité inter-visites, représentant la variabilité à long terme, et la variabilité intra-visite, reflétant la variabilité à court terme. Ce travail a mené à la rédaction d'un article scientifique soumis dans une revue internationale. Ce chapitre correspond à la version courante de l'article qui est disponible librement en tant que pre-print (Courcoul et al., 2024a) [1]. Après un résumé en français, la suite de ce chapitre sera rédigée en anglais.

Le modèle conjoint proposé dans ce chapitre combine un modèle *location-scale* linéaire à effets mixtes, qui traite la variance résiduelle en distinguant la variabilité à long terme de celle à court terme, et un modèle *illness-death* permettant de gérer des risques semi-compétitifs et la censure par intervalle. Les deux variabilités résiduelles sont définies via l'inclusion d'un effet aléatoire individuel pour chacune d'entre elles. Les intensités de transitions entre les différents états de santé peuvent dépendre simultanément de la valeur courante et de la pente du marqueur, ainsi que de chacune des deux composantes de la variance résiduelle. L'estimation du modèle est réalisée en maximisant la fonction de vraisemblance à l'aide de l'algorithme de Marquardt-Levenberg. L'intégrale sur les effets aléatoires est approchée par la méthode de QMC et les intégrales sur le temps pour les différentes transitions et la gestion de la censure par intervalle sont approchées par Gauss-Kronrod. Une étude de simulation valide la procédure et la compare à une méthode naïve du traitement de la censure par intervalle. Cette méthode naïve consiste à imputer le temps de survenue de la démence par le milieu de l'intervalle entre la dernière visite à laquelle l'individu est vu sans démence et la visite de diagnostic, et de considérer comme décédés sans démence les individus décédés sans diagnostic de démence. Cette comparaison met en évidence des résultats biaisés pour la méthode naïve, contrairement au modèle que l'on propose. Le modèle est finalement appliqué à la cohorte des Trois-Citées pour étudier l'impact de la variabilité de la PAS sur le risque de démence et de décès. Cette application montre un effet significatif de la variabilité inter-visites sur le risque de démence et de décès sans démence. En revanche, aucun lien n'est trouvé entre la variabilité intra-visite et les événements considérés.

1. *Joint model for interval-censored semi-competing events and longitudinal data with subject-specific within and between visits variabilities*, L. Courcoul, C. Helmer, A. Barbieri et H. Jacqmin-Gadda. arXiv :2408.06769.

IV.1 Introduction

Dementia affects over 50 million people worldwide, and the number of cases continues to rise with increasing life expectancy. With very few treatment's options, prevention of modifiable vascular risk factors remains a crucial action. Numerous studies have already highlighted the link between hypertension and the risk of dementia. Furthermore, an increasing number of studies are focusing on blood pressure variability as a risk factor, independently of blood pressure level (Alpérovitch et al., 2014; Ma et al., 2019, 2020). These studies predominantly examine long-term blood pressure variability, calculated from repeated measures over several years. In comparison, few studies have investigated short or medium-term blood pressure variability (Ma et al., 2020). Moreover, most studies fail to rigorously account for these different types of variability and suffer from methodological weaknesses (de Courson et al., 2021). Typically, variability is calculated as the standard deviation of blood pressure measurements and included as a time-dependent risk factor in a Cox model. This approach introduces bias by neglecting measurement error in the standard deviation of blood pressure and requiring imputation of the standard deviation for all event times. Additionally, blood pressure and its standard deviation are endogenous variables for which the Cox model is not suitable (Prentice, 1982).

To address these biases, a joint model combining a location-scale linear mixed model with a subject-specific residual variance and a proportional hazards model has been proposed (Gao et al., 2011; Barrett et al., 2019; Courcoul et al., 2024b). However, this model only accounts for a single blood pressure measurement at each measurement time, whereas in most studies, blood pressure is measured at least twice at each visit. From a clinical perspective, it would be interesting to assess whether short-term (intra-visit) variability predicts dementia, as its measurement is straightforward.

Furthermore, given the common risk factors between dementia and death, it is important to consider the competing risk of death in such studies. However, in cohort studies, the onset age of dementia is not precisely known since dementia is only diagnosed at scheduled visit times. This introduces uncertainty about the exact time of dementia onset, which occurs between the last visit where the individual was seen without symptoms and the visit of diagnosis. Moreover, an individual may develop dementia and die between two visits, and thus dementia may not have been diagnosed. Consequently, when interval censoring is not rigorously accounted for, the risk of dementia may be underestimated (Leffondré et al., 2013). To address this issue, Joly et al. (2002) proposed an illness-death model to adjust both the risk of dementia and death, taking into account interval censoring. Later, Rouanet et al. (2016) extended this work to consider longitudinal data via a joint latent class model. However, this

latter work does not allow to consider individual-specific variability and to assess its impact on event risk.

The aim of this work is to propose a joint model combining a location-scale mixed model decomposing individual short- and long-term residual variance and an illness-death model dealing with both interval censoring and left truncation. In this model, the risks of each event can depend simultaneously on the current value, the slope and both variabilities of the marker.

The paper is organized as follows. Section 2 describes the model and the estimation procedure that is then assessed in a simulation study in Section 3. In section 4, the model is applied to the data from the Three-City (3C) cohort (3C Study Group, 2003) to study the impact of within and between visits blood pressure variabilities on the risk of dementia and death. Finally, Section 5 concludes this work with some elements of discussion.

IV.2 Method

Let us consider a sample of N individuals. For each subject i ($i = 1, \ldots, N$), Y_{ijl} is the marker value for measure l ($l = 1, \ldots, n_{ij}$), at visit j ($j = 1, \ldots, n_i$) and time t_{ij}. For each visit j, subject i can have n_{ij} measurements of the longitudinal marker, with $n_{ij} \geq 1$. Y_i is a vector of dimension $\sum_{j=1}^{n_i} n_{ij}$ containing all marker measurements for the subject i.

We denote T_i^{Dem} the unobserved age at dementia onset and T_i^{Death} the age of death. We assume that age at dementia onset is interval-censored while age at death is known since in most cohorts the exact age at death can be collected. The vector of collected data for time-to-events is given by $D_i = (T_{0i}, L_i, R_i, \delta_i^{Dem}, T_i, \delta_i^{Death})^\top$ where T_{0i} is the age at inclusion, L_i is the age at the last visit where the subject was seen free of dementia, R_i is the age at the visit of diagnosis if the subject was diagnosed with dementia (in case of no diagnosis of dementia, R_i is undefined), T_i is the minimum between the age at death and the age of right censoring (e.g. the end of the follow-up), δ_i^{Dem} is the indicator of dementia diagnosis and δ_i^{Death} is the indicator of death.

IV.2.1 Joint model with inter and intra-visit individual variabilities

We propose a joint modelling for a longitudinal outcome and semi-competing events using a shared random-effect approach. The longitudinal submodel is defined by a location-scale

linear mixed-effect model decomposing within and between individual residual variance:

$$\begin{cases} Y_{ijl} = \widetilde{Y}_i(t_{ij}) + \epsilon_{ij} + \nu_{ijl} = X_{ij}^\top \beta + Z_{ij}^\top b_i + \epsilon_{ij} + \nu_{ijl}, \\ \epsilon_{ij} \sim \mathcal{N}(0, \sigma_i^2) \quad \text{with} \quad \log(\sigma_i) = \mu_\sigma + \tau_{\sigma i}, \\ \nu_{ijl} \sim \mathcal{N}(0, \kappa_i^2) \quad \text{with} \quad \log(\kappa_i) = \mu_\kappa + \tau_{\kappa i}, \end{cases} \tag{IV.1}$$

with X_{ij} and Z_{ij} two vectors of explanatory variables for subject i at visit j, respectively associated with the fixed-effect vector β and the subject-specific random-effect vector b_i, and μ_σ and μ_κ two fixed effects associated with the intercept for the between visits and within visit variabilities respectively. The subject-specific random-effect b_i and $\tau_i = (\tau_{\sigma i}, \tau_{\kappa i})^\top$ are both normally distributed and could be supposed to be correlated, such as

$$\begin{pmatrix} b_i \\ \tau_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_b & \Sigma_{\tau b} \\ \Sigma_{\tau b}^\top & \Sigma_\tau \end{pmatrix} \right) \tag{IV.2}$$

To be able to account for the interval censoring of dementia, the risks of dementia and death are modeled according to an illness-death model (Figure IV.1A). The transition intensities from state $k \in \{0, 1\}$ to state $l \in \{1, 2\}$ are defined by a proportional hazards model under the Markovian hypothesis:

$$\lambda_i^{kl}(t|b_i, \tau_i) = \lambda_0^{kl}(t) \exp\left(W_i^{kl\top} \gamma^{kl} + \alpha_1^{kl} \tilde{y}_i(t) + \alpha_2^{kl} \tilde{y}_i'(t) + \alpha_\sigma^{kl} \sigma_i + \alpha_\kappa^{kl} \kappa_i \right) \tag{IV.3}$$

with $\lambda_0^{kl}(t)$ the baseline risk function, W_i^{kl} a vector of baseline covariates associated with the regression coefficient γ^{kl}, and α_1^{kl}, α_2^{kl}, α_σ^{kl}, and α_κ^{kl} the regression coefficients associated with the current value $\tilde{y}_i(t)$, the current slope $\frac{\partial \widetilde{y}_i(t)}{\partial t} = \tilde{y}_i'(t)$, the long-term residual variability σ_i and the short-term residual variability κ_i, respectively. Different parametric forms for the baseline risk functions can be considered, such as exponential, Weibull, or for more flexibility, a B-splines base.

IV.2.2 Estimation Procedure

Log-likelihood

Let θ be the set of parameters to be estimated including parameters of the Cholesky decomposition of the covariance matrix of the random effects, β, $\mu = (\mu_\sigma, \mu_\kappa)^\top$, $\alpha = (\alpha_1^{kl}, \alpha_2^{kl}, \alpha_\sigma^{kl}, \alpha_\kappa^{kl})^\top$, $\gamma = (\gamma^{kl})^\top = (\gamma_{01}, \gamma_{02}, \gamma_{12})^\top$ for $k \in \{0, 1\}$ and $l \in \{1, 2\}$ and the parameters of the three baseline risk functions. Considering the frequentist approach, the parameter estimation is obtained by maximizing the likelihood function. Under the assumption of in-

dependence between D_i and Y_i conditionally on random effects and assuming independent censoring, the contribution of individual i to the marginal likelihood is defined by:

$$\mathcal{L}_i(\theta; Y_i, D_i) = \int f(Y_i|b_i, \tau_i; \theta) f(D_i|b_i, \tau_i; \theta) f(b_i, \tau_i; \theta) db_i d\tau_i, \quad \text{(IV.4)}$$

and in case of delayed entry, the log-likelihood is divided by the probability of being alive and healthy at entry:

$$\mathcal{L}_i^{DE}(\theta; Y_i, D_i) = \frac{\mathcal{L}_i(\theta; Y_i, D_i)}{\int \exp(-\Lambda_{01i}(T_{0i}|b_i, \tau_i; \theta) - \Lambda_{02i}(T_{0i}|b_i, \tau_i; \theta)) f(b_i, \tau_i; \theta) db_i d\tau_i}$$

with:

— $f(b_i, \tau_i; \theta)$ is a multivariate Gaussian density

— $f(Y_i|b_i, \tau_i; \theta) = \prod_{j=1}^{n_i} f(Y_{ij}|b_i, \tau_{\sigma i}, \tau_{\kappa i}; \theta)$ where $f(Y_{ij}|b_i, \tau_i; \theta)$ is a multivariate Gaussian density with symmetric covariance matrix $\Sigma = \sigma_i^2 \mathbb{1}_{n_i}^\top \mathbb{1}_{n_i} + \kappa_i^2 I_{n_i}$, with $\mathbb{1}_{n_i}$ the unity vector of size n_i and I_{n_i} the identity matrice of size $n_i \times n_i$.

— $\Lambda_i^{01}(T_{0i}|b_i, \tau_i; \theta)$ and $\Lambda_i^{02}(T_{0i}|b_i, \tau_i; \theta)$ are the cumulative risk functions respectively for transition Healthy to Dementia and Healthy to Death with for $k = 1, 2$
$\Lambda_i^{0k}(T_{0i}|b_i, \tau_i; \theta) = \int_0^{T_{0i}} \lambda_i^{0k}(t|b_i, \tau_i; \theta) dt$ and $\lambda_i^{0k}(t|b_i, \tau_i; \theta)$ defined in equation (IV.3)

— $f(D_i|b_i, \tau_i; \theta)$ is the survival part of the individual contribution to the likelihood that depends on the subject trajectory as illustrated on Figure IV.1B.

To present the different definitions of $f(D_i|b_i, \tau_i; \theta)$ according to subjects observations, we deliberately omit the conditioning on random effects and parameters for ease of notation:

Cases 1 and 2: subject healthy and alive until L_i, diagnosed with dementia at R_i, remained alive until T_i, possibly died at T_i:

$$f(T_{0i}, L_i, R_i, \delta_i^{Dem} = 1, T_i, \delta_i^{Death}) = \int_{L_i}^{R_i} e^{-\Lambda_i^{01}(u) - \Lambda_i^{02}(u)} \lambda_i^{01}(u) e^{-(\Lambda_i^{12}(T_i) - \Lambda_i^{12}(u))} \lambda_i^{12}(T_i)^{\delta_i^{Death}} du$$

Cases 3 and 4: subject healthy and alive until T_i ($T_i = L_i$), possibly died at T_i:

$$f(T_{0i}, L_i, R_i, \delta_i^{Dem} = 0, T_i, \delta_i^{Death}) = e^{-\Lambda_i^{01}(T_i) - \Lambda_i^{02}(T_i)} \lambda_i^{02}(T_i)^{\delta_i^{Death}}$$

Cases 5 and 6: subject healthy at his/her last visit $L_i = R_i < T_i$, remained alive until T_i

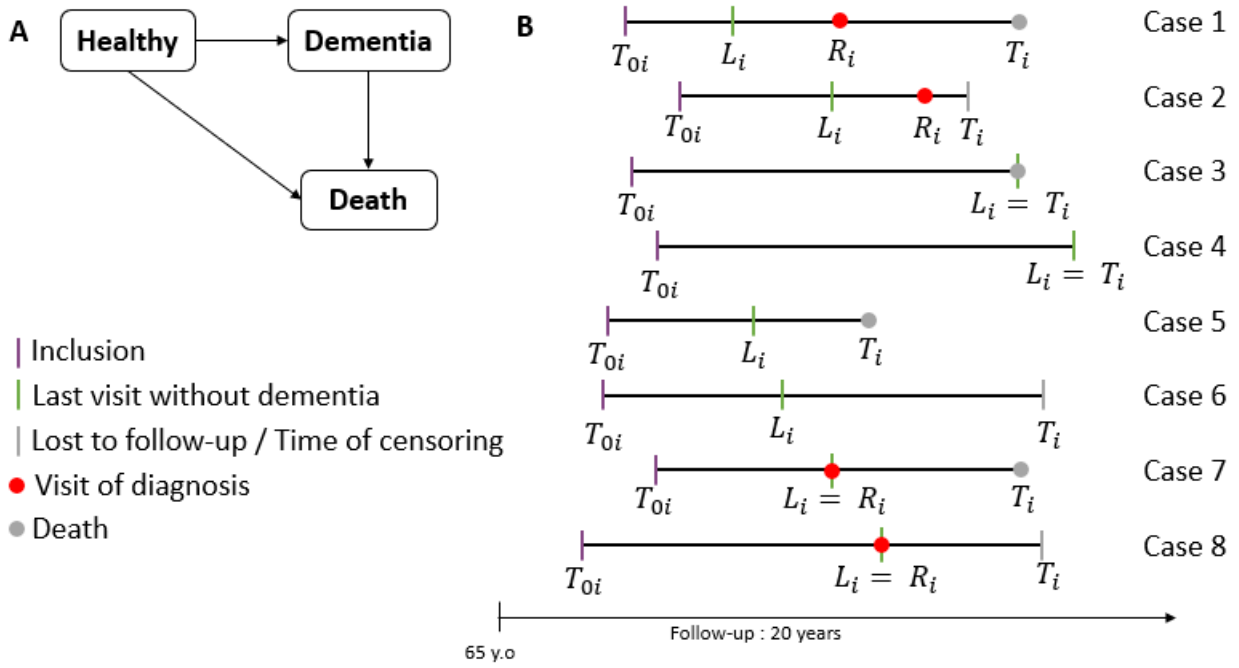

Figure IV.1 – A. Illness-death model. B. Possible patterns of dementia and death.

and possibly died at T_i:

$$f(T_{0i}, L_i, R_i, \delta_i^{Dem} = 0, T_i, \delta_i^{Death}) = e^{-\Lambda_i^{01}(T_i)-\Lambda_i^{02}(T_i)}\lambda_i^{02}(T_i)^{\delta_i^{Death}} + \int_{L_i}^{T_i} e^{-\Lambda_i^{01}(u)-\Lambda_i^{02}(u)}\lambda_i^{01}(u)e^{-(\Lambda_i^{12}(T_i)-\Lambda_i^{12}(u))}\lambda_i^{12}(T_i)^{\delta_i^{Death}}du$$

The likelihood accounts for the two possible trajectories of this subject: either direct transition from healthy to death or unobserved transition to dementia between L_i and T_i.

Cases 7 and 8: subject diagnosed with dementia at $L_i = R_i$ (exact date of dementia onset), remained alive until T_i, possibly died at T_i:

$$f(T_{0i}, L_i, R_i, \delta_i^{Dem} = 1, T_i, \delta_i^{Death}) = e^{-\Lambda_i^{01}(L_i)-\Lambda_i^{02}(L_i)}\lambda_i^{01}(L_i)e^{-(\Lambda_i^{12}(T_i)-\Lambda_i^{12}(L_i))}\lambda_i^{12}(T_i)^{\delta_i^{Death}}$$

These cases do not occur in the context of dementia but could arise in other types of events.

Optimisation

Since the integral over the random effects does not have an analyticial solution, it is computed using a Quasi Monte Carlo (QMC) approximation (Pan and Thompson, 2007), employing

deterministic quasi-random sequences. This leads to the equation (IV.4) being approximated by:

$$\mathcal{L}_i^{DE}(\theta; Y_i, D_i) \simeq \frac{\frac{1}{S}\sum_{s=1}^{S} p(Y_i, D_i | b_i^s, \tau_i^s; \theta)}{\frac{1}{S}\sum_{s=1}^{S} \exp(-\Lambda_i^{01}(T_{0i}|b_i^s, \tau_i^s; \theta) - \Lambda_i^{02}(T_{0i}|b_i^s, \tau_i^s; \theta))}$$

where $(b_i^1, ..., b_i^S)$ and $(\tau_{\sigma i}^1, ..., \tau_{\sigma i}^S, \tau_{\kappa i}^1, ..., \tau_{\kappa i}^S)$ are draws of a S-sample in the sobol sequel for the distribution $f(b_i, \tau_{\sigma i}, \tau_{\kappa i}; \theta)$. Moreover, to approximate the cumulative risk functions, we use the Gauss-Kronrod quadrature approximation with 15 points (Gonnet, 2012).

Parameter estimation is obtained by maximizing the log-likelihood function $\ell(\theta; Y, D) = \log\left(\prod_{i=1}^{N} \mathcal{L}_i(\theta; Y_i, D_i)\right)$. The maximization is performed using the `marqLevAlg` R-package based on the Marquardt-Levenberg algorithm (Philipps et al., 2021).

The variances of the estimates are estimated by the inverse of the Hessian matrix computed by finite differences. The variances of the parameters of the random effects covariance matrix are obtained from those of the parameters of the Cholesky decomposition using the Delta-Method (Meyer et al., 2013).

In order to insure precise estimates of parameters and their standard error and to limit computation time, the estimation is performed in three steps:

1. **Parameters initialisation**. estimation of the joint model without handling interval censoring:
   - taking the middle of the interval as time of dementia onset and assuming that subjects who died before diagnosis of dementia directly made the transition health to death;
   - using the estimated parameters of the estimation from the longitudinal model (equation (IV.1))
2. **Parameters estimation** of the joint model accounting for interval censoring with a sufficient number $S1$ of QMC to avoid bias using the estimated parameters from step (1);
3. **Precision improvement** with a number $S2 > S1$ of QMC draws using the estimated parameters from step (2).

The choice of the number of QMC draws $S1$ and $S2$ has a notable effect on computational time. Therefore, for model selection, we suggest that users compare different models using the results of step (2) (using likelihood or information criteria) with a small value for $S1$ and

perform step (3), which entails a larger number of QMC draws, solely for the final selected model.

IV.3 Simulations

In order to evaluate the performance of the estimation procedure and compare the estimations with those obtained using a joint model that does not account for interval censoring we carried out simulations driven by the application.

IV.3.1 Design

For each subject, age at entry is generated from a beta distribution, over a window of 65 to 85 years old. Visit times are generated using a uniform distribution centered around each specified time, with a variation of ± 2.4 months in either direction. For each visit time, two measurements ($l = 1, 2$) of the marker are generated, using a location-scale linear mixed-effects model with fixed and random intercept and slope, and two heterogeneous variances:

$$\begin{cases} Y_{ijl} = \widetilde{Y}_i(t_{ij}) + \epsilon_{ij} + \nu_{ijl} = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) \times t_{ij} + \epsilon_{ij} + \nu_{ijl} \\ \epsilon_{ij} \sim \mathcal{N}(0, \sigma_i^2) \quad \text{with} \quad \log(\sigma_i) = \mu_\sigma + \tau_{\sigma i}, \\ \nu_{ijl} \sim \mathcal{N}(0, \kappa_i^2) \quad \text{with} \quad \log(\kappa_i) = \mu_\kappa + \tau_{\kappa i}, \end{cases} \tag{IV.5}$$

where t is a scaled transformation of age $t : t = \frac{age-65}{10}$.

The follow-up visits, $(t_{ij})_{j=(1,\dots,n_i)}$, are scheduled regularly from inclusion until the minimum between death and the administrative right-censoring which is 20 years after inclusion. The process of age of dementia onset and death generation is described on Supplementary Materials (IV.6.1) with each function of transition defined by the following proportional hazards models:

$$\lambda_i^{kl}(t) = \lambda_0^{kl}(t) \exp\left(\alpha_1^{kl}\tilde{y}_i(t) + \alpha_2^{kl}\tilde{y}_i'(t) + \alpha_\sigma^{kl}\sigma_i + \alpha_\kappa^{kl}\kappa_i\right) \tag{IV.6}$$

with $\lambda_0^{kl}(t) = \eta^{kl} t^{\eta^{kl}-1} e^{\alpha_0^{kl}}$ being a Weibull function.

Three scenarii were performed, varying the parameters and the number of repeated measures:

— Scenario A: parameters driven by the estimation on the model defined by (IV.5) and (IV.6) on the 3C cohort, and visit times driven by 3C cohorts: at 2, 4, 7, 10, 12, 14 and 17 years from inclusion.

— Scenario B: same parameters as in Scenario A but visit times at 4, 8, 12 and 16 years from inclusion to assess the impact of the interval between visits on the results.

— Scenario C: increased signal on interest parameters to simulate a positive effect of the current value and between-visits variability on the risk of dementia, and same visit times as in Scenario A.

For each scenario, 500 samples of 1000 subjects have been generated. We also estimated a naive joint illness-death model without handling interval censoring considering the exact time of dementia onset known as the middle of the interval between the last visit without symptoms and the visit of diagnosis for diagnosed individuals and considering that subjects who died before diagnosis of dementia directly made the transition healthy to death. All estimations are performed with $S1 = 1000$ QMC draws and $S2 = 5000$ QMC draws.

IV.3.2 Results

Tables IV.1, IV.2 and IV.3 displays the results of the simulation studies for each scenario. For each scenario, we compared the results of our model properly handling interval censoring into account with those obtained from the model without taking interval censoring into account. For the three scenarios, the estimation procedure for the model accounting for the interval censoring provided satisfactory results as the bias was minimal, the mean asymptotic and the empirical standard errors were close, and the coverage rates of the 95% confidence interval were close to the nominal value. When comparing these results with those of the model that does not account for interval censoring, we observe higher biases and underestimation of the standard errors, leading to poor coverage rates. These trends are more pronounced for Scenario B and Scenario C as the interval between visits is larger or the signal increase. Especially for scenario C, we observe large biases for the association parameters between inter-visit variability and event risks. As shown by Leffondré et al. (2013), the bias is stronger for scenario C because the inter-visit variability is strongly associated with both dementia and death after dementia. Thus, subjects with high variability are more prone to die before the first visit following dementia onset and consequently to be undiagnosed. They are thus considered as died without dementia in the naive analysis.

Table IV.1 – Results of the simulation study for Scenario A, comparing estimates of the joint illness-death model for interval-censored events and the naive joint illness-death model. A total of 500 samples of 1000 subjects were generated with a joint illness-death model with visit times driven by 3C design. ASE is the asymptotic standard error, ESE is the empirical standard error and the coverage rate is calculated from the 95% confidence interval.

			Interval censoring[a]				**Naive model**[b]			
Parameter		θ	$\hat{\theta}$	ESE	ASE	CR (95%)	$\hat{\theta}$	ESE	ASE	CR (95%)
Longitudinal process										
Intercept	β_0	14.0	14.00	0.10	0.10	95.2	13.98	0.10	0.10	94.2
Slope	β_1	0.17	0.170	0.064	0.063	95.6	0.193	0.062	0.061	92.8
Variability inter	μ_σ	0.30	0.300	0.017	0.017	94.8	0.300	0.017	0.017	94.8
Variability intra	μ_κ	−0.23	−0.231	0.013	0.013	94.0	−0.232	0.013	0.013	94.4
Transition 0-1										
Current value	α_1^{01}	−0.06	−0.059	0.045	0.045	95.0	−0.071	0.045	0.045	95.0
Current slope	α_2^{01}	0.0	−0.002	0.093	0.094	94.6	−0.017	0.093	0.093	95.2
Inter-variability	α_σ^{01}	0.50	0.521	0.320	0.297	94.4	0.569	0.317	0.300	95.4
Intra-variability	α_κ^{01}	0.01	−0.013	0.478	0.453	94.0	−0.005	0.479	0.461	94.4
Weibull	$\sqrt{\eta^{01}}$	2.00	2.012	0.055	0.057	94.8	1.907	0.056	0.058	60.5
	ζ^{01}	−4.00	−4.080	0.764	0.781	94.8	−3.815	0.744	0.776	95.0
Transition 0-2										
Current value	α_1^{02}	−0.10	−0.107	0.051	0.052	95.2	−0.069	0.038	0.039	86.7
Current slope	α_2^{02}	−0.40	−0.411	0.119	0.115	95.8	−0.293	0.082	0.084	72.1
Inter-variability	α_σ^{02}	0.46	0.440	0.368	0.347	95.6	0.412	0.264	0.265	96.0
Intra-variability	α_κ^{02}	0.21	0.181	0.553	0.516	94.6	0.133	0.387	0.394	96.8
Weibull	$\sqrt{\eta^{02}}$	1.70	1.708	0.060	0.060	94.6	1.822	0.051	0.049	32.3
	ζ^{02}	−2.50	−2.398	0.906	0.897	95.0	−3.031	0.691	0.695	89.4
Transition 1-2										
Current value	α_1^{12}	0.04	0.047	0.056	0.056	94.8	0.027	0.046	0.055	97.6
Current slope	α_2^{12}	0.02	0.011	0.120	0.120	94.8	−0.021	0.081	0.107	98.6
Inter-variability	α_σ^{12}	−0.12	−0.127	0.352	0.335	95.0	0.038	0.211	0.278	97.0
Intra-variability	α_κ^{12}	−0.18	−0.158	0.571	0.531	93.8	−0.142	0.372	0.463	98.2
Weibull	$\sqrt{\eta^{12}}$	1.70	1.717	0.104	0.106	94.2	1.788	0.095	0.103	87.2
	ζ^{12}	−2.20	−2.385	1.022	1.005	95.0	−2.929	0.823	0.976	95.2

[a]496 samples with convergence criteria fulfilled; [b]499 samples with convergence criteria fulfilled.

Table IV.2 – Results of the simulation study for Scenario B, comparing estimates of the joint illness-death model for interval-censored events and the naive joint illness-death model. A total of 500 samples of 1000 subjects were generated with a joint illness-death model with visit times every 4 years. ASE is the asymptotic standard error, ESE is the empirical standard error and the coverage rate is calculated from the 95% confidence interval.

			Interval censoring[a]				**Naive model**[b]			
Parameter		θ	$\hat{\theta}$	ESE	ASE	CR (95%)	$\hat{\theta}$	ESE	ASE	CR (95%)
Longitudinal process										
Intercept	β_0	14.0	14.01	0.12	0.12	95.4	13.97	0.12	0.12	94.2
Slope	β_1	0.17	0.162	0.074	0.076	95.8	0.198	0.071	0.071	93.1
Variability inter	μ_σ	0.30	0.301	0.026	0.027	95.8	0.301	0.027	0.026	94.8
Variability intra	μ_κ	-0.23	-0.230	0.016	0.017	95.6	-0.230	0.015	0.016	96.5
Transition 0-1										
Current value	α_1^{01}	-0.06	-0.060	0.056	0.054	94.7	-0.078	0.067	0.057	91.3
Current slope	α_2^{01}	0.0	0.006	0.125	0.124	96.0	-0.093	1.051	0.147	95.2
Inter-variability	α_σ^{01}	0.50	0.559	0.595	0.531	96.6	0.757	2.116	0.601	97.0
Intra-variability	α_κ^{01}	0.01	-0.057	0.780	0.725	95.2	-0.176	2.237	0.763	94.4
Weibull	$\sqrt{\eta^{01}}$	2.00	2.022	0.067	0.068	96.2	1.885	0.172	0.069	43.3
	ζ^{01}	-4.00	-4.129	1.099	1.057	96.2	-3.919	2.116	1.168	93.3
Transition 0-2										
Current value	α_1^{02}	-0.10	-0.107	0.067	0.066	95.2	-0.057	0.041	0.043	83.5
Current slope	α_2^{02}	-0.40	-0.440	0.156	0.159	97.2	-0.269	0.099	0.100	68.6
Inter-variability	α_σ^{02}	0.46	0.421	0.659	0.648	98.2	0.392	0.432	0.398	96.1
Intra-variability	α_κ^{02}	0.21	0.230	0.828	0.843	97.0	0.149	0.510	0.526	96.8
Weibull	$\sqrt{\eta^{02}}$	1.70	1.709	0.063	0.071	96.8	1.855	0.048	0.052	10.8
	ζ^{02}	-2.50	-2.454	1.217	1.264	96.6	-3.237	0.786	0.838	88.7
Transition 1-2										
Current value	α_1^{12}	0.04	0.049	0.074	0.072	95.8	0.011	0.152	0.088	96.8
Current slope	α_2^{12}	0.02	0.010	0.170	0.167	96.0	-0.285	3.514	0.303	97.8
Inter-variability	α_σ^{12}	-0.12	-0.065	0.690	0.578	96.8	0.582	6.825	0.847	97.6
Intra-variability	α_κ^{12}	-0.18	-0.207	0.915	0.837	95.8	-0.565	6.508	1.126	98.3
Weibull	$\sqrt{\eta^{12}}$	1.70	1.754	0.139	0.136	94.5	1.971	0.588	0.150	66.0
	ζ^{12}	-2.20	-2.622	1.528	1.399	93.9	-4.529	8.827	1.865	86.4

[a]495 samples with convergence criteria fulfilled; [b]462 samples with convergence criteria fulfilled.

Table IV.3 – Results of the simulation study for Scenario C, comparing estimates of the joint illness-death model for interval-censored events and the naive joint illness-death model. A total of 500 samples of 1000 subjects were generated with a joint illness-death model with visit times driven by 3C. ASE is the asymptotic standard error, ESE is the empirical standard error and the coverage rate is calculated from the 95% confidence interval.

			Interval censoring[a]				**Naive model**[a]			
Parameter		θ	$\hat{\theta}$	ESE	ASE	CR (95%)	$\hat{\theta}$	ESE	ASE	CR (95%)
Longitudinal process										
Intercept	β_0	14.0	14.00	0.10	0.09	93.4	14.01	0.10	0.10	93.8
Slope	β_1	0.17	0.167	0.062	0.061	94.6	0.155	0.060	0.059	94.2
Variability inter	μ_σ	0.30	0.299	0.021	0.020	95.6	0.301	0.020	0.020	95.2
Variability intra	μ_κ	-0.23	-0.230	0.017	0.016	95.2	-0.231	0.016	0.016	95.0
Transition 0-1										
Current value	α_1^{01}	0.20	0.204	0.046	0.043	93.6	0.151	0.043	0.042	77.5
Current slope	α_2^{01}	0.0	-0.003	0.078	0.075	94.4	-0.047	0.081	0.075	90.2
Inter-variability	α_σ^{01}	0.80	0.819	0.149	0.148	95.0	0.583	0.155	0.145	64.3
Intra-variability	α_κ^{01}	0.01	0.014	0.203	0.190	93.2	-0.048	0.209	0.187	93.0
Weibull	$\sqrt{\eta^{01}}$	2.00	2.006	0.046	0.047	95.0	1.851	0.045	0.047	10.4
	ζ^{01}	-7.00	-7.105	0.773	0.733	94.8	-5.867	0.722	0.704	64.5
Transition 0-2										
Current value	α_1^{02}	0.30	0.302	0.109	0.097	93.2	0.305	0.048	0.048	95.8
Current slope	α_2^{02}	0.10	0.098	0.199	0.184	92.4	0.066	0.079	0.080	92.4
Inter-variability	α_σ^{02}	0.20	0.129	0.424	0.394	94.2	0.811	0.152	0.155	1.6
Intra-variability	α_κ^{02}	0.20	0.148	0.426	0.394	93.6	0.141	0.190	0.193	94.2
Weibull	$\sqrt{\eta^{02}}$	1.70	1.697	0.089	0.086	93.8	2.011	0.056	0.052	0.0
	ζ^{02}	-8.00	-7.959	1.797	1.608	93.2	-9.144	0.839	0.836	75.1
Transition 1-2										
Current value	α_1^{12}	0.15	0.158	0.052	0.051	95.2	0.145	0.038	0.049	99.2
Current slope	α_2^{12}	0.10	0.101	0.085	0.085	95.8	0.025	0.051	0.072	90.4
Inter-variability	α_σ^{12}	0.80	0.813	0.192	0.181	94.0	0.565	0.093	0.140	63.5
Intra-variability	α_κ^{12}	0.10	0.128	0.227	0.220	94.4	0.107	0.146	0.198	99.4
Weibull	$\sqrt{\eta^{12}}$	1.70	1.717	0.095	0.092	94.4	1.725	0.074	0.086	96.2
	ζ^{12}	-4.50	-4.712	1.033	1.007	94.2	-4.576	0.718	0.919	99.4

[a] 498 samples with convergence criteria fulfilled.

IV.4 Application

The proposed model was then applied to the French prospective 3C Cohort to study the effect of the inter- and intra-visit blood pressure variability on the risk of dementia.

IV.4.1 The Three-City Cohort

The 3C study is a population-based prospective cohort which aimed at assessing the relation between vascular factors and dementia in the elderly (3C Study Group, 2003). Participants, aged 65 years and older, were randomly selected in 1999 from the electoral lists of three French cities, Bordeaux, Dijon and Montpellier. This analysis was performed on the Bordeaux subsample because the follow-up was longer. This Bordeaux sample included 2104 participants at baseline who were followed every 2-3 years up to 20 years. At each visit, systolic blood pressure (SBP) was measured two or three times. Each measurement was taken in a seated position after a rest period.
The final sample of analysis included 1788 non-demented participants at baseline for whom there is at least one SBP measurement and sex, educational level and APOE are known. Among these participants 493 developed dementia, 1002 died without previous dementia diagnosis and 401 died after dementia diagnosis. Participants were 74 years old at baseline on average, 61% were women, 62% had an educational level lower than secondary school and 19% carried the APOE allele.

IV.4.2 Specification of the model

The aim of the study was to evaluate the impact of within and between visits blood pressure variability on the risk of transition to dementia and death (before and after dementia). We estimated the proposed joint model defined by (IV.7) using the age as time scale. The mean trajectory of blood pressure was described over time by a linear location-scale mixed effect model. The individual time trend of the marker was modelled by a linear trend. The baseline hazard functions of each transition were defined by a Weibull function. The models for transition intensities for each event depended on both inter and intra-visit variabilities, the individual current value and the current slope of the blood pressure, and they were also adjusted for sex (male versus female), educational level (higher than secondary school ($\geq$ 10

years of study or not) and on carrying the APOE allele:

$$\begin{cases} Y_{ijl} = \widetilde{Y}_i(t_{ij}) + \epsilon_{ij} + \nu_{ijl} = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) \times t_{ij} + \epsilon_{ij} + \nu_{ijl}, \\ \epsilon_{ij} \sim \mathcal{N}(0, \sigma_i^2) \quad \text{with} \quad \log(\sigma_i) = \mu_\sigma + \tau_{\sigma i}, \\ \nu_{ijl} \sim \mathcal{N}(0, \kappa_i^2) \quad \text{with} \quad \log(\kappa_i) = \mu_\kappa + \tau_{\kappa i}, \\ \lambda_i^{kl}(t) = \lambda_0^{kl}(t) \exp\left(\gamma_1^{kl} Sex_i + \gamma_2^{kl} Edu_i + \gamma_3^{kl} Apoe4_i + \alpha_1^{kl} \tilde{y}_i(t) + \alpha_2^{kl} \tilde{y'}_i(t) + \alpha_\sigma^{kl} \sigma_i + \alpha_\kappa^{kl} \kappa_i\right) \end{cases} \tag{IV.7}$$

with $t = \frac{age-65}{10}$ and the random effects b_i and $\tau_i = (\tau_{\sigma i}, \tau_{\kappa i})^\top$ are supposed to be independent. The estimation was performed with $S1 = 1000$ and $S2 = 5000$ draws of QMC to ensure a greater accuracy.

IV.4.3 Results

Table IV.4 provides estimates of the regression parameters from the joint model and equation (IV.8) presents the covariance matrix of the random effects and their standard errors computed through the Delta-Method. Mean blood pressure increased with time ($\widehat{\beta_1} = 0.211$, $p < 0.001$, table IV.4). Both between and within visits variances of the residual part were heterogenerous between the subjects ($\widehat{Var}(\tau_{\sigma i}) = 0.07$, $sd = 0.01$ and $\widehat{Var}(\tau_{\kappa i}) = 0.07$, $sd = 0.006$) which confirms the importance of taking heteroscedasticity into account (Figures S1 on Supplementary Materials IV.6.2). The risk of dementia was lower for subjects with a high educational level (Hazard Ratio: $\widehat{HR} = 0.79$, $\widehat{IC} = [0.66; 0.96]$) but 50% higher for subjects carrying the APOE4 allele ($\widehat{HR} = 1.50$, $\widehat{IC} = [1.21; 1.86]$). There was no effect of the sex ($p = 0.80$). Adjusting for these covariates, the risk of dementia increased with the between-visits blood pressure variability ($\widehat{HR} = 1.82$, $\widehat{IC} = [1.10; 3.00]$) but we found no significant effect of within-visit blood pressure variability, the current slope and the current value. Then, considering both the risk of death after dementia and without dementia, the risk of death was lower for women ($\widehat{HR} = 0.47$, $\widehat{IC} = [0.38; 0.58]$ without dementia and $\widehat{HR} = 0.57$, $\widehat{IC} = [0.47; 0.70]$ with dementia) but none of the other covariates and none of the characteristics of the blood pressure trajectory was significantly associated with the risk of death.

Table IV.4 – Parameter estimates of the joint model on the 3C-Bordeaux data ($N = 1788$).

Parameter	**Estimate**	**Standard error**	**p-value**
Dementia (transition 0-1)			
BP current value	-0.071	0.051	0.163
BP slope	-0.070	0.077	0.364
BPV between-visits	0.597	0.256	0.019
BPV within-visit	-0.084	0.335	0.802
Woman	-0.027	0.109	0.804
Education ($\geq$ 10 years)	-0.230	0.098	0.020
APOE4	0.404	0.110	<0.001
Death without dementia (transition 0-2)			
BP current value	-0.086	0.072	0.235
BP slope	-0.196	0.103	0.058
BPV between-visits	0.396	0.331	0.232
BPV within-visit	0.319	0.411	0.437
Woman	-0.760	0.1097	<0.001
Education ($\geq$ 10 years)	-0.087	0.104	0.404
APOE4	-0.015	0.138	0.915
Death after dementia (transition 1-2)			
BP current value	0.035	0.048	0.464
BP slope	0.009	0.079	0.909
BPV between-visits	-0.231	0.218	0.290
BPV within-visit	-0.154	0.298	0.607
Woman	-0.556	0.101	<0.001
Education ($\geq$ 10 years)	0.081	0.096	0.404
APOE4	-0.067	0.109	0.538
Longitudinal submodel			
Intercept	13.90	0.081	<0.001
Time	0.211	0.053	<0.001
Intercept of between-visits variability	0.299	0.015	<0.001
Intercept of within-visits variability	-0.228	0.010	<0.001

$$
\begin{aligned}
\widehat{\Sigma} &= \begin{bmatrix}
\widehat{Var}(b_{0i}) & & & \\
\widehat{Cov}(b_{0i}, b_{1i}) & \widehat{Var}(b_{1i}) & & \\
\widehat{Cov}(b_{0i}, \tau_{\sigma i}) & \widehat{Cov}(b_{1i}, \tau_{\sigma i}) & \widehat{Var}(\tau_{\sigma i}) & \\
\widehat{Cov}(b_{0i}, \tau_{\kappa i}) & \widehat{Cov}(b_{1i}, \tau_{\kappa i}) & \widehat{Cov}(\tau_{\sigma i}, \tau_{\kappa i}) & \widehat{Var}(\tau_{\kappa i})
\end{bmatrix} \\
&= \begin{bmatrix}
4.57_{(0.50)} & & & \\
-1.86_{(0.61)} & 1.22_{(0.86)} & & \\
0 & 0 & 0.07_{(0.01)} & \\
0 & 0 & 0.01_{(6e-3)} & 0.07_{(6e-3)}
\end{bmatrix}
\end{aligned}
\tag{IV.8}
$$

IV.4.4 Goodness-of-fit

To assess fit of the longitudinal sub-model we computed the empirical Bayes estimates of the random effects and the predicted value of the marker for each individual at their respective visiting times. Figure IV.2 A compares the mean of marker predictions at each visit time to the mean of the observed measurements. It shows that the joint model adequately fit the trajectory of the blood pressure.

Then to evaluate the fit of the illness-death submodel, we plugged the empirical Bayes estimates of the random effects in the formula for the risk functions to compute the predicted cumulative hazard function in a grid of times. Figure IV.2 B compares the mean of this predicted cumulative hazard function for each transition to a non-parametric estimator obtained using the `SmoothHazard` package (Touraine et al., 2017). It highlights that the proposed joint model adequately fitted each transition.

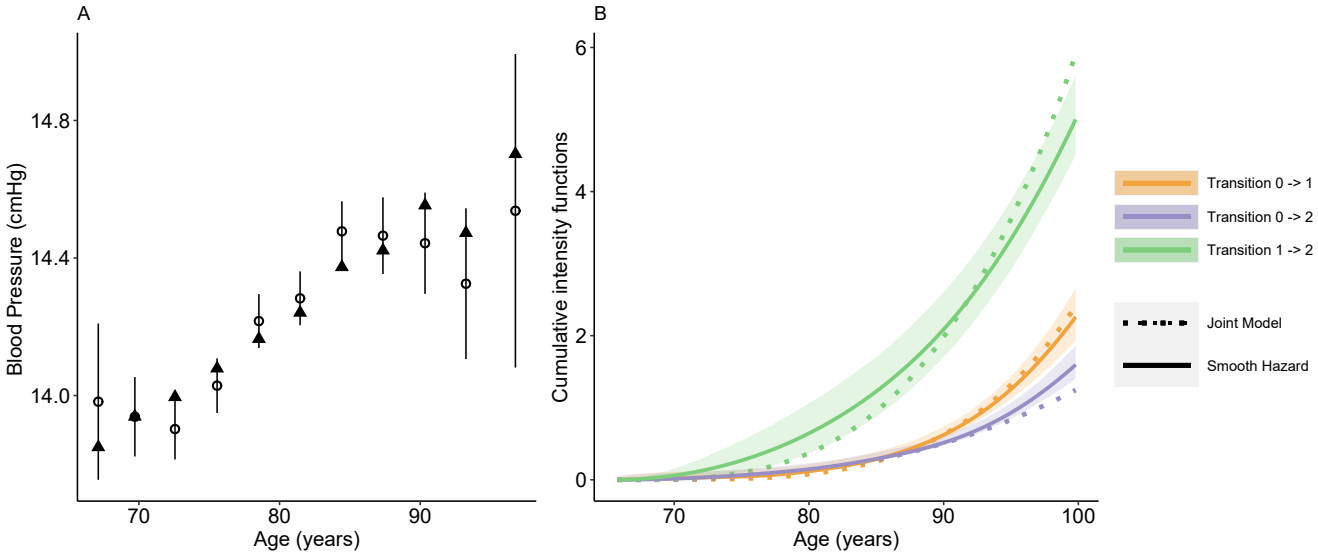

Figure IV.2 – Goodness-of-fit of the model for the 3C-Bordeaux cohort ($N = 1788$): A. Mixed effects submodel: Comparison between predicted value of the marker from the joint model (black triangles) and the observations (mean in white circles with 95% confidence interval (by 3-year age intervals). B. Illness-death model: Comparison between the predicted cumulative hazard function from the joint model for each transition and a non-parametric estimator with 95% confidence interval (Illness-death model estimated by penalized likelihood accounting for interval censoring with the `R`-package `SmoothHazard`).

IV.5 Discussion

We proposed a new joint model for semi-competing interval censored events, and longitudinal data with heteroscedasticity assuming two subject-specific residual variances: one within-visit and one between-visits. This model allows to study the impact of the residual variabilities on the risks of the events. Simulations studies demonstrated the good performance of the estimation procedure dealing with interval censoring and emphasized biased estimates obtained with a naive model estimated by imputing the middle of the censoring interval for the time of dementia onset and assuming that all the subjects who died without dementia diagnosis were really not demented at death. The `R`-package `LSJM` has been developped to allow the estimation of such models and is available on Github at the following link: `https://github.com/LeonieCourcoul/LSJM`.

The analysis of the 3C cohort with this model has shown that a high between-visits blood pressure variability increases the risk of dementia but the within-visit variability seems to have no impact. This suggests that intra-visit variability could mainly quantifies measurement error rather than true variability. This does not support the hypothesis that intra-visit variability, which is easily measured could be an interesting indicator of the dementia risk.

This model relies on some hypothesis. Regarding missing data for marker measurements, it is important to note that this model deals with the informative dropouts due to dementia and death thanks to the joint modeling of the risks of death and dementia. But it assumes that missing marker measurements for other causes are missing at random. Moreover, this model relies on a Markovian assumption but semi-Markovian hypothesis could be considered. However, Rouanet et al. (2016) shown previously that mortality among subjects with dementia depends more on age than on duration of dementia, suggesting that the Markovian hypothesis is better.

Joint models that account for inter-visit and intra-visit subject-specific residual variance are highly valuable for examining the relationship between both variabilities of markers or risk factors and the risk of health events across various areas of medical research. Furthermore, the proposed model can be extended to consider not only inter-visit and intra-visit variability but also long-term versus short-term variability via a temporal window with different measurements times for the short-term variability. It could be useful to study the variability of a marker during different period of time, for instance the short-term variability could be the intra-day variability and the long-term variability the variability during all the follow-up.

IV.6 Supplementary materials

IV.6.1 Appendix A: Process of data generation

For each individual, the age of dementia onset and death were generated in the following steps:

1. Generate T_{01i} and T_{02i} using the Brent's univariate root-finding method according to the following proportional hazards models:

$$\lambda_i^{kl}(t) = \lambda_0^{kl}(t) \exp\left(\alpha_1^{kl}\tilde{y}_i(t) + \alpha_2^{kl}\tilde{y}_i'(t) + \alpha_\sigma^{kl}\sigma_i + \alpha_\kappa^{kl}\kappa_i\right) \tag{IV.9}$$

   with $\lambda_0^{kl}(t) = \eta^{kl}t^{\eta^{kl}-1}e^{\alpha_0^{kl}}$ being a Weibull function.

2. If $T_{01i} > A_{0i} + C_i$ and $T_{02i} > A_{0i} + C_i$ (with C_i the right censoring and A_{0i} the age at inclusion) then the individual is free of any event at the end of the follow-up: $T_i = A_{0i} + C_i$, $\delta_i^{dem} = 0$, $\delta_i^{death} = 0$ and $L_i = \max(t_{ij})$ (the last visit).
3. Else, if $T_{02i} < T_{01i}$ and $T_{02i} \leq A_{0i} + C_i$, then the individual is dead at T_{02} without dementia diagnosis: $T_i = T_{02}$, $\delta_i^{dem} = 0$, $\delta_i^{death} = 1$ and $L_i = \max(t_{ij})$.
4. Else, if $T_{01i} \leq T_{02i}$ and $T_{01i} \leq A_{0i} + C_i$, then the individual developed dementia and we generate the time to death from dementia T_{12i} according to the intensity model for transition 1-2;
   (a) if $T_{12i} > A_{0i} + C_i$ then the subject is alive at the end of the follow-up:
      — if $T_{01i} > \max(t_{ij})$: the dementia was not diagnosed so $T_i = A_{0i}+C_i$, $\delta_i^{dem} = 0$, $\delta_i^{death} = 0$ and $L_i = \max(t_{ij})$.
      — else, dementia is diagnosed: $T_i = A_{0i} + C_i$, $\delta_i^{dem} = 1$, $\delta_i^{death} = 0$, $L_i = \max\{t_{ij}/t_{ij} \leq T_{01i}\}$ (ie. the last visit before T_{01i}) and $R_i = \min\{t_{ij}/t_{ij} \geq T_{01i}\}$ (ie. the first visit after T_{01i}).
   (b) else ($T_{12i} \leq A_{0i} + C_i$), then the subject died at T_{12i}:
      — if $T_{01i} > \max(t_{ij})$: the dementia was not diagnosed so $T_i = T_{12}$, $\delta_i^{dem} = 0$, $\delta_i^{death} = 1$ and $L_i = \max(t_{ij})$.
      — if T_{01i} and T_{12i} are in the same interval between two consecutive visits, then the subject was not diagnosed: $T_i = T_{12i}$, $\delta_i^{dem} = 0$, $\delta_i^{death} = 1$ and $L_i = \max\{t_{ij}/t_{ij} < T_{12i}\}$ (ie the last visit before death).
      — else the subject is diagnosed: $T_i = T_{12i}$, $\delta_i^{dem} = 1$, $\delta_i^{death} = 1$, $L_i = \max\{t_{ij}/t_{ij} \leq T_{01i}\}$ and $R_i = \min\{t_{ij}/t_{ij} \geq T_{01i}\}$.
5. Delete the measurements of the longitudinal marker measured after T_{02i} or T_{12i}.

IV.6.2 Appendix B: Histograms of inter and intra-visit variabilities

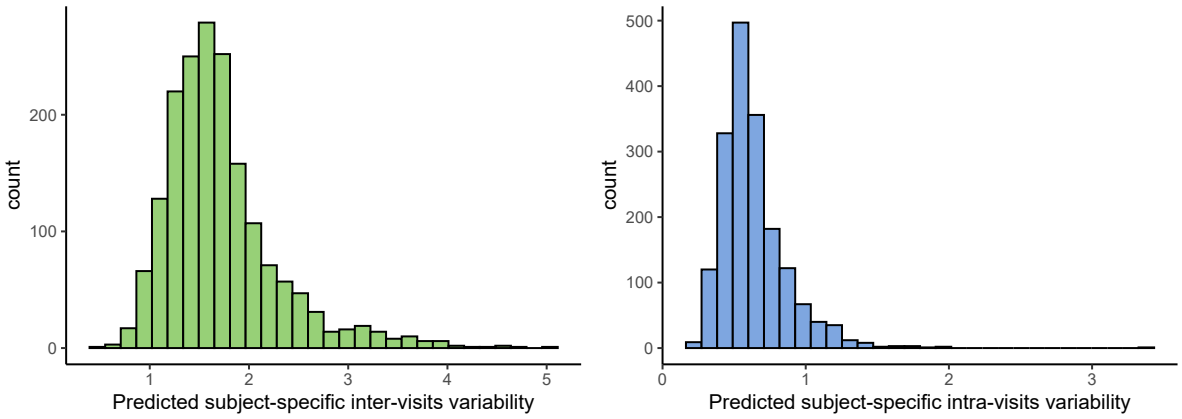

Figure IV.3 – Histogram of predicted subject-specific inter-visit variability (left) and intra-visits variability (right) from the 3C-Bordeaux cohort ($N = 1788$).

Chapitre V

LSJM : package R pour modèles avec variance résiduelle hétérogène

Sommaire

RÉSUMÉ

Ce chapitre présente le package `LSJM` développé en `R` et `C++`. Ce package met à disposition plusieurs fonctions afin d'estimer des modèles basés sur la théorie des modèles *location-scale* linéaires mixtes et des modèles conjoints *location-scale*. Il généralise la présentation des modèles proposés en chapitres III et IV pour mettre en évidence toutes les variantes des modèles *location-scale*, définis à la fois par le choix du modèle linéaire mixte *location-scale* et la nature du modèle de survie. Il permet ainsi d'estimer des modèles avec variance résiduelle hétérogène, soit dépendante du temps et d'autres covariables soit distinguant les variabilités inter-visites et intra-visite. Pour les données de survie, le package `LSJM` permet de traiter un ou deux événements (semi-)compétitifs, la censure par intervalle ainsi que l'entrée retardée. Après avoir rappelé les notations et la procédure d'estimation, les méthodes d'évaluation des modèles et les prédictions dynamiques déjà présentées dans les chapitres précédents, ce chapitre présente l'implémentation des différentes fonctions `R`. Enfin, deux applications sur un sous-échantillon de la cohorte des Trois-Citées permettent d'illustrer l'utilisation du package en étudiant l'effet de la variabilité de la PAS sur le risques de démence et de décès. La première application illustre un modèle conjoint pour risques compétitifs dans lequel la variance résiduelle se distingue entre les variabilités inter-visites et intra-visite. La deuxième illustration est un modèle conjoint pour risques semi-compétitifs et traitant la censure par intervalle avec une variance résiduelle dépendante du temps.

Ce chapitre fera l'objet d'un troisième article qui sera soumis à une revue spécialisée pour les logiciels statistiques. Ainsi, la rédaction de ce chapitre correspond à la version courante de l'article rédigé en anglais.

V.1 Introduction

Joint modeling of longitudinal and time-to-event data has become an essential tool to simultaneously analyze longitudinal data (measured repeatedly over time) and survival events (such as disease occurrence or death) (Wulfsohn and Tsiatis, 1997; Tsiatis and Davidian, 2004). The strength of joint models relies on their ability to capture the relationship between the evolution of a biomarker and the risk of an event occurring. Traditional approaches that analyse these two types of data separately often miss critical interactions, whereas joint models integrate both, offering a more comprehensive understanding of how an individual's biomarker trajectory influences their risk of future events. The joint models combine a mixed model fitting the mean behavior over time of the longitudinal outcome and a survival model for the risk of developing one or more health events. Their estimations is based on the joint likelihood of these two models, whose dependency structure defines the type of joint model used (Wulfsohn and Tsiatis, 1997; Proust-Lima et al., 2009). The first concerns latent class joint models (Proust-Lima et al., 2009). They assume that the population of interest is heterogeneous, and that there are homogeneous sub-populations that differ both on the mean trajectory of the longitudinal marker and on the risk of developing the event of interest. In addition, conditionally on class membership, the longitudinal variable and event times are considered independent. The second type of joint models concerns shared-random effect models (Rizopoulos, 2012). Here, the dependence structure is defined by one or more functions of the random effects from the mixed model which are included as covariates in the survival model. They thus make it possible to assess the impact of a characteristic of the mean trajectory of the marker on the risk of developing the event. In this approach, the longitudinal variable and time-to-event are assumed to be independent conditionally on random effects.

Over the years, several `R` packages have been developed to fit joint models, each offering specific features and varying estimation methods. The `JM` package (Rizopoulos, 2010) was one of the first `R` packages designed for fitting joint models with shared random effects. It enables the estimation of a joint model with a single event risk, using only the current value as the association function between the two submodels. The `JMbayes` package (Rizopoulos, 2016) can be seen as an extension of `JM`, incorporating Bayesian methods using Markov Chain Monte Carlo (MCMC) techniques to estimate the posterior distributions of model parameters. This package allows for the consideration of different types of longitudinal outcomes by using a generalized linear mixed model for the longitudinal part. The baseline risk is modeled using penalized B-splines, and left truncation, as well as time-dependent and exogenous covariates, can be accounted for in the survival submodel. Furthermore, various associations between

the marker and survival can be specified, such as the current value, slope, or the cumulative effect. The integral in the survival function is computed using either Gauss-Kronrod or Gauss-Legendre quadrature. The `JMbayes2` (Rizopoulos et al., 2023) package is an evolution of `JMbayes`, offering a broader range of longitudinal and survival submodels. It allows for the estimation of multi-marker joint models and the inclusion of competing risks. In a frequentist framework, the `joineR` package allows to fit joint model by maximizing the likelihood using EM algorithm and Gauss-Hermite integration on the random effects. The authors considered a semi-parametric approach, meaning that the baseline hazard is unspecified. It also handles competing risks, and a second version of the package, `joineRML` (Hickey et al., 2018), allows for multi-marker models, using Monte Carlo integration on the random effects. The `rstanarm` package (Brilleman et al., 2018) also provides a Bayesian approach using advanced sampling techniques, such as the No-U-Turn Sampler (NUTS). The `INLAjoint` package (Rustand et al., 2023) introduces an alternative Bayesian approach by leveraging Integrated Nested Laplace Approximation (INLA) for fast estimation. It supports multiple markers with different distributions, as well as competing risks and multi-state models. Lastly, the `lcmm` package (Proust-Lima et al., 2017) allows the estimation of latent class joint models, including competing risks, by maximizing the likelihood using the Marquardt-Levenberg algorithm (Levenberg, 1944; Marquardt, 1963). However, all these packages always assume homogeneity in the residual variance, which might not hold in more complex data settings. More recently, Gao et al. (2011) and Barrett et al. (2019) have extended the joint model with shared random effects to include a subject-specific residual variability of the marker in the risk of the event. These models combine a location-scale linear mixed model and a proportional hazard model. They have been motivated by the study of the impact of the variability of blood pressure on the risk of cardiovascular events and by the study of the variability of intraocular pressure as an independent risk of primary open-angle glaucoma. In medical research, many current hypotheses regards the role of the variability of markers or risk factors (blood pressure, glycemia ...) on the risk of clinical events (cardiovascular diseases, dementia ...). However, the models proposed by Gao et al. (2011) and Barrett et al. (2019) have strong constraints (constant residual variance, no competing risks ...) and no package was proposed for estimating these models.

On the other hand, several softwares exist to estimate linear location-scale mixed models that accounts for heteroscedasticity in the data by modeling the residual variance as a function of covariates or time. Softwares `MIXREGLS` (Hedeker and Nordgren, 2013), `MixWild` (Dzubur et al., 2020) and the `R`-package `LMMEL` (Martin and Rast, 2024) have been proposed to estimate such kind of models. However, they do not support jointly modeling the impact of a marker's residual variability on event risk, even though, in biomarker studies, it is clinically

relevant to consider this variability as a potential risk factor.
Despite the wide range of available software, there remains a growing need for more flexible joint models that can simultaneously manage longitudinal data in the presence of heteroscedasticity and (semi-)competitive interval-censored events, where the risk is influenced by the variability of the longitudinal outcome. These advanced models are crucial for addressing contemporary clinical hypotheses, which are often closely tied to significant public health challenges.
In response to this gap, we developed the `LSJM` package, designed to estimate joint models with subject-specific residual variance while addressing the challenges discussed in Courcoul et al. (2024b) and Courcoul et al. (2024a). Firstly, this package enables the modeling of longitudinal data with heteroscedasticity, allowing both the mean trajectory and the residual variance, to depend on time, covariates and random effects. Secondly, it supports joint modeling of a continuous longitudinal marker with heteroscedasticity, combined with survival data. The survival component is highly flexible, allowing for the modeling of a single event risk, competing risks, or semi-competing risks using either a cause-specific model or an illness-death model for interval-censored data. This last feature is unique among packages estimating joint models with shared-random effects. Additionally, the joint model estimation accommodates left truncation.
The paper is organized as follows. Section 2 defines the statistical models implemented in the package and Section 3 details the estimation process. Section 4 describes the post-fit analyses. Section 5 details the implementation of the estimation functions and the post-fit functions. Finally, Section 6 provides two examples based on the `threeC` dataset, a random subsample of 500 subjects from the French Three-Cities cohort (3C Study Group, 2003) available in the package and Section 7 concludes.

V.2 Location-Scale Joint Models

This section describes each family of statistical models implemented in the package. The first subsection describes the location-scale linear mixed models including the standard linear mixed model. The second subsection is dedicated to the presentation of the location-scale joint models for jointly analyzing a longitudinal marker with heteroscedasticity and complex survival data.

V.2.1 The location-scale mixed models (LSMM)

Section V.2.1.1 presents the standard linear mixed model, which is a special case of the LSMM managed by the package. This section presents the general notations of the longitudinal model.

V.2.1.1 The standard linear mixed model

Let us consider a sample of N individuals. For each individual $i \in \{1, ..., N\}$, we consider the n_i-vector of repeated measures $Y_i = (Y_{i1}, ..., Y_{in_i})^\top$ with Y_{ij} the value of the longitudinal outcome of individual i at time t_{ij} ($j = 1, \ldots, n_i$). Laird and Ware (1982) proposed the following linear mixed model:

$$Y_{ij} = Y_i(t_{ij}) = \widetilde{Y}_i(t_{ij}) + \epsilon_{ij} = X_{ij}^\top \beta + Z_{ij}^\top b_i + \epsilon_i(t_{ij}) \tag{V.1}$$

where X_{ij} and Z_{ij} vectors of explanatory variables for subject i at visit j, respectively associated with the fixed-effect vector β and the subject-specific random-effect vector b_i such as $b_i \sim \mathcal{N}(0, B)$, with B an unspecified matrix. The measurements error ϵ_{ij} are independent Gaussian errors with variance σ_ϵ^2.

V.2.1.2 LSMM with time-dependent variability

Some authors have proposed to overcome the homoscedasticity assumption in the linear mixed model by defining the residual variance as a function of covariates and random effects (Aitkin, 1987; Foulley et al., 1992; Lin et al., 1997; Hedeker et al., 2008). These models are typically called location-scale mixed model. It is defined by equation (V.1) completed by the following specification for the residual error:

$$\epsilon_{ij}(t_{ij}) \sim \mathcal{N}(0, \sigma_i^2(t_{ij})) \quad \text{with} \quad \log(\sigma_i(t_{ij})) = O_{ij}^\top \mu + M_{ij}^\top \tau_i \tag{V.2}$$

with O_{ij} and M_{ij} two vectors of explanatory variables for subject i at visit j, respectively associated with the fixed-effect vector μ, and the subject-specific random-effect vector τ_i, such as

$$\begin{pmatrix} b_i \\ \tau_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_b & \Sigma_{\tau b} \\ \Sigma_{\tau b}^\top & \Sigma_\tau \end{pmatrix} \right) \tag{V.3}$$

As conventionally assumed in practice, the random effects b_i can be considered independent of the errors by setting $\Sigma_{\tau b} = 0$.

V.2.1.3 LSMM distinguishing within and between visits variabilities

In some studies, several measures of the marker are collected at each measurement time. For instance, in medical studies, 2 or 3 measures of blood pressure are generally collected at each visit and intra-visit variability can be an interesting indicator. Thus, we propose a LSMM distinguishing within and between visits variabilities. Let us consider an additional level in longitudinal data, grouping repeated measurements according to measurement time. For example, for each subject i, $(i = 1, \ldots, N)$, Y_{ijl} is the marker value for measure l $(l = 1, \ldots, n_{ij})$, at visit j $(j = 1, \ldots, n_i)$ and time t_{ij}. For each visit j, subject i can have n_{ij} measurements of the longitudinal marker. Y_i is a vector of dimension $\sum_{j=1}^{n_i} n_{ij}$ containing all marker measurements for the subject i.

We then defined the following LSMM model decomposing within and between visits individual residual variance:

$$\begin{cases} Y_{ijl} = \widetilde{Y}_i(t_{ij}) + \epsilon_{ij} + \nu_{ijl} = X_{ij}^\top \beta + Z_{ij}^\top b_i + \epsilon_{ij} + \nu_{ijl}, \\ \epsilon_{ij} \sim \mathcal{N}(0, \sigma_i^2) \quad \text{with} \quad \log(\sigma_i) = \mu_\sigma + \tau_{\sigma i}, \\ \nu_{ijl} \sim \mathcal{N}(0, \kappa_i^2) \quad \text{with} \quad \log(\kappa_i) = \mu_\kappa + \tau_{\kappa i}, \end{cases} \tag{V.4}$$

with X_{ij} and Z_{ij} two vectors of explanatory variables for subject i at visit j, respectively associated with the fixed-effect vector β and the subject-specific random-effect vector b_i. The parameter μ_σ and μ_κ correspond to the subject-specific fixed intercepts for the between-visits and within-visit variances respectively. The subject-specific random-effect b_i and $\tau_i = (\tau_{\sigma i}, \tau_{\kappa i})^\top$ are assumed to be Gaussian as presented in equation (V.3).

In the following, we denote r_i the vector of all random effects. For the standard mixed model, $r_i = b_i$ and for location-scale mixed models with heteroscedasticity $r_i = (b_i, \tau_i)^\top$.

V.2.2 The location-scale joint models (LSJM)

A joint model is composed of two submodels: a linear mixed model as defined by (V.1), or a LSMM defined either by (V.2) or by (V.4) and a survival model. Three main survival processes represented in Figure V.1 are considered and detailed in this section. Each of the defined joint model falls within the framework of shared random effect models where the time-to-event may depend on the value of the marker, and/or the trajectory slope, and/or its variability. In the following we will describe the different cases. Note that Section V.2.2.1 presents the basic survival model and allows the introduction of both association structures linking LSMM and survival model and the different baseline risk functions proposed by the

LSJM package.

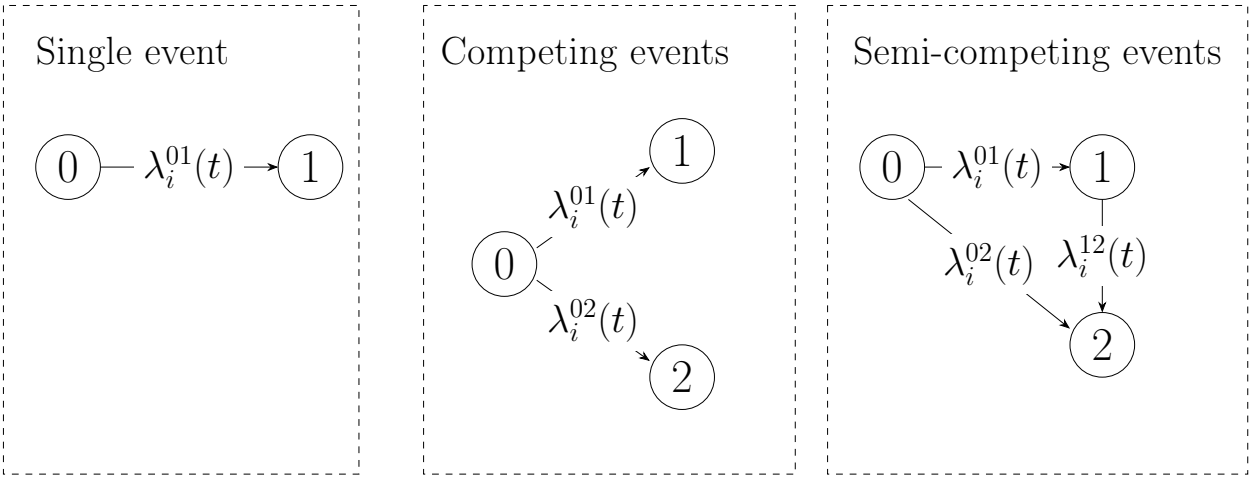

Figure V.1 – Graphical representations of the three different patterns for the survival part.

V.2.2.1 A single event: a proportional hazard model

In the simplest case we consider only a single event. We denote $T_i = \min(T_{i1}^*, C_i)$ the observed time with T_{i1}^* the real time for the event and C_i the censoring time for the ith individual. Censoring time is assumed to be non-informative (i.e. independent from the time to event given characteristics of marker trajectories and covariates). We then denote $\delta_i \in \{0, 1\}$ the individual event indicator such as $\delta_i = 1$ if $T_{i1}^* \leq C_i$ and $\delta_i = 0$ otherwise. The risk of event is then described using a proportional hazard model:

$$\lambda_i^{01}(t|r_i) = \lambda_0(t) \exp\left(W_i^{01\top}\gamma^{01} + g_y^{01}(b_i, t)^\top \alpha_b^{01} + g_\tau^{01}(\tau_i, t)^\top \alpha_\tau^{01}\right), \tag{V.5}$$

with $\lambda_0(t)$ the baseline risk function, W_i a vector of baseline covariates associated with the regression coefficient γ. The α_b is a vector corresponding to the regression coefficients associated with the function g_y, possibly multi-dimentional, which measures the association between the risk and the mean evolution of Y. This function can be defined in the package by :

- the current value: $g_y(b_i, t) = \tilde{y}_i(t)$
- the current slope: $g_y(b_i, t) = \tilde{y}_i'(t) = \frac{\partial \tilde{y}_i(t)}{\partial t}$
- both current value and slope: $g_y(b_i, t) = (\tilde{y}_i(t), \tilde{y}_i'(t))$
- the random effects: $g_y(b_i, t) = b_i$

The α_τ is a vector corresponding to the regression coefficients associated with the function g_τ which measures the association between the event risk and the evolution of the residual

variance of the marker. If a standard linear mixed model with an homogeneous variability is estimated for the longitudinal part then $g_\tau(\tau_i, t) = 0$. In the case of a LSMM:

— $g_\tau(\tau_i, t) = \sigma_i(t)$ if a covariate or time-dependent variability is considered as in equation (V.2). Thus the risk depends on the current value of the variability $\sigma_i(t)$ of the marker.

— $g_\tau(\tau_i, t) = (\sigma_i, \kappa_i)^\top$ if within and between visits variabilities are considered as in equation (V.4). The vector α_τ thus includes α_σ and α_κ, the regression coefficients associated with the between-visits residual variability σ_i and the within-visit residual variability κ_i, respectively. Note that the model can be adjusted on both variability or only on one.

For the baseline risk function, different parametric forms can be considered :

— Exponential specified by $\lambda_0(t) = \exp(\alpha_0)$

— Weibull specified by $\lambda_0(t) = \zeta^2 t^{\zeta^2 - 1} \exp(\alpha_0)$

— Gompertz specified by $\lambda_0(t) = \kappa_1^2 \exp(\kappa_2 t)$

— Cubic B-splines, with Q knots, specified by $\lambda_0(t) = \exp\left(\sum_{q=1}^{Q+4} \eta_q B_q(t, \nu)\right)$ where $B_q(t, \nu)$ is the q-th basis function of B-splines with the knot vector ν and η_q is the associated parameter to be estimated.

V.2.2.2 Competing events: cause-specific model

Instead of considering a single event, we can consider two causes of event by denoting $T_i = \min(T_{i1}^*, T_{i2}^*, C_i)$ the observed time with T_{ik}^* the real time for the event k ($k = 1, 2$) and C_i the censoring time for the ith individual. We denote also $\delta_i \in \{0, 1, 2\}$ the individual event indicator such as $\delta_i = k$ if $T_i = T_{ik}^*$ for $k \in \{1, 2\}$ and $\delta_i = 0$ otherwise. In this case, the cause-specific proportional hazard model is defined by:

$$\lambda_i^{0k}(t) = \lambda_0^{0k}(t) \exp\left(W_i^{0k\top} \gamma^{0k} + g_y^{0k}(b_i, t)^\top \alpha_b^{0k} + g_\tau^{0k}(\tau_i, t)^\top \alpha_\tau^{0k}\right) \tag{V.6}$$

As previously, $\lambda_0^{0k}(t)$ is the baseline risk function for event k, W_i^{0k} a vector of baseline covariates associated with the regression coefficients γ^{0k}, α_b^{0k} and α_τ^{0k} the vectors of parameters associated with functions $g_y^{0k}(b_i, t)$ and $g_\tau^{0k}(\tau_i, t)$ defined as previously in section V.2.2.1, respectively.

V.2.2.3 Semi-competing event: illness-death model

The illness-death model is a multi-state model which describes the transitions from an initial state (e.g., alive and disease-free) to an absorbing state (e.g death) either directly or

via an intermediate state (e.g. disease such as dementia). In some applications, for example in the study of dementia and death (state (0) for healthy, state (1) for dementia and (2) for death), the transition times from the initial state (0) (healthy) to the intermediate state (1) (dementia) can be interval-censored for some or all subjects. This typically occurs when the intermediate state is only checked at visit times. For instance, if a subject is diagnosed with dementia at a visit time R, and was free of dementia at previous visit time L, the onset of dementia is interval-censored between L and R for that subject. The occurrence of death can lead to a complex observational pattern: if a subject dies without a prior diagnosis of dementia, it is not known whether the disease onset occurred between the last visit and the time of death. We assume that time of transition to state (1) could be interval-censored while time of transition to state (2) is known. The vector of collected data for time-to-events is given by $D_i = (T_{0i}, L_i, R_i, \delta_i^{(1)}, T_i, \delta_i^{(2)})^\top$ where T_{0i} is the time at inclusion in case of delayed entry, L_i is the time at the last visit where the subject was still in state (0), R_i is the time at the first visit where the subject is seen in state (1) (in case of no observation in state (1), R_i is undefined), T_i is the minimum between the time of transition to state (2) and the time of right censoring (e.g. the end of the follow-up), $\delta_i^{(1)} = \mathbb{1}_{R_i<T_i}$ is the indicator of state (1) observation and $\delta_i^{(2)}$ is the indicator of state (2).

The transition intensities from state $k \in \{0,1\}$ to state $l \in \{1,2\}$ are defined by a proportional hazards model under the Markovian hypothesis:

$$\lambda_i^{kl}(t|b_i,\tau_i) = \lambda_0^{kl}(t)\exp\left(W_i^{kl^\top}\gamma^{kl} + g_y^{kl}(b_i,t)^\top\alpha_b^{kl} + g_\tau^{kl}(\tau_i,t)^\top\alpha_\tau^{kl}\right) \tag{V.7}$$

with $\lambda_0^{kl}(t)$ the baseline risk function, W_i^{kl} a vector of baseline covariates associated with the regression coefficients γ^{kl}, α_b^{kl} and α_τ^{kl} the vectors of regression parameters associated with functions $g_y^{kl}(b_i,t)$ and $g_\tau^{kl}(\tau_i,t)$ defined as previously in section V.2.2.1, respectively.

Figures V.2, V.3 and V.4 illustrate the graphical representations for the LSMMs (right box) and LSJMs for each different considered fit of variability (standard, time-dependent or combining between and within visits) to get a better view of the dependence structure for the model..

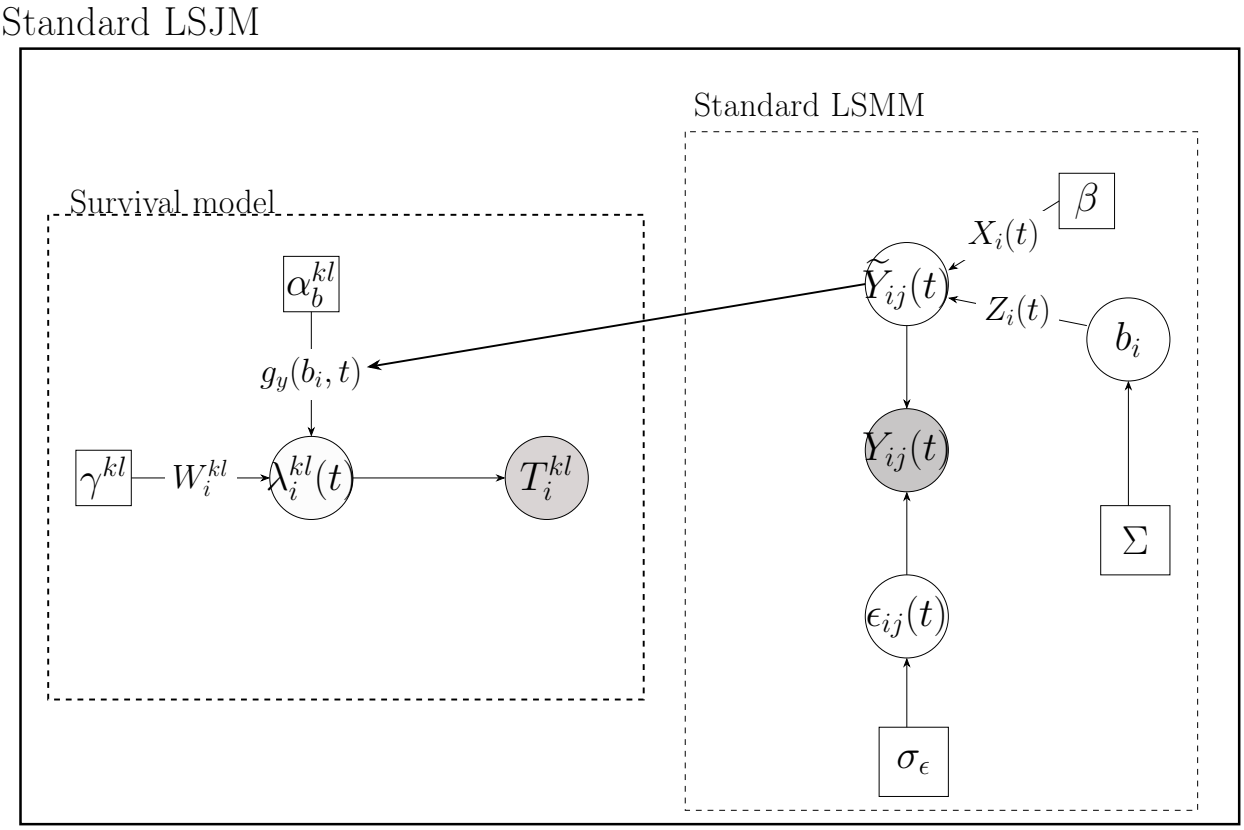

Figure V.2 – Graphical representation of the linear mixed model (right box) and the joint model (global box). The survival part could be defined either by only one cause-specific risk function (for one event), two cause-specific risk functions (for two competing events) or an illness-death model (for two semi-competing events). Circles denote random variables (white circles denote latent variables) and square boxes denote unknown parameters.

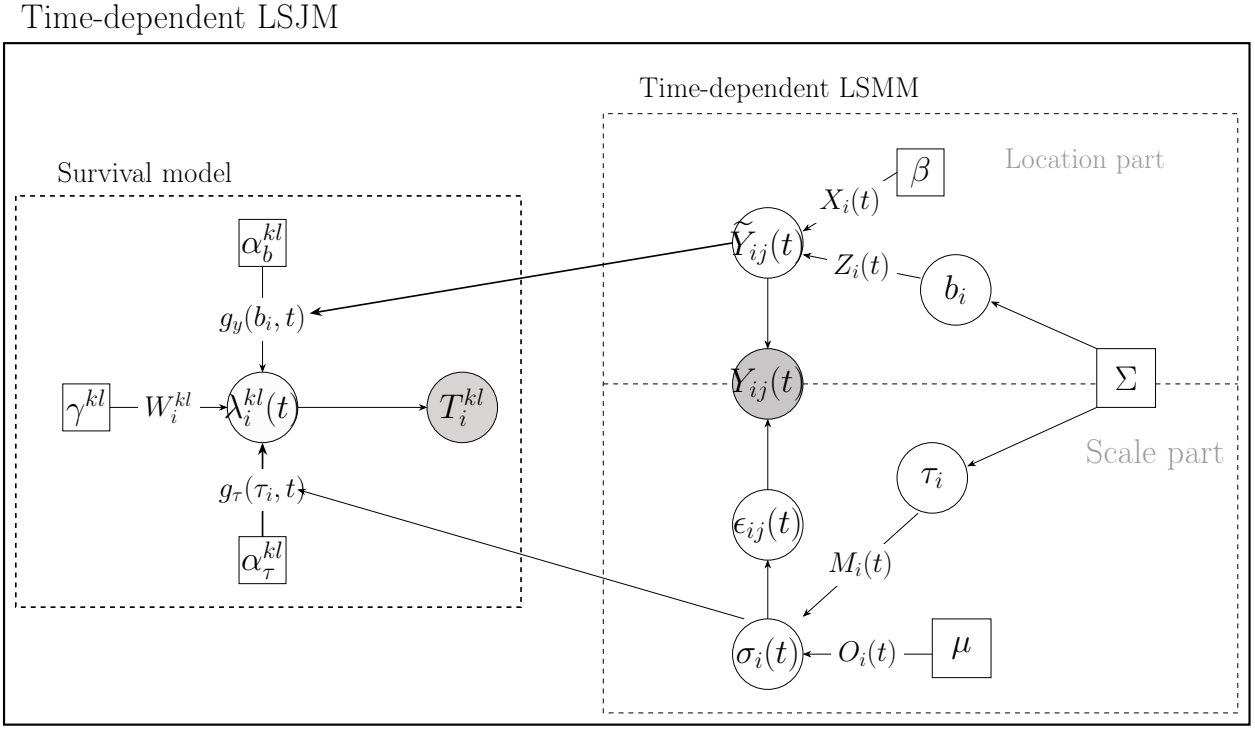

Figure V.3 – Graphical representation of the location-scale mixed model (right box) and location-scale joint model (global box) with time-dependent residual variance. The survival part (left box) could be defined either by only one cause-specific risk function (for one event), two cause-specific risk functions (for two competing events) or three transition intensities of an illness-death model (for two semi-competing events). Circles denote random variables, the white ones denote latent variables, and square boxes denote unknown parameters. $X_i(t), Z_i(t), O_i(t)$ and $M_i(t)$ are the matrices of covariates for individual i at time t.

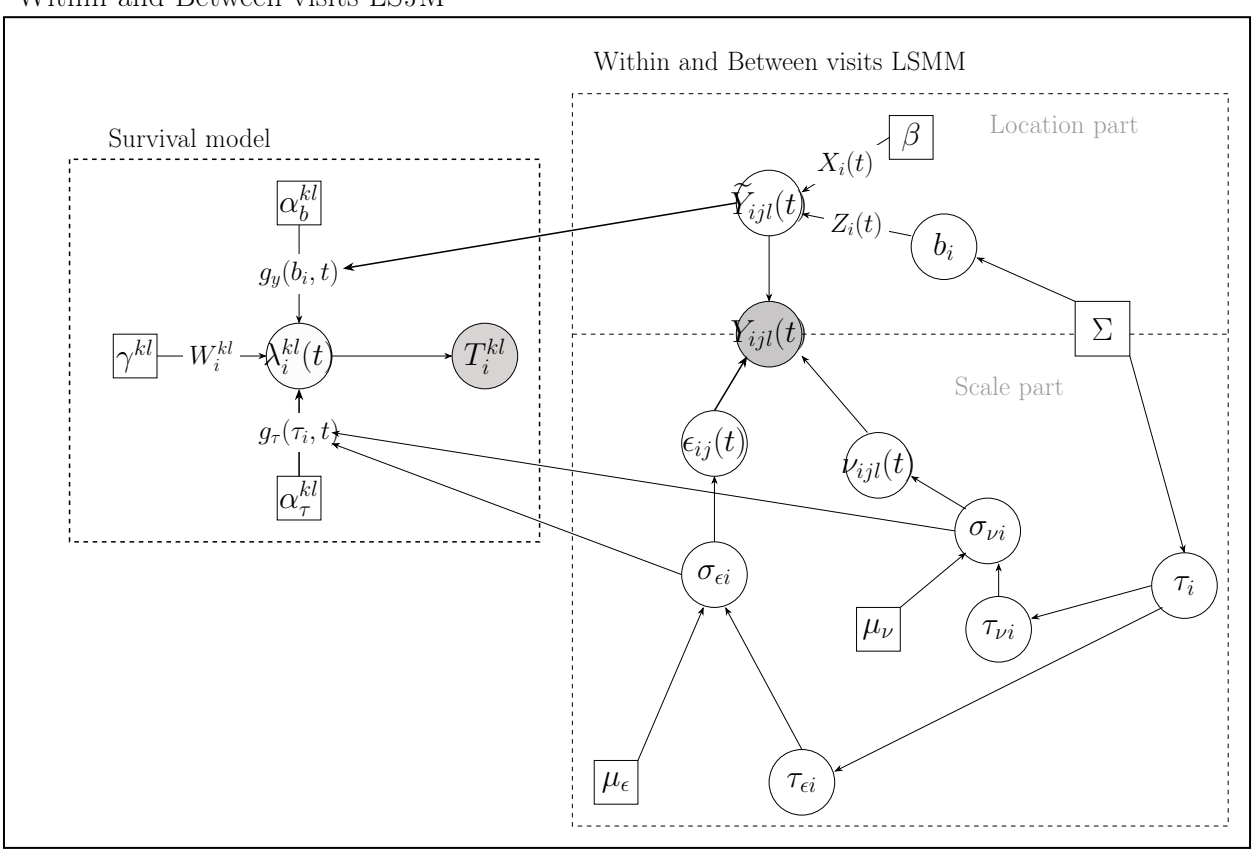

Figure V.4 – Graphical representation of the location-scale mixed model (right box) and location-scale joint model (global box) distinguishing within and between visits residual variabilities. The survival part (left box) could be defined either by only one cause-specific risk function (for one event), two cause-specific risk functions (for two competing events) or three transition intensities of an illness-death model (for two semi-competing events). Circles denote random variables, the white ones denote latent variables, and square boxes denote unknown parameters. $X_i(t), Z_i(t), O_i(t)$ and $M_i(t)$ are the matrices of covariates for individual i at time t.

V.3 Estimation

All these models are estimated within the maximum likelihood framework. We denote θ the entire vector of parameters involved in the model estimation. We consider the log-likelihood defined by $\ell(\theta) = \sum_{i=1}^{N} \log(\mathcal{L}_i(\theta))$ with $\mathcal{L}_i$ the individual contribution to the likelihood, which will be defined in this section for each of the models presented above.

V.3.1 Individual contributions to the likelihoods

V.3.1.1 Linear location-scale mixed model

The individual contribution to the likelihood of a LSMM is:

$$\mathcal{L}_i(\theta; Y_i) = \int f(Y_i|r_i;\theta) f(r_i;\theta) dr_i = \int \prod_{j=1}^{n_i} f(Y_{ij}|r_i;\theta) f(r_i;\theta) dr_i. \tag{V.8}$$

For the standard linear mixed model and the LSMM with a time-dependent residual variability, $f(Y_{ij}|r_i;\theta)$ is an univariate Gaussian density and $f(r_i;\theta)$ is a multivariate Gaussian density with zero-mean and covariance Σ. Note that $r_i = b_i$ for the standard model and $r_i = (b_i, \tau_i)^\top$ for the LSMM defined by equation (V.3).

For the LSMM with both between and within visits variabilities, Y_{ij} is a vector of dimension n_{ij} and $f(Y_{ij}|r_i;\theta) = f(Y_{ij}|b_i, \tau_{\sigma_i}, \tau_{\kappa_i};\theta)$ is a multivariate Gaussian density with covariance matrix $\Sigma_{Y_{ij}|r_i} = \sigma_i^2 \mathbb{1}_{n_{ij}}^\top \mathbb{1}_{n_{ij}} + \kappa_i^2 I_{n_{ij}}$, where $\mathbb{1}_{n_{ij}}$ is the n_{ij}-length vector of 1 and $I_{n_{ij}}$ is the identity matrix of size $n_{ij} \times n_{ij}$.

V.3.1.2 Joint models

Denoting D_i the vector of survival data, under the assumption of independence between D_i and Y_i, conditionally on the random effects, and assuming independent censoring, the individual contribution to the likelihood for the LSJM is defined by:

$$\mathcal{L}_i(\theta; Y_i, D_i) = \int f(Y_i|r_i;\theta) f(r_i;\theta) f(D_i|r_i;\theta) dr_i \tag{V.9}$$

where $f(Y_i|r_i;\theta)$ and $f(r_i;\theta)$ are defined in section V.3.1.1 following the LSMM sub-model considered.

A single or two competing events

In case of a single event ($K = 1$) or two competing events ($K = 2$) then :

$$f(D_i|r_i;\theta) = \exp\left(-\sum_{k=1}^{K}\Lambda_i^{0k}(T_i|r_i;\theta)\right)\prod_{k=1}^{K}\lambda_i^{0k}(T_i|r_i;\theta)^{\mathbb{1}_{\delta_i=k}} \tag{V.10}$$

with $\Lambda_i^{0k}(T_i|r_i;\theta)$ the cumulative risk function given by:

$$\Lambda_i^{0k}(T_i|r_i;\theta) = \int_0^t \lambda_i^{0k}(u|r_i;\theta)du \tag{V.11}$$

An illness-death model

In case of interval-censored semi-competing events and illness-death model, $f(D_i|r_i;\theta)$ depends on the subject trajectory as illustrated on Figure V.5. To present the different definitions of $f(D_i|r_i;\theta)$ according to subjects observations, we deliberately omit the conditioning on random effects and parameters for ease of notations:

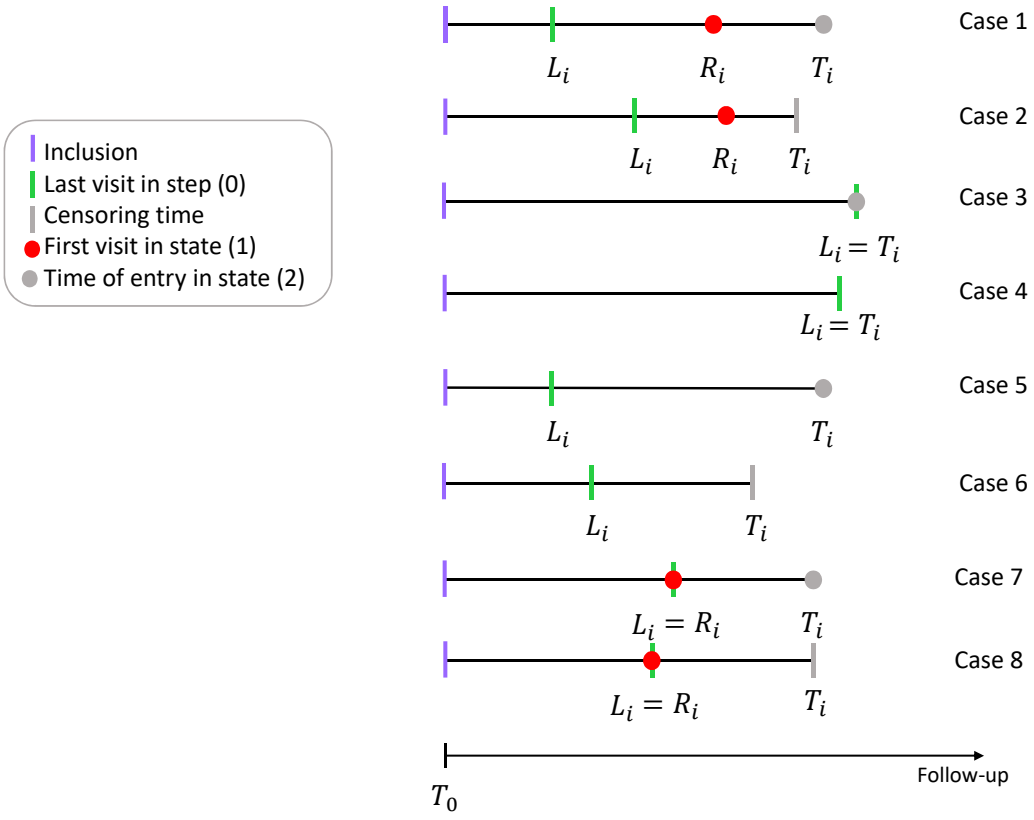

Figure V.5 – Possible patterns of state of observation for interval censored semi-competing events.

Cases 1 and 2: subject in state (0) until L_i, observed in state (1) at R_i, remained in state (1) until T_i, and either entered in state (2) (case 1) or censored at T_i (case 2):

$$f(T_{0i}, L_i, R_i, \delta_i^{(1)} = 1, T_i, \delta_i^{(2)}) = \int_{L_i}^{R_i} e^{-\Lambda_i^{01}(u)-\Lambda_i^{02}(u)} \lambda_i^{01}(u) e^{-(\Lambda_i^{12}(T_i)-\Lambda_i^{12}(u))} \lambda_i^{12}(T_i)^{\delta_i^{(2)}} du$$

Cases 3 and 4: subject in state (0) until T_i ($T_i = L_i$), and either entered in state (2) (case 3) or censored at T_i (case 4):

$$f(T_{0i}, L_i, R_i, \delta_i^{(1)} = 0, T_i, \delta_i^{(2)}) = e^{-\Lambda_i^{01}(T_i)-\Lambda_i^{02}(T_i)} \lambda_i^{02}(T_i)^{\delta_i^{(2)}}$$

Cases 5 and 6: subject in state (0) at his/her last visit $L_i < T_i$, and either entered in state (2) (case 5) or censored at T_i (case 6):

$$\begin{aligned} f(T_{0i}, L_i, R_i, \delta_i^{(1)} = 0, T_i, \delta_i^{(2)}) &= e^{-\Lambda_i^{01}(T_i)-\Lambda_i^{02}(T_i)} \lambda_i^{02}(T_i)^{\delta_i^{(2)}} + \\ &\int_{L_i}^{T_i} e^{-\Lambda_i^{01}(u)-\Lambda_i^{02}(u)} \lambda_i^{01}(u) e^{-(\Lambda_i^{12}(T_i)-\Lambda_i^{12}(u))} \lambda_i^{12}(T_i)^{\delta_i^{(2)}} du \end{aligned}$$

The likelihood accounts for the two possible trajectories of this subject: either direct transition from state (0) to state (2) or unobserved transition to state (1) between L_i and T_i.

Cases 7 and 8: subject observed in state (1) at $L_i = R_i$ (exact date of transition), remained in state (1) until T_i, and either entered in state (2) (case 7) or censored at T_i (case 8):

$$f(T_{0i}, L_i, R_i, \delta_i^{(1)} = 1, T_i, \delta_i^{(2)}) = e^{-\Lambda_i^{01}(L_i)-\Lambda_i^{02}(L_i)} \lambda_i^{01}(L_i) e^{-(\Lambda_i^{12}(T_i)-\Lambda_i^{12}(L_i))} \lambda_i^{12}(T_i)^{\delta_i^{(2)}}$$

Left truncation
In case of delayed entry, the individual contribution to the likelihood is divided by the probability to be free of any event at entry time T_{0i}:

$$\mathcal{L}_i^{DE}(\theta; Y_i, D_i) = \frac{\mathcal{L}_i(\theta; Y_i, D_i)}{\int \exp(-\sum_{k=1}^{K} \Lambda_i^{0k}(T_{0i}|r_i; \theta) f(r_i; \theta) dr_i} \tag{V.12}$$

V.3.2 Computational aspects

V.3.2.1 Integral computation

In each model, the integral over the random effects is computed by a Quasi Monte Carlo (QMC) approximation (Pan and Thompson, 2007), using deterministic quasi-random se-

quences. The approximation of the integral is defined by:

$$\mathcal{L}_i(\theta; Y_i, D_i) \simeq \frac{1}{S}\sum_{s=1}^{S} p(Y_i, D_i|b_i^s, \tau_i^s; \theta) \tag{V.13}$$

where $(b_i^1, ..., b_i^S)$ and $(\tau_i^1, ..., \tau_i^S)$ are draws of a S-sample in the sobol sequel for the distribution $f(b_i, \tau_i; \theta)$. The sobol sequel is generated using the `spacefillr` `R` package (Morgan-Wall, 2024). Moreover, the cumulative risk functions are approached using the Gauss-Kronrod quadrature with 7 or 15 points (Gonnet, 2012).

V.3.2.2 Optimization algorithm

Parameter estimation is performed by maximizing the log-likelihood function $\ell(\theta; Y_i, D_i) = \log\left(\prod_{i=1}^{N} \mathcal{L}_i(\theta; Y_i, D_i)\right)$. This optimization is carried out using the `marqLevAlg` R-package, which applies the Marquardt-Levenberg algorithm (Philipps et al., 2021), a robust variant of the Newton-Raphson algorithm (Levenberg, 1944; Marquardt, 1963). This algorithm iteratively refines the parameter estimates θ until convergence is reached. Strict convergence criteria are used, based on the stability of both the parameters and the log-likelihood function, along with the relative distance to the maximum, calculated from the first and second derivatives of the log-likelihood. The convergence is achieved when $\frac{\nabla\left(\ell\left(\theta^{(l)}\right)\right)\left(H\left(\theta^{(l)}\right)\right)^{-1}\nabla\left(\ell\left(\theta^{(l)}\right)\right)}{m} < \varepsilon_d$ at an iteration l, with m the number of parameters, ε_d a predefined threshold, H the Hessian matrix and ∇ the gradient function.

The estimation is performed using the Cholesky decomposition of the covariance matrix of random effects to ensure positive-definite matrix. The inverse of the Hessian matrix is computed by finite difference to obtain the variances of the estimates. The variances of the estimated parameters of the covariance matrix in its natural scale are obtained using the Delta-Method. Let denote C the Cholesky matrix with l_{ij} its parameters and Σ the covariance matrix such as $\Sigma = CC^\top$ with σ_{ij} its factors. Given $Cov(\hat{l}_{ij}, \hat{l}_{km})$, $Cov(\hat{\sigma}_{ij}, \hat{\sigma}_{km})$ is equal to

$$\sum_{t=1}^{f(i,j)}\sum_{s=1}^{f(k,m)}\left[\hat{l}_{jt}\hat{l}_{ms}Cov(\hat{l}_{it}, \hat{l}_{ks}) + \hat{l}_{jt}\hat{l}_{ks}Cov(\hat{l}_{it}, \hat{l}_{ms}) + \hat{l}_{it}\hat{l}_{ms}Cov(\hat{l}_{jt}, \hat{l}_{ks}) + \hat{l}_{it}\hat{l}_{ks}Cov(\hat{l}_{jt}, \hat{l}_{ms})\right]$$

V.3.2.3 Initialisation

In each estimation function, it is necessary to provide a set of parameter initial value.

LSMM:
The first step is to estimate the considered LSMM. The initial values of the fixed effects are obtained by estimating a standard linear regression model. The cholesky of the covariance matrix Σ of random effects is initialised by the identity matrix. Then the estimation of the LSMM is performed using $S1$ (in general $S1 = 500$ or $S1 = 1000$) QMC draws until convergence is achieved.

LSJM:
The initial values of the location-scale mixed sub-model are the estimates obtained by fitting the LSMM alone. For the survival part, the values of the baseline risk function and the fixed covariates W are initialised using an estimation of a parametric survival model. Then, the association parameters, α_b and α_τ are intialised at 0. For the case of an illness-death model with interval censoring, the joint model is first estimated without handling interval cencoring, taking the middle of the interval as time of transition to state (1) and assuming that subjects who go to state (2) before being observed in state (1) directly made the transition (0) to (2).

V.3.2.4 Strategy

For both LSMM and LSJM models, the estimation can be performed with a relatively small number $S1$ of QMC draws (e.g. $S1 = 500$ or $S1 = 1000$) to obtain estimated regression parameters without bias. However, as we have previously shown (Courcoul et al., 2024b,a), the estimated variances of parameters could be biased. We have therefore proposed to perform a second step for precision improvement, then few additional iterations are performed with a higher number $S2 > S1$ of QMC draws (e.g. $S2 = 5000$ or $S2 = 10000$) until the Hessian matrix is invertible.

The choice of the number of QMC draws $S1$ and $S2$ has a notable effect on computational time. Therefore, for model selection, we suggest that users compare different models using the results of step 1 (using likelihood or information criteria) with a small value for $S1$ and perform step 2, which entails a larger number of QMC draws, solely for the final selected model. Finally, we recommend to increase the number of QMC draws with the number of random effects and in case of convergence issues.

V.4 Post-fit computations

The package includes a range of post-fit analyses and computations, most of which are shared between the different estimation functions. The upcoming subsections detail these post-fit computations.

V.4.1 Predictions

For linear mixed models and joint models empirical Bayes estimates of the random effects r_i can be obtained. For each subject the empirical Bayes estimates of the random effects, denoted by $\widetilde{r}_i = \text{argmax}_{r_i} f(r_i|Y_i, \widehat{\theta})$ for LSMM and $\widetilde{r}_i = \text{argmax}_{r_i} f(r_i|Y_i, T_i, \delta_i, \widehat{\theta})$ for LSJM correspond to the mode of their estimated conditional posterior given the data and estimated parameters. They are both computed with the Marquardt-Levenberg algorithm.

The package allows also to compute the individual predicted trajectory corresponding to the conditional expectation given the random effects, defined by:

$$\widehat{\mathbb{E}}(Y_i(t)|\widetilde{b}_i, \widetilde{\tau}_i) = X_i(t)\widehat{\beta} + Z_i(t)\widetilde{b}_i$$

For LSJM, the predicted cumulative hazard function can be computed for a grid of times, by plugging the empirical Bayes estimates of the random effects in the formula for the risk functions:

$$\widehat{\Lambda}_i^{kl}(t|\widetilde{r}_i; \widehat{\theta}) = \int_0^t \widehat{\lambda}_0^{kl}(u; \widehat{\theta}) \exp(W_i^{kl\top}\widehat{\gamma}^{kl} + g_y^{kl}(\widetilde{b}_i, u)^\top \widehat{\alpha}_b^{kl} + g_\tau^{kl}(\widetilde{\tau}_i, u)^\top \widehat{\alpha}_\tau^{kl})du \tag{V.14}$$

computed using the Gauss-Kronrod approximations.

V.4.2 Goodness-of-fit

V.4.2.1 Longitudinal fit

To assess the fit of the longitudinal sub-model, we compute the predicted marker values for each individual at their respective visit times using the empirical Bayes estimates. Then it is possible to display the comparison of the mean of the predicted marker values with the mean of the observed measurements grouping the observations by chosen time windows.

V.4.2.2 Survival fit

To evaluate the fit of the survival submodels, it is possible to provide a plot of the mean of the predicted cumulative hazard function for each transition versus the Nelson-Aalen estimator in case of competing risks or versus a non-parametric estimator obtained using the `SmoothHazard` package (Touraine et al., 2017).

V.4.3 Dynamic predictions

This section is specific to LSJM. Individual dynamic predictions can be computed and plotted. They are defined as the predicted probability of having event k between time s and $s+t$ given that the subject i did not experience any event before time s, and knowing all marker measures collected until time s, denoted by $\mathcal{Y}_i(s)$, and the set of estimated parameters. The prediction is defined for subject i by:

$$\begin{aligned}\pi_i^{0k}(s,t;\widehat{\theta}) &= P(s<T_i<s+t,\delta_i=k|T_i>s,\mathcal{Y}_i(s),\widehat{\theta}) \\ &= \frac{\int\left[\int_s^{s+t}\exp\left(-\sum_{c=1}^K\Lambda_i^{0c}(u|r_i,\widehat{\theta})\right)\lambda_i^{0k}(u|r_i,\widehat{\theta})du\right]f(\mathcal{Y}_i(s)|r_i,\widehat{\theta})f(r_i|\widehat{\theta})dr_i}{\int\exp\left(-\sum_{c=1}^K\Lambda_i^{0c}(s|r_i,\widehat{\theta})\right)f(\mathcal{Y}_i(s)|r_i,\widehat{\theta})f(r_i|\widehat{\theta})dr_i}\end{aligned} \tag{V.15}$$

For illness-death model and competing events, $K=2$ whereas for a model with a single event, $K=1$. For illness-death model and for a model with a single event, only the prediction from state (0) to state (1) can be computed whereas for competing risks both transition (0) to (1) and (0) to (2) can be predicted.
As in the estimation procedure, the integral over the random effects is computed by QMC approximation and the integral over time with the Gauss-Kronrod quadrature.
The 95% confidence interval of predictions is obtained by the following Monte Carlo algorithm. This can be quite time-consuming. For L large enough and $l=1,...,L$ ($L=1000$ for instance):

— Generate $\widetilde{\theta}^{(l)}\sim\mathcal{N}(\widehat{\theta},V(\widehat{\theta}))$ where $V(\widehat{\theta})$ is given by the inverse of the Hessian matrix at $\widehat{\theta}$;

— Compute $\widetilde{\pi}_i^{(l)}(s,t;\widetilde{\theta}^{(l)})$ from equation (V.15);

— Compute the 95% confidence interval from the 2.5th and 97.5th percentiles of the L-sample of $\widetilde{\pi}_i^{(l)}(s,t;\widetilde{\theta}^{(l)})$.

V.5 Implementation and Examples

The package allows to estimate the models presented in section V.3 and provide the output of post-fit computations detailed in section V.4. The two estimation functions, `lsmm` and `lsjm`, rely on programs written in `R` and `C++`. The description of the calls of these functions with some illustrations based on fictive data can be founded in Annexes VI.4. For users, it is worth noting that the data preparation format is illustrated using real data examples in the following section, depending on the estimated models.

V.6 Application on real data

This section presents some examples of the application of `LSJM` `R`-package functions, whether for model estimation, graphical output, or dynamic prediction. These examples are illustrated using the real `threeC` data supplied with our `R`-package.

After installing the package, the first step consists in loading package `LSJM`:

Code R

```
library("LSJM")
```

V.6.1 `threeC` data

The `threeC` dataset from `LSJM` `R`-package is a random subsample of 500 subjects, identified by `ID`, from the French Three-Cities cohort aimed at assessing the relation between vascular factors and dementia in the elderly (3C Study Group, 2003). The sample include participants without dementia at baseline and aged 65 years old or older. Repeated measures of systolic blood pressure (`SBP`) were collected over a maximum period of 20 years. At each visit, systolic blood pressure was measured two or three times. The age at entry in the cohort (`age0`), the age at the visit (`age.visit`), the age of last visit without dementia (`age.last`), the age of visit diagnosis (`age.first`), the age of death or censoring (`age.final`) and the indicator of dementia and death (`dem`, `death`) were also collected. The sex variable (`sex`) is also reported (187 men and 313 women).
For computation and interpretation purposes, all `age` variables will be replaced by the age minus 65 divided by 10. The values of blood pressure are also divided by 10 which leads to a measurement unit in cmHg instead of mmHg.
The following lines create the variables relative to the new `age`, and display the first lines of

the `threeC` dataset:

Code R

```
data(threeC)
threeC$age.visit65 <- (threeC$age.visit-65)/10
threeC$age.final65 <- (threeC$age.final-65)/10
threeC$age0_65 <- (threeC$age0-65)/10
threeC$age.last65 <- (threeC$age.last-65)/10
threeC$age.first65 <- (threeC$age.first-65)/10
threeC$SBP <- threeC$SBP/10
```

Output R

```
head(threeC)

    ID SBP age.visit age.final     age0 age.last age.first   sex dem
10003 14.2  72.17522  90.44216 72.17522 89.85352  89.85352 Woman   0
10003 13.4  74.57632  90.44216 72.17522 89.85352  89.85352 Woman   0
10003 12.8  74.57632  90.44216 72.17522 89.85352  89.85352 Woman   0
10003 14.2  76.56400  90.44216 72.17522 89.85352  89.85352 Woman   0
10003 12.0  76.56400  90.44216 72.17522 89.85352  89.85352 Woman   0
10003 11.5  76.56400  90.44216 72.17522 89.85352  89.85352 Woman   0
death num.visit age.final65 age.visit65    age0_65 age.last65 age.first65
    1    V0    2.544216   0.7175222 0.7175222   2.485352    2.485352
    1    V1    2.544216   0.9576318 0.7175222   2.485352    2.485352
    1    V1    2.544216   0.9576318 0.7175222   2.485352    2.485352
    1    V2    2.544216   1.1563997 0.7175222   2.485352    2.485352
    1    V2    2.544216   1.1563997 0.7175222   2.485352    2.485352
    1    V2    2.544216   1.1563997 0.7175222   2.485352    2.485352
```

V.6.2 A LSJM model with time-dependent variability and an illness-death model

This subsection illustrates the estimation of a LSJM. The LSMM sub-model is estimated suggesting a covariate and time-dependent variability and the survival sub-model is an illness-death model.

V.6.2.1 Data management

In this example, we consider only one measure at each visit. So we create first a new variable: the mean of all measures collected at the corresponding visit, `SBPmean`. Then, we delete the measurements of blood pressure collected after the diagnosis time of dementia. The following lines create the new dataset and display its first lines:

Code R

```
library(dplyr)
threeC_ex1 <- threeC
threeC_ex1 <- threeC_ex1 %>% group_by(ID, num.visit) %>% mutate(SBPvisit =
                                                                mean(SBP))
threeC_ex1 <- threeC_ex1[!duplicated(threeC_ex1[, c("ID", "num.visit")]),
                         c("ID", "SBPvisit", "age.visit65", "age0_65",
                         "age.final65","age.last65",  "age.first65", "dem",
                         "death", "sex")]
```

Output R

```
head(threeC_ex1)
```

ID	SBPvisit	age.visit65	age0_65	age.final65	age.last65	age.first65	dem	death	sex
10003	14.2	0.718	0.718	2.54	2.49	2.49	0	1	woman
10003	13.1	0.958	0.718	2.54	2.49	2.49	0	1	woman
10003	12.6	1.16	0.718	2.54	2.49	2.49	0	1	woman
10003	11.7	1.48	0.718	2.54	2.49	2.49	0	1	woman
10003	12.6	1.73	0.718	2.54	2.49	2.49	0	1	woman
10003	10.8	2.00	0.718	2.54	2.49	2.49	0	1	woman
10003	12.4	2.19	0.718	2.54	2.49	2.49	0	1	woman
10003	12.7	2.49	0.718	2.54	2.49	2.49	0	1	woman
10005	14	0.975	0.975	2.07	1.21	2.02	0	1	woman
10005	12.3	1.21	0.975	2.07	1.21	2.02	0	1	woman

```
0     1  woman
0     1  woman
0     1  woman
1     1  woman
1     1  woman
```

Here, subject 1 is a woman, enter in the study at 0.718, was not diagnosed with dementia and died at time 2.54, whereas individual 2 was diagnosed with dementia at 2.02 and died at 2.07.

V.6.2.2 LSMM with time-dependent variability

The first step is to fit a LSMM using the `lsmm` function with linear trajectories of `SBP` and its residual variability with `age.visit`, assuming uncorrelated random effects between the random effects of the mean and the one of the variance. The next lines estimates the corresponding model:

Code R

```
m1 <- lsmm(formFixed = SBPvisit ~ age.visit65,
              formRandom = ~ age.visit65,
              formGroup = ~ ID,
              timeVar = 'age.visit65',
              data.long = threeC_ex1,
              formVar = "cov-dependent",
              formFixedVar = ~ age.visit65,
              formRandomVar = ~ age.visit65,
              correlated_re = FALSE,
              S1 = 500,
              S2 = 1000,
              nproc = 2)
```

Output R

```
summary(m1)
Location-scale linear mixed model fitted by maximum likelihood method

Statistical Model:
    Number of subjects: 500
```

```
    Number of observations: 2458

Iteration process:
    Convergence criteria satisfied
     Number of iterations:
          Step 1:  17
          Step 2:  1
     Convergence criteria (Step1): parameters = 4.93e-09
                                 : likelihood = 2.92e-07
                                 : second derivatives = 6.4e-14
     Time of computation : 1.145953 mins

Goodness-of-fit statistics:
    Likelihood:  -5027.3
    AIC:  10074.599

Maximum Likelihood Estimates:
Longitudinal model:
      Fixed effects of the location part:
               Coeff     SE     Wald Pvalue
(Intercept) 13.8661 0.1387 100.0068 <0.001
age.visit65  0.2868 0.0836   3.4324 <0.001

     Fixed effects of the scale part:
              Coeff     SE   Wald Pvalue
(Intercept) 0.2262 0.0555 4.0739 <0.001
age.visit65 0.1183 0.0331 3.5699 <0.001

     Covariance matrix of the location random effects:
             (Intercept) age.visit65
(Intercept)    4.283720   -1.757586
age.visit65   -1.757586    1.157830

     Covariance matrix of the scale random effects:
             (Intercept)  age.visit65
```

```
(Intercept)  0.08713937 -0.016774493
age.visit65 -0.01677449  0.007028832
```

First, the `summary()` method details the dataset, the number of subjects and the number of observations. The next block provides information about the convergence process with the number of iterations for both step 1 and 2, the convergence criteria of step 1, the time of computation and whether the model converged correctly ("Convergence criteria satisfied"). Then some goodness-of-fit statistics are given: the maximum log-likelihood and the Akaike information criteria (AIC). Finally, tables of estimates are provided including the estimated parameter (`Coeff`), the estimated standard error (`SE`), the Wald test statistics with the normal approximation (`Wald`) and the corresponding p-value (`Pvalue`), for both the fixed effects associated with the mean (loaction part) and the fixed effects associated with the residual variability (scale part). For the random effect distributions, we distinguish two estimated covariance matrices, one for the location part and one for the scale part of the model, since these two parts have been assumed to be uncorrelated (`correlated_re = FALSE`). If `correlated_re = TRUE`, a single covariance matrix is returned for all the random effects in the model.

All estimated LSMM parameters and their standard errors are available in the `table.res` object of the list returned by the `lsmm` function. This concerns the fixed parameters, the parameters of the Cholesky decomposition of the random effects covariance matrix, as well as the parameters of the latter obtained via the Delta-method. The prediction of the random effects and the predicted value of the marker for each individual at their respective measurement times are computed using the `ranef()` or `predict()` functions and the trajectory of the marker can be displayed by the `plot()` function. Since the same functions can be applied to LSJM models, we will detail their usage in the following section.

V.6.2.3 LSJM for semi-competing events and interval censoring

Then `lsjm()` function allows to estimate the illness-death joint model, assuming baseline hazard functions defined using a Weibull function for transitions (01) and (02) and splines with one internal knot for transition (12). Data are left truncated by the age of entry in the study, `age0_65`. All transitions are adjusted on the current value and the variability of the marker. The transition from healthy to death is also adjusted on the slope. And finally, both transitions to death are adjusted on the sex. The next lines estimate the corresponding model:

Code R

```
l1 <- lsjm(m1,
           survival_type = 'IDM',
           formSurv_01=~1,
           formSurv_02=~sex,
           formSurv_12=~sex,
           sharedtype_01 = c("value", "variability"),
           sharedtype_02 = c("value", "slope", "variability"),
           sharedtype_12 = c("value", "variability"),
           hazardBase_01 = "Weibull",
           hazardBase_02 = "Weibull",
           hazardBase_12 = "Splines",
           delta1=~dem,
           delta2=~death,
           Time_T =~age.final65,
           Time_L =~age.last65,
           Time_R =~age.first65,
           Time_T0 =~age0_65,
           formSlopeFixed =~1,
           formSlopeRandom =~1,
           index_beta_slope = c(2),
           index_b_slope = c(2),
           nb.knots.splines = c(0,0,1),
           S1 = 1000,
           S2 = 2000,
           nproc = 10)
```

Output R

```
summary(l1)
Location-scale joint model for an illness-death model
fitted by maximum likelihood method

Statistical Model:
    Number of subjects:  500
```

```
    Number of observations:  2458

Iteration process:
    Convergence criteria satisfied
     Number of iterations:
          Step 1:  20
          Step 2:  1
     Convergence criteria (Step1): parameters = 4.47e-05
                                 : likelihood = 9.44e-06
                                 : second derivatives = 1.76e-10
     Time of computation : 6.874512 hours

Goodness-of-fit statistics:
    Likelihood:  -5883.375
    AIC:  11822.751

Maximum Likelihood Estimates:
Longitudinal model:
      Fixed effects of the location part:
               Coeff     SE    Wald Pvalue
(Intercept) 13.9551 0.1566 89.1354 <0.001
age.visit65  0.2131 0.1083  1.9680  0.049

     Fixed effects of the scale part:
              Coeff     SE   Wald Pvalue
(Intercept) 0.1537 0.0538 2.8568  0.004
age.visit65 0.1791 0.0339 5.2860 <0.001

     Covariance matrix of the location random effects:
            (Intercept) age.visit65
(Intercept)    4.366972   -1.810359
age.visit65   -1.810359    1.193415

     Covariance matrix of the scale random effects:
             (Intercept)   age.visit65
```

```
(Intercept)  0.065553512 -0.0037253869
age.visit65 -0.003725387  0.0004496534

Survival models:
    Transition 0-1:
       Regression:
                  Coeff     SE    Wald Pvalue
value 01       -0.0342 0.0933 -0.3669  0.714
variability 01  1.6529 0.7477  2.2108  0.027

     Baseline:  Weibull

            Coeff     SE    Wald Pvalue
intercept -6.7827 1.3252 -5.1184 <0.001
shape_01   2.0655 0.0954 21.6561 <0.001

    Transition 0-2:
       Regression:
                  Coeff     SE    Wald Pvalue
value 02       -0.3480 0.1459 -2.3846  0.017
slope 02       -0.0877 0.2257 -0.3885  0.698
variability 02  2.1693 0.9036  2.4007  0.016
sexWoman_02    -0.9835 0.2240 -4.3909 <0.001

     Baseline:  Weibull

            Coeff     SE    Wald Pvalue
intercept -2.0109 2.0339 -0.9887  0.323
shape_02   1.8216 0.0992 18.3585 <0.001


    Transition 1-2:
       Regression:
                  Coeff     SE    Wald Pvalue
value 12        0.1028 0.0719  1.4301  0.153
```

```
variability 12 -0.1229 0.4026 -0.3052   0.76
sexWoman_12    -0.7427 0.1907 -3.8939 <0.001

     Baseline:  Splines

              Coeff     SE    Wald Pvalue
splines12_1 -6.2992 7.5075 -0.8390  0.401
splines12_2 -0.7688 2.7564 -0.2789  0.780
splines12_3 -0.3364 2.3142 -0.1454  0.884
splines12_4  0.5493 1.7673  0.3108  0.756
splines12_5  4.1182 4.1341  0.9961  0.319
```

Similarly to the LSMM, the first part of the `summary()` method provides details about the dataset, the convergence process and the results of the longitudinal submodel. For this example, the convergence criteria where satisfied in 20 iterations for the first step and 1 for the second. There are 500 subjects and 2458 obervations. The AIC is equal to 11822.751. According to the table on the fixed effects of the scale part, blood pressure residual variability increases with age. The last part of the summary gives the estimated values of the parameters associated with each transition functions: the regression parameters and the baseline parameters. First, for the transition to dementia, it appears that dementia occurrence seems to be not associated with the current value of blood pressure but the risk increases with the variability of blood pressure. Then, the risk of death without dementia decreases with the current value of blood pressure but increases with the variability. This risk is also lower for women. Finally, the risk of death with dementia seems to be associated only with the sex. The estimated Cholesky parameters, their standard errors and the standard errors of the covariance matrices obtained with the Delta-Method are given by `l1$table.res`.

Then, `predict()` function allows to compute the empirical Bayes estimates of the random effects, the predicted values for each individual at their respective visiting times and the cumulative risks for each considered transitions:

Code R

```
predictl1 <- predict(l1, which = c('RE','Y','Cum'))
```

Output R

```
head(predictl1$predictRE)
 id    (Intercept)_Location age.visit65_Location (Intercept)_Scale
```

```
10005           -0.7218100            0.1361275        0.05100922
10026            1.3343421            0.2648115        0.22039363
10060            1.0049743           -0.4892723        0.20685208
10091            1.7085881           -0.8189352       -0.16157812
10098            0.2941827           -0.8438771       -0.02399378
10152           -1.9701846            0.4015402        0.03410410
   age.visit65_Scale
        -0.002302383
        -0.010803398
        -0.010553859
         0.006838082
         0.001073053
        -0.001663643

head(predictl1$predictY)
   id      time    predY   predSD
10005 0.9754278 13.57392 1.458094
10005 1.2138946 13.65720 1.520873
10026 1.5854209 16.04713 1.898126
10026 1.8156742 16.15717 1.973113
10026 2.0160849 16.25296 2.040788
10060 0.4451061 14.83712 1.545819
```

The goodness-of-fit of the longitudinal part can be assess by displaying the mean of marker predictions at each visit times and the mean of the observed measurements as in Figure V.6. Then, Figure V.7 presents the predicted trajectory and its confidence interval for two selected individual. The following lines provide the code to create these graphics:

Code R

```
plot(l1, which = "traj.fit", Objectpredict = predictl1,
                                    break.times = (seq(65,95,by = 2.5)-65)/10)
plot(l1, which = "traj.ind", Objectpredict = predictl1,
                                                    ID.ind = c(10003,120010))
```

For the survival part, the mean of the predicted cumulative hazard function should be compared for each transition to a non-parametric estimator obtained using the `SmoothHazard` package (Touraine et al., 2017):

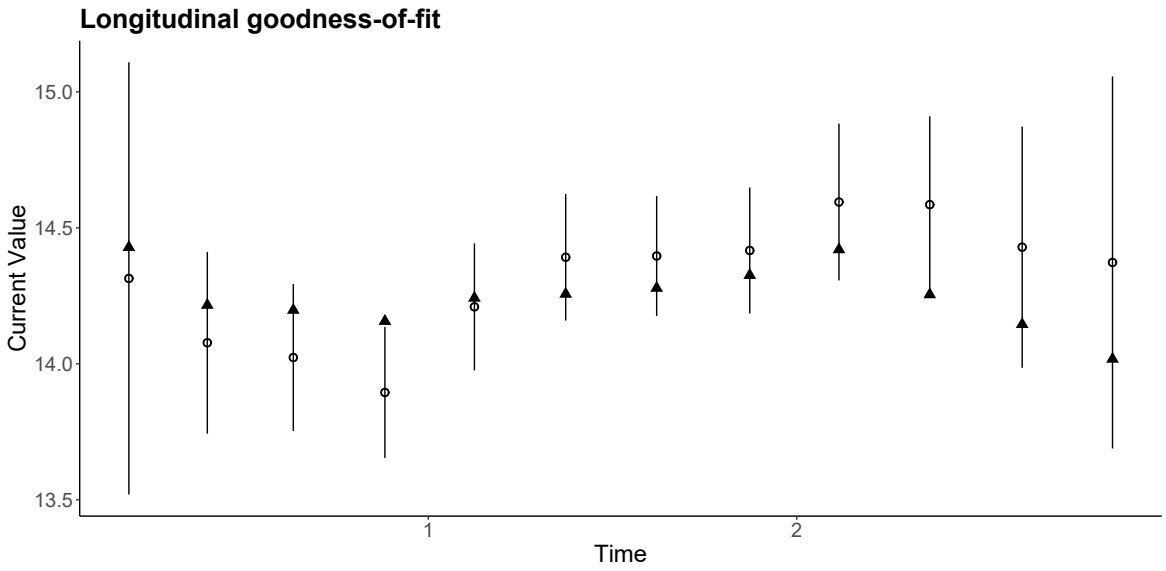

Figure V.6 – Comparison between predicted value of the marker from the joint model (black triangles) and the observations (mean in white circles with 95% confidence interval (by 2.5-year age intervals).

Code R

```
library(SmoothHazard)
threeC_ex1.id <- threeC_ex1[!duplicated(threeC_ex1$ID),]

sm1 <- idm(formula02=Hist(time=age.final65,event=death,entry=age0_65)~1,
           formula01=Hist(time=list(age.last65,age.first65),event=dem,
           entry = age0_65)~1,
           formula12=Hist(time=age.final65,event=death,entry=age0_65)~1,
           method="Splines", CV = 1, n.knots = c(15,15,15),
           kappa = c(10, 10, 2), data=threeC_ex1.id)

plot(l1, which = 'survival.fit', Objectpredict = predictl1,
                                          ObjectSmoothHazard = sm1)
```

Figure V.8 presents this assessment. Here, the adjustment between the model and the data could be improved.

Finally, the `dynpred()` function allows to compute the individual dynamic prediction of an event. As described in section V.4.3, for a specific subject (e.g a new one), whose data are provided in the input, the probabilities of occurrence of the first event from a landmark time

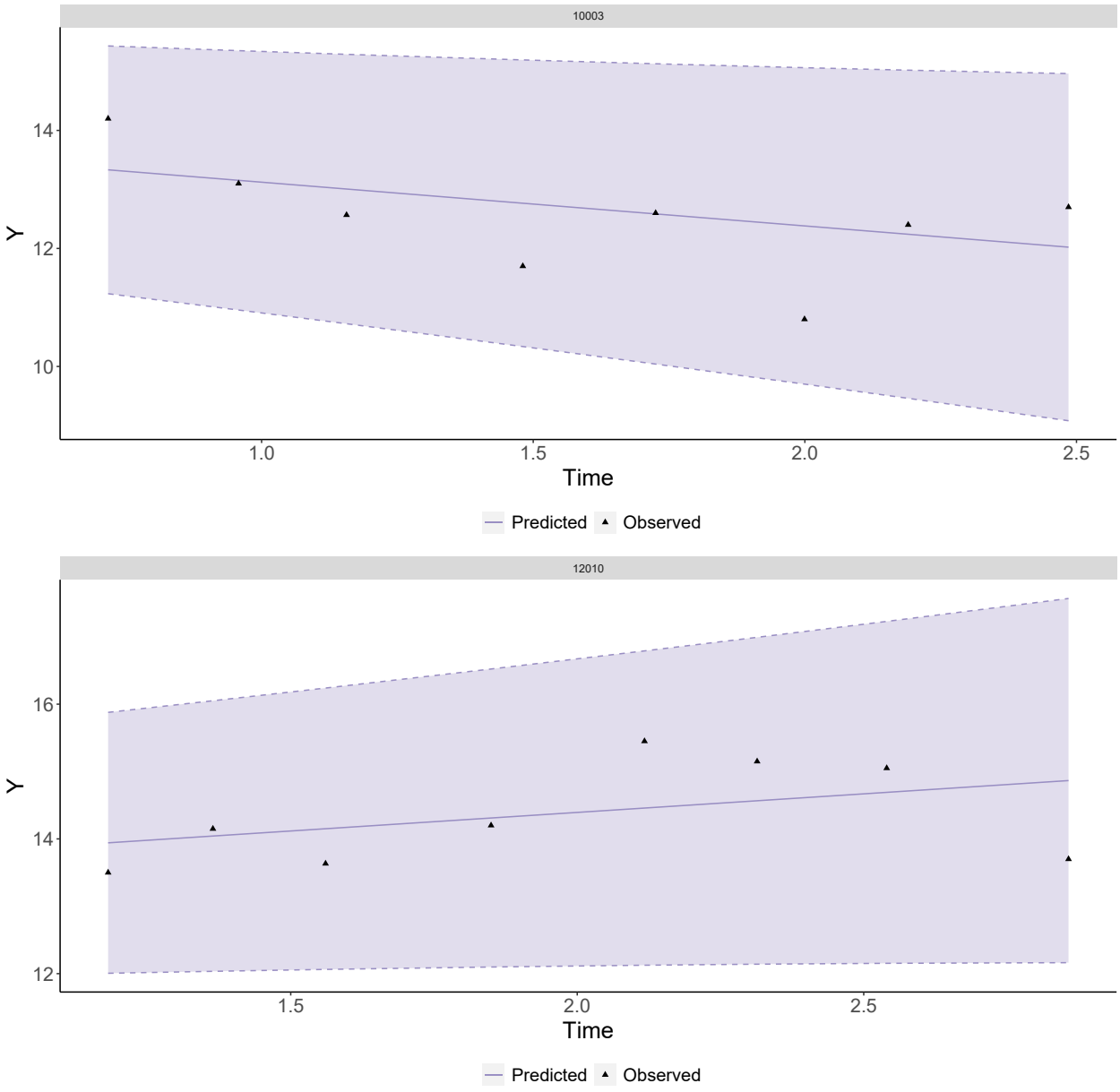

Figure V.7 – Prediction over time of the individual blood pressure and its prediction interval at 95% for two subjects. The black triangles are the observed measurements.

indicated in `s` at horizons indicated in `horizon` are computed using the estimated model and the longitudinal information of the new subject up to the landmark time `s`. The following example is for a subject of the estimation dataset but it could be perform for a new subject. We compute the probability that dementia occurs between landmark time 2 and time 3. Here we computed also the interval confidence of this prediction. Figure V.9 illustrates the result and the following lines provide the code for one individual :

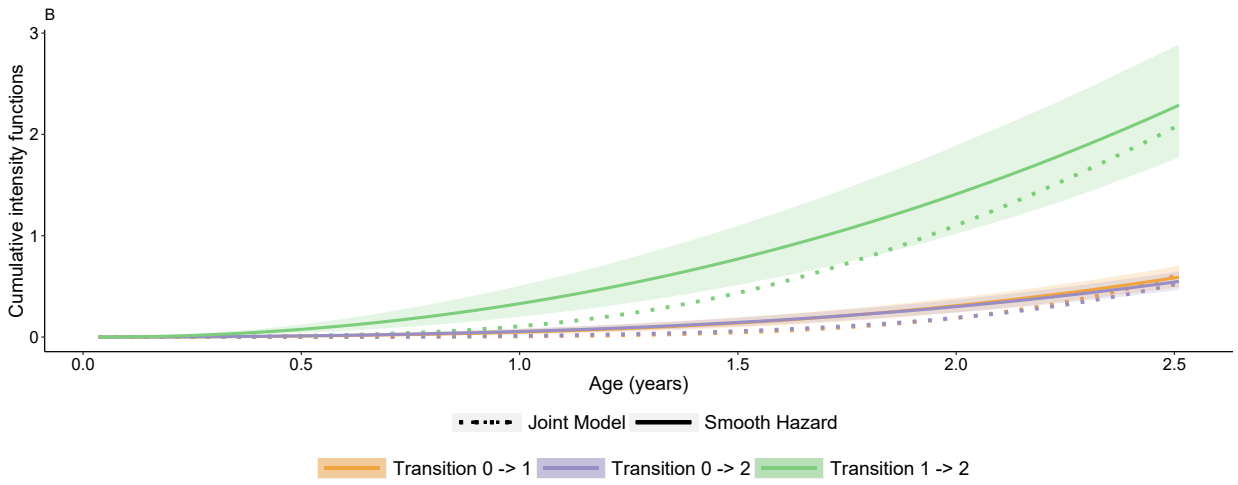

Figure V.8 – Comparison between the predicted cumulative hazard function from the joint model for each transition and a non-parametric estimator with 95% confidence interval (Illness-death model estimated by penalized likelihood accounting for interval censoring with the R package `SmoothHazard`).

Code R

```
ind3 <- threeC_ex1[which(threeC_ex1$ID == 10003),]
dynp <- dynpred(ind1, l1, s = 2, horizon = seq(2.1,3,0.1),
        nb.draws = 1000)
```

Output R

```
dynp
   ID Time Prediction     Median      ICinf      ICsup Empirical SD
10003  2.1 0.03668934 0.03711915 0.02341961 0.05515911  0.008347391
10003  2.2 0.07454292 0.07489770 0.04803115 0.11390105  0.016467563
10003  2.3 0.11238074 0.11093355 0.07166993 0.16920983  0.025350245
10003  2.4 0.14898390 0.14924575 0.09761294 0.22580855  0.033167628
10003  2.5 0.18321395 0.18236693 0.11472371 0.27167955  0.040913491
10003  2.6 0.21412335 0.21263945 0.12935679 0.33223693  0.049535750
10003  2.7 0.24103969 0.23474732 0.14684326 0.35918877  0.054841381
10003  2.8 0.26361194 0.25857337 0.15603418 0.39115197  0.061001419
10003  2.9 0.28181406 0.28113784 0.16466651 0.42909889  0.067716372
10003  3.0 0.29590811 0.28717605 0.17564340 0.45149581  0.070177043
```

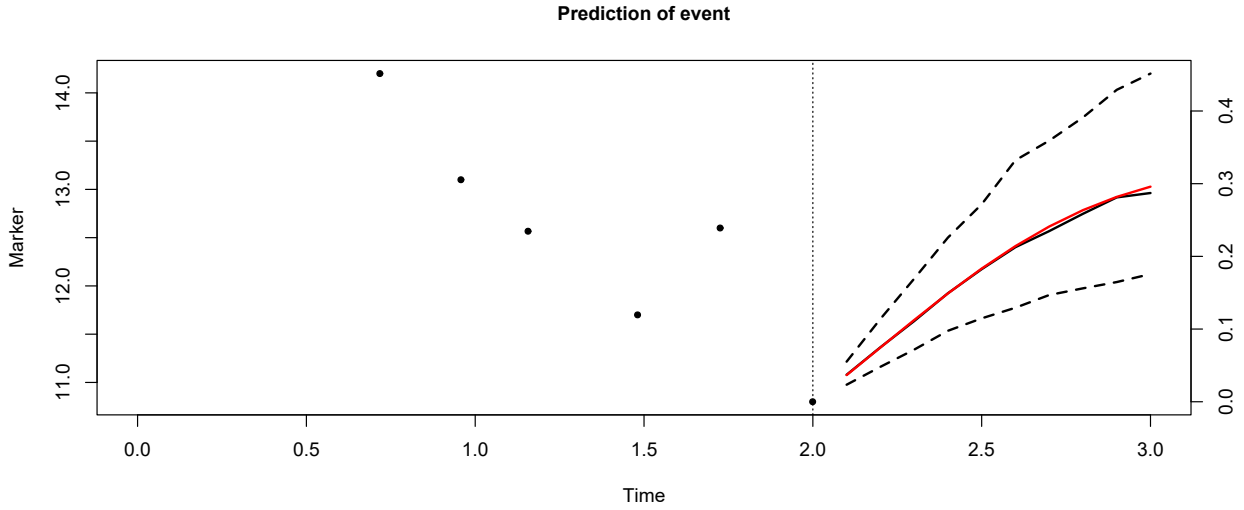

Figure V.9 – Prediction of the risks of Dementia between 2 and 3 years (with its 95% confidence interval indicated by dashed lines and the median by black line), for a patient at risk at 2 years.

V.6.3 A LSJM distinguishing inter and intra-visit variabilities for competing events

This second example illustrates the estimation of a LSJM combining within and between visits variabilities and competing events by studying blood pressure trajectory and the risk of dementia and death.

V.6.3.1 Data management

As we consider dementia and death as competing risk without handling interval censoring, we suppose as known the exact age of dementia onset taking the middle of interval between the last visit without dementia and the visit of diagnosis [1]. If an individual is diagnosed with dementia, he is then censored for death and conversely. Then, as the example distinguishes within and between visits variabilities, all measurements are kept, excepted ones collected after the time of the event. The following lines create the new dataset and display its first lines:

Code R

```
threeC_ex2 <- threeC[,c("ID", "SBP", "age.visit65", "age0_65",
```

1. this corresponds to the naive method discussed in section V.3.2.3 to present the parameter initialization of the illness-death model dealing with interval censoring;

```
                        "age.final65", "age.last65", "age.first65",
                        "dem", "death", "sex","num.visit")]

threeC_ex2$age65_CR <- NA
threeC_ex2$age65_CR[which(threeC_ex2$dem == 1)] <-
  (threeC_ex2$age.last65[which(threeC_ex2$dem == 1)] +
     threeC_ex2$age.first65[which(threeC_ex2$dem == 1)])/2
threeC_ex2$age65_CR[which(threeC_ex2$dem == 0)] <-
  threeC_ex2$age.final65[which(threeC_ex2$dem == 0)]

threeC_ex2$demCR <- threeC_ex2$dem
threeC_ex2$deathCR <- NA
threeC_ex2$deathCR[which(threeC_ex2$dem == 1)] <- 0
threeC_ex2$deathCR[which(threeC_ex2$dem == 0)] <-
  threeC_ex2$death[which(threeC_ex2$dem == 0)]

threeC_ex2 <- threeC_ex2 %>% group_by(ID) %>% filter(age.visit65 <= age65_CR)

threeC_ex2 <- threeC_ex2[,c("ID","SBP","age.visit65","num.visit","age0_65",
                            "demCR","deathCR","age65_CR","sex")]
```

Output R

```
head(threeC_ex2)
   ID  SBP age.visit65 num.visit   age0_65 demCR deathCR age65_CR   sex
10003  14.2   0.7175222        V0 0.7175222     0       1 2.544216 Woman
10003  13.4   0.9576318        V1 0.7175222     0       1 2.544216 Woman
10003  12.8   0.9576318        V1 0.7175222     0       1 2.544216 Woman
10003  14.2   1.1563997        V2 0.7175222     0       1 2.544216 Woman
10003  12.0   1.1563997        V2 0.7175222     0       1 2.544216 Woman
10003  11.5   1.1563997        V2 0.7175222     0       1 2.544216 Woman
```

Here the individual is considered as dead without dementia at 2.54.

V.6.3.2 LSMM with inter and intra-visit variabilities

As is the first example, we run the LSMM to initialize the parameters corresponding to the longitudinal sub-model:

Code R

```
m2 <- lsmm(formFixed = SBP ~ age.visit65,
           formRandom = ~ age.visit65,
           formGroup = ~ ID,
           timeVar = 'age.visit65',
           data.long = threeC_ex2,
           formVar = "inter-intra",
           random_inter = T,
           random_intra = T,
           formGroupVisit = ~num.visit,
           correlated_re = FALSE,
           S1 = 500,
           S2 = 1000,
           nproc = 4)
```

Output R

```
summary(m2)
Location-scale linear mixed model fitted by maximum likelihood method

Statistical Model:
    Number of subjects: 500
    Number of observations: 4857

Iteration process:
    Convergence criteria satisfied
     Number of iterations:
          Step 1:  11
          Step 2:  1
     Convergence criteria (Step1): parameters = 1.44e-06
                                 : likelihood = 4.46e-05
                                 : second derivatives = 1.32e-10
     Time of computation : 1.427672 mins

Goodness-of-fit statistics:
    Likelihood:  -8635.61
```

```
    AIC:  17291.22

Maximum Likelihood Estimates:
Longitudinal model:
      Fixed effects of the location part:
               Coeff     SE    Wald Pvalue
(Intercept) 13.7152 0.1402 97.8547 <0.001
age.visit65  0.3927 0.0874  4.4904 <0.001

     Fixed intercept of the scale part(inter/intra variabilities):
        Coeff     SE     Wald Pvalue
inter  0.2931 0.0275  10.6572 <0.001
intra -0.2216 0.0182 -12.1901 <0.001



     Covariance matrix of the random effects of the mean:
             (Intercept) age.visit65
(Intercept)    3.843600   -1.575881
age.visit65   -1.575881    1.186363

     Covariance matrix of the random effects of the variance:
            inter      intra
inter 0.06036374 0.02102480
intra 0.02102480 0.04600852
```

As in the previous example (section V.6.2.2), the `summary()` function provides details about the dataset, the convergence process, and the tables of estimates. The `ranef()` and `predict()` functions computes the predicted random effects and for the second one, it also provides the predicted value of the marker.

V.6.3.3 LSJM for two competing events

The following lines provide the code to estimate the joint model with competing risks and using the location-scale linear mixed model estimated previously. We assume Weibull functions for both baseline hazard functions. Both risks are adjusted on the sex, the current value of the marker and the between visits variability. The risk of death is also associated on the within visit variability. Data are left truncated by the age of entry in the study `age0_65`:

Code R

```
l2 <- lsjm(Objectlsmm = m2,
           survival_type = 'CR',
           formSurv_01 = ~ sex,
           formSurv_02 = ~ sex,
           sharedtype_01 = c("value", "variability inter"),
           sharedtype_02 = c("value", "variability inter",
                                      "variability intra"),
           hazardBase_01 = "Weibull",
           hazardBase_02 = "Weibull",
           delta1 = ~ demCR,
           delta2 = ~ deathCR,
           Time_T0 = ~ age0_65,
           Time_T = ~ age65_CR,
           nproc = 5,
           S1 = 1000,
           S2 = 2000)
```

Output R

```
summary(l2)
Location-scale joint model with competing events
fitted by maximum likelihood method

Statistical Model:
    Number of subjects:  500
    Number of observations:  4857

Iteration process:
    Convergence criteria satisfied
     Number of iterations:
          Step 1:  18
          Step 2:  1
     Convergence criteria (Step1): parameters = 3.25e-08
                                 : likelihood = 9.52e-09
                                 : second derivatives = 1.44e-14
```

```
     Time of computation : 1.526567 hours

Goodness-of-fit statistics:
    Likelihood:  -9337.714
    AIC:  18717.428

Maximum Likelihood Estimates:
Longitudinal model:
      Fixed effects of the location part:
               Coeff     SE    Wald Pvalue
(Intercept) 13.7715 0.1438 95.7902 <0.001
age.visit65  0.3411 0.0918  3.7149 <0.001

     Fixed intercept of the scale part(inter/intra variabilities):
        Coeff     SE     Wald Pvalue
inter  0.3192 0.0286  11.1444 <0.001
intra -0.2185 0.0187 -11.6750 <0.001



     Covariance matrix of the random effects of the mean:
             (Intercept) age.visit65
(Intercept)    3.912766   -1.696949
age.visit65   -1.696949    1.267781

     Covariance matrix of the random effects of the variance:
           inter      intra
inter 0.06732289 0.01799515
intra 0.01799515 0.04611648

Survival models:
    Transition 0-1:
       Regression:
                         Coeff     SE    Wald Pvalue
value 01              -0.1127 0.0915 -1.2313  0.218
variability inter 01   1.7386 0.7074  2.4577  0.014
```

```
sexWoman_01          -0.1600 0.2100 -0.7617  0.446

     Baseline:  Weibull

            Coeff     SE    Wald Pvalue
intercept -4.5011 1.4080 -3.1968  0.001
shape_01   1.6786 0.1091 15.3870 <0.001

    Transition 0-2:
       Regression:
                        Coeff     SE    Wald Pvalue
value 02             -0.0826 0.0512 -1.6143  0.106
variability inter 02  0.9763 0.4540  2.1507  0.032
variability intra 02 -0.6852 0.7087 -0.9669  0.334
sexWoman_02          -0.7785 0.1304 -5.9702 <0.001

     Baseline:  Weibull

            Coeff     SE    Wald Pvalue
intercept -3.1455 0.8768 -3.5875 <0.001
shape_02   1.9826 0.0674 29.3976 <0.001
```

Both risks of dementia and death increase with the inter-visit variability of blood pressure but not with the current value or the intra-visit variability.

Then, using the same functions than in the previous example, `predict()` function allows to compute the predicted random effects, values of the marker and cumulative functions and `plot()` functions displays the Figures V.10, V.11, V.12 to assess the goodness-of-fit of the model:

Code R

```
predictl2 <- predict(l2, which = c('RE','Y','Cum'))

plot(l2, which = "traj.fit", Objectpredict = predictl2,
                                  break.times = (seq(65,95,by = 2.5)-65)/10)
plot(l2, which = "traj.ind", Objectpredict = predictl2,
```

```
                                                    ID.ind = c(10003,120010))
plot(l2, which = "survival.fit", Objectpredict = predictl2)
```

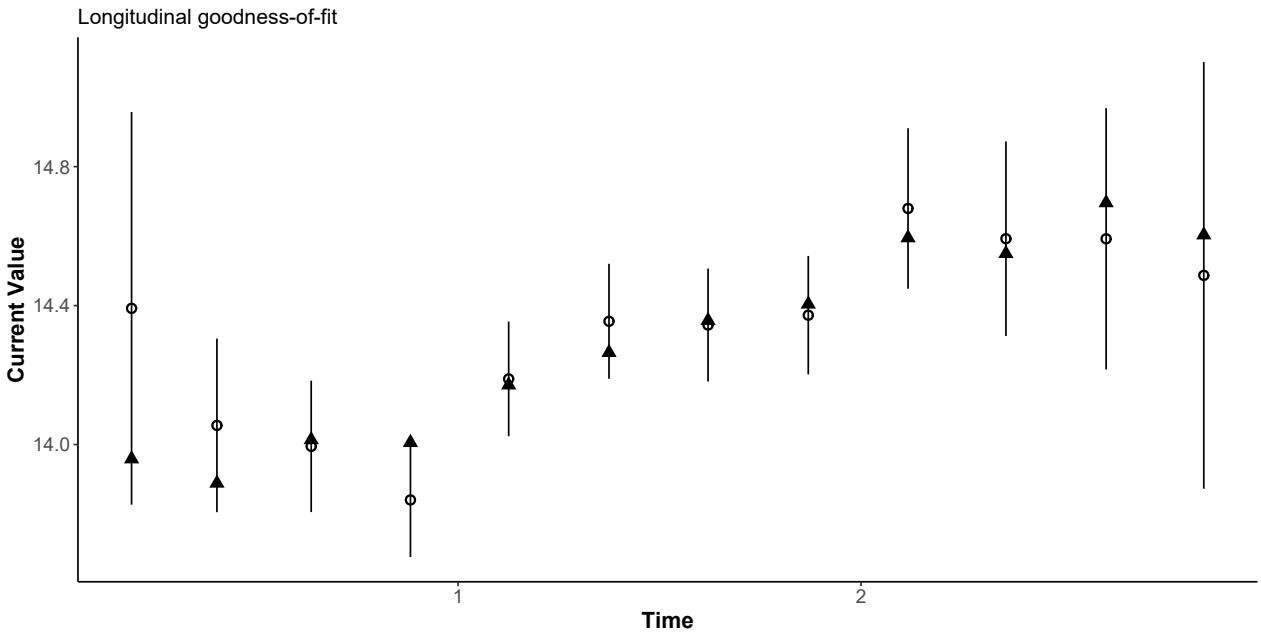

Figure V.10 – Comparison between predicted value of the marker from the joint model (black triangles) and the observations (mean in white circles with 95% confidence interval (by 2.5-year age intervals).

Finally, for a competing risk LSJM, dynamic predictions could be computed for both risks:

Code R

```
ind3 <- threeC_ex2[which(threeC_ex2$ID == 10003),]
dynpDementia <- dynpred(ind1, l1, s = 2, horizon = seq(2.1,3,0.1),
                                                event = 1, nb.draws = 1000)


dynpDeath <- dynpred(ind1, l1, s = 2, horizon = seq(2.1,3,0.1),
                                                event = 2, nb.draws = 1000)
```

Figures V.13 display the computed prediction of dementia and death between 2 and 3 years.

V.7 Discussion

The LSJM package introduces a versatile set of functions designed to estimate new location-scale linear mixed models incorporating various definition of residual variability,

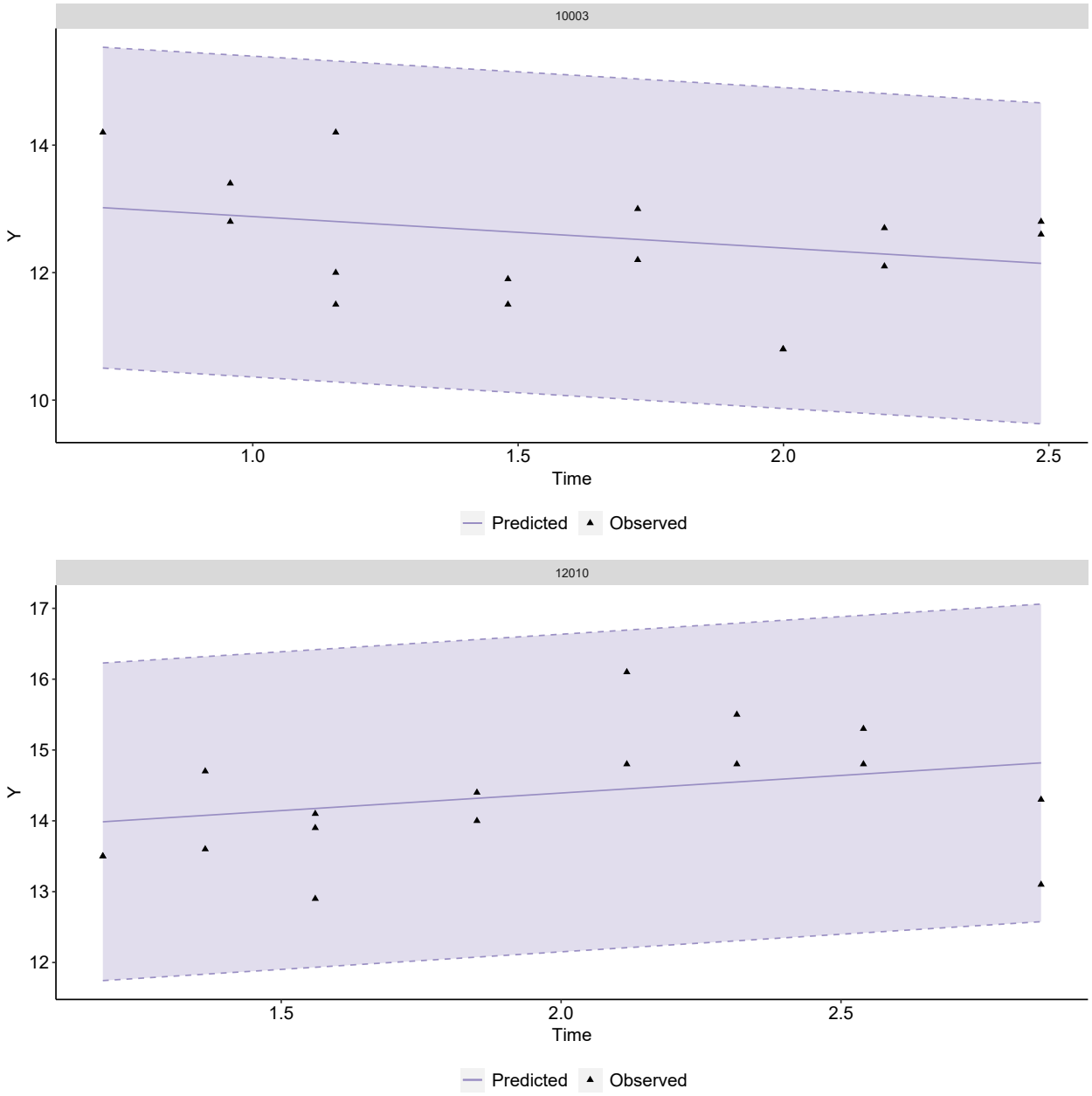

Figure V.11 – Prediction over time of the individual blood pressure and its prediction interval at 95% for two subjects. The black triangles are the observed measurements.

whether it is time-dependent or whether it can be divided into two components, one short-term and one long-term. To our knowledge, no other `R` package allows all these different fits of the residual variability. It allows also to adjust LSJM with shared random effects combining a LSMM and a complex survival model. The first particularity is the ability to adjust the risk of events based on the individual variability of the marker, in addition to the current

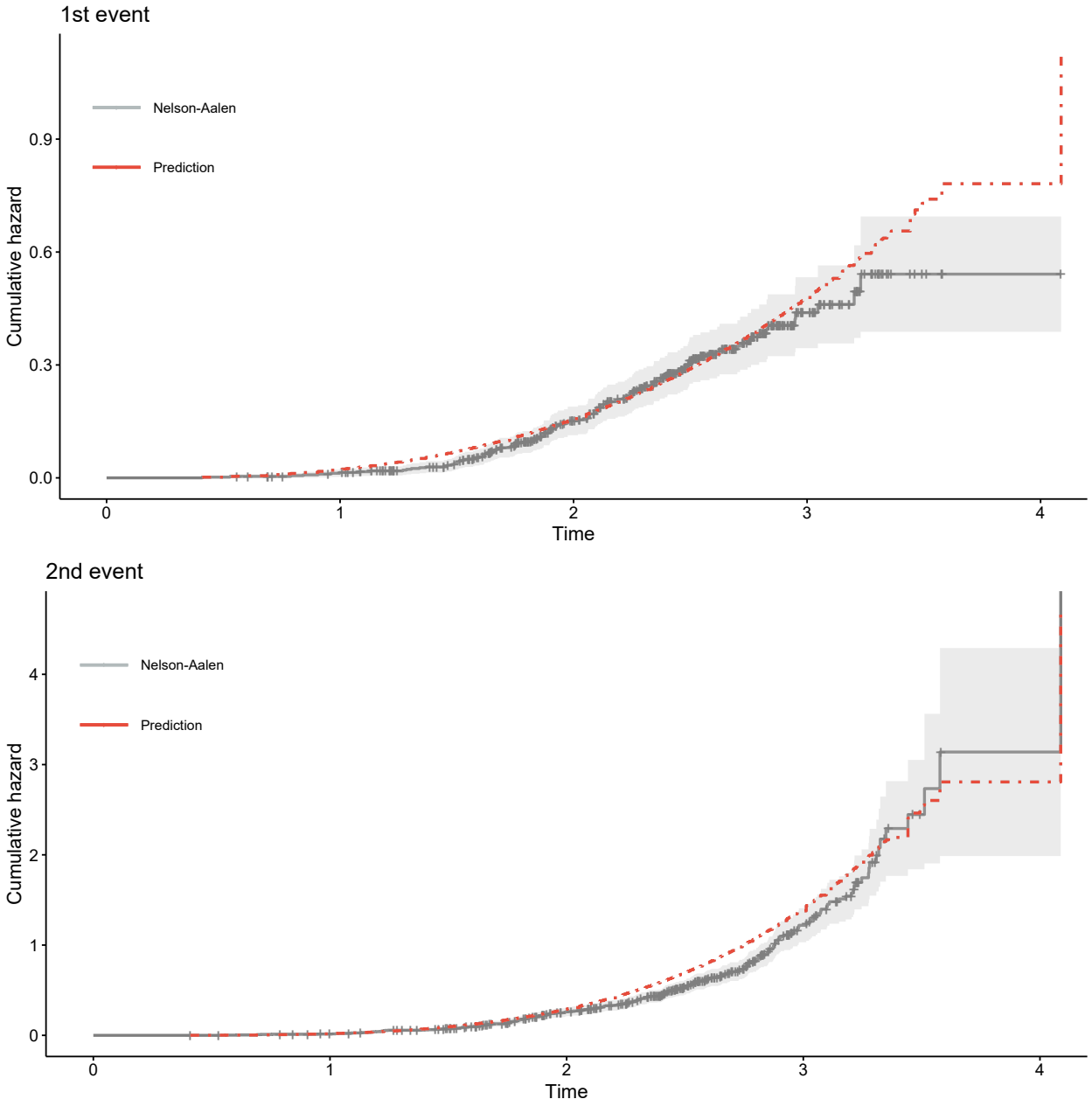

Figure V.12 – Comparison between predictive cumulative hazard function (in red) and Nelson-Aalen estimator (in black) for the risk of dementia (top) and death (bottom).

value or slope, which is not offered elsewhere. The second is that an illness-death model with interval censoring can be adjusted in this package, something that is not provided by other packages for standard shared random effects joint model. Note that our package can handle the case where residual variability is assumed to be homoscedastic, allowing users to run a classic joint model but with interval censoring.

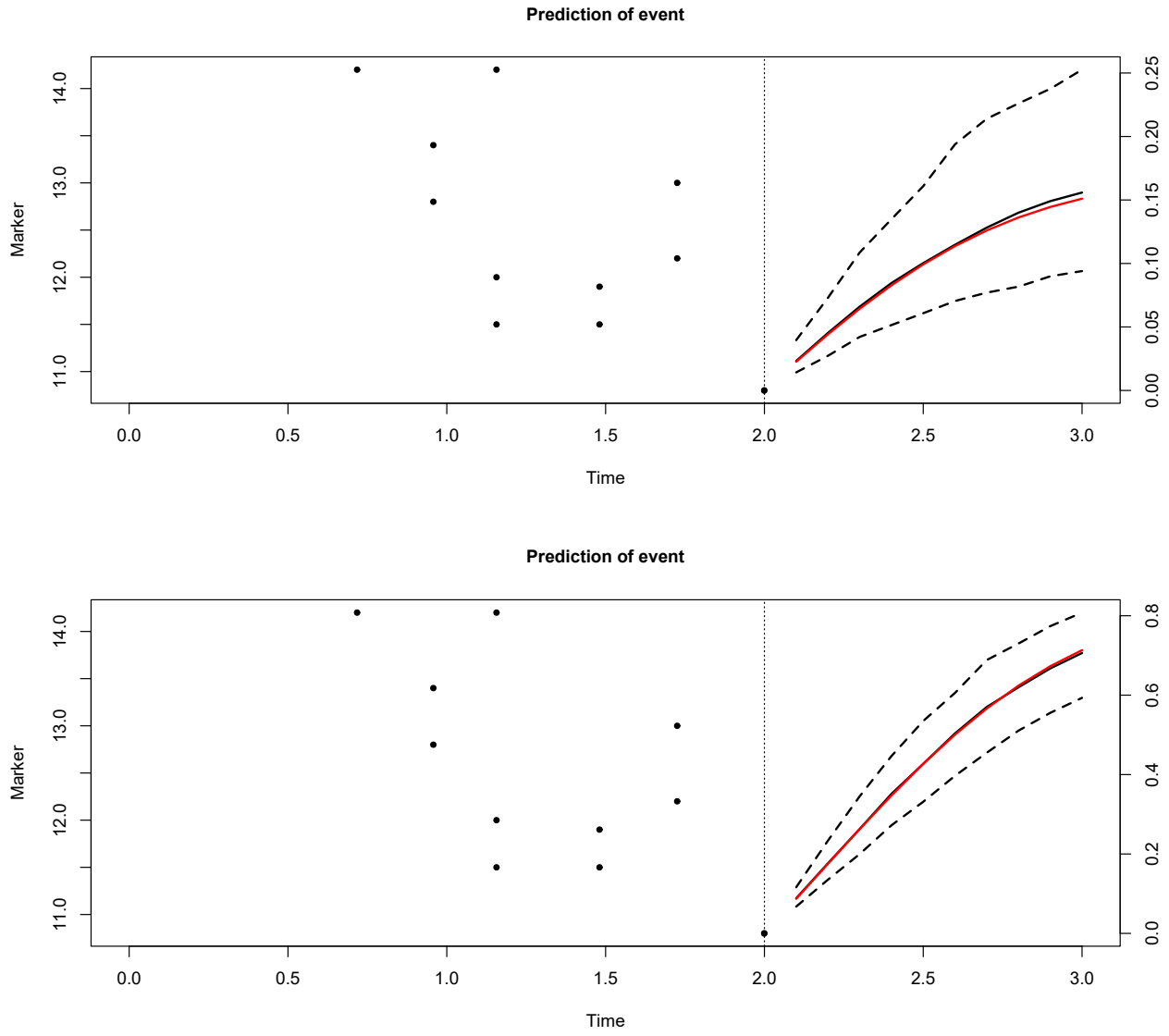

Figure V.13 – Prediction of the risk of dementia (top) and death (bottom) between 2 and 3 years (with its 95% confidence interval indicated by dashed lines and the median in black), for a patient at risk at 2 years.

All of this has been detailed earlier (Courcoul et al., 2024b,a) and illustrated using real `threeC` data from the 3C cohort which is included in the package.

However, the package has some computational limitations, particularly regarding computation time, which can become significantly prolonged when multiple random effects are included. Additionally, convergence issues may arise as the model complexity increases, especially with more intricate random effect specifications. In such cases, increasing the number of QMC draws is often necessary, though this will further extend computation time. Finally,

the proposed package allows only one continuous longitudinal marker which is assumed to be Gaussian.

Looking ahead, future developments to the `LSJM` package could include multi-markers models, expanding its capacity to handle more complex datasets. Additionally, extending the package to accommodate location-scale joint models with binary outcomes, as suggested by Elliott et al. (2012), would further broaden its utility. Then, it would be interesting to extend the estimation to more general family of mixed models either than just the linear one or to consider non linear effects of the current value or the variability on the risk of event.

Finally, while the package was initially conceived to analyse blood pressure variability in the context of dementia in a cohort study, as data available in `threeC` (Courcoul et al., 2024b), or cardiovascular events in a clinical trial (Courcoul et al., 2024a), its flexible framework makes it applicable to a wide range of other research areas and clinical applications. The development version of the package is available at `https://github.com/LeonieCourcoul/LSJM`.

Chapitre VI

Discussion

Sommaire

VI.1 Résumé des travaux de thèse

Dans ce travail de thèse, nous nous sommes intéressés à la prise en compte de l'effet de la variabilité individuelle d'un marqueur longitudinal sur des risques d'événements de santé. Pour répondre à cet objectif, nous avons développé deux modèles conjoints. Le premier modèle permet la prise en compte d'une variabilité individuelle pouvant dépendre du temps et d'autres covariables afin d'étudier son effet sur des événements de santé compétitifs. Dans le deuxième modèle nous nous sommes focalisés sur la distinction entre deux variabilités individuelles : la variabilité intra-visite, représentant le court-terme, et la variabilité inter-visites, associée au long terme. Pour la partie survie du modèle, un modèle multi-état permet de traiter les risques semi-compétitifs ainsi que la censure par intervalle afin d'étudier proprement l'effet de ces deux variabilités sur des événements de santé.

Ces deux nouveaux modèles ont été validés par des études de simulations. Ces dernières ont également permis de mettre en évidence l'importance de traiter la censure par intervalle lorsque la date exacte de survenue de maladie n'est pas connue. Des outils de prédictions individuelles des effets aléatoires mais aussi de prédictions individuelles dynamiques du risque d'événement ont également été développés. Dans l'étude de l'impact de la variabilité de la PAS sur le risque de CCVD ou de démence, ces modèles ont été appliqués à deux jeux de données. Nous avons ainsi pu confirmer que la variabilité individuelle de la PAS était un facteur de risque des CCVD mais aussi de décès. En comparaison à un modèle conjoint standard, il est apparu que le modèle avec variance résiduelle hétérogène avait une meilleure adéquation au modèle d'un point de vue individuel. Enfin, nous avons montré que la variabilité inter-visites de la PAS était un facteur de risque pour la démence mais nous n'avons en revanche trouvé aucun lien entre la variabilité intra-visite de la PAS et le risque de démence ou de décès.

Dans une optique de diffusion et de reproductibilité de la recherche, les modèles développés ont été implémentés dans un package `R` diffusé librement et accesible sur GitHub (bientôt sur le CRAN) sous le nom `LSJM`. Ce package permet d'estimer les deux modèles présentés dans cette thèse mais également de généraliser à l'ensemble des variantes des modèles *location-scale*, conjoints ou non, en fonction du choix de définition de la variabilité et du modèle de survie. Il permet également d'estimer des modèles conjoints standards avec une variance résiduelle commune entre les individus et au cours du temps. Enfin, le package propose des outils complémentaires à l'estimation de ces modèles, comme des sorties graphiques permettant d'évaluer l'ajustement du modèle aux données, ou encore des fonctions permettant de réaliser des prédictions dynamiques des risques étudiés.

VI.2 Limites

Ce travail présente cependant certaines limites. La première limite est computationnelle. Au début de ce travail de thèse, l'implémentation du premier modèle a été réalisée sous le paradigme bayésien. Cependant, le temps de calcul et les nombreux problèmes de convergences nous ont conduit à changer pour l'inférence fréquentiste, induisant moins de problèmes de convergence. Concernant le temps de calcul, il a pu être réduit en calculant la vraisemblance en `C++` plutôt qu'en `R`, puis en proposant des stratégies d'initialisation des paramètres et de choix du nombre de QMC. Le choix du nombre de QMC en deux étapes permet de jouer à la fois sur la précision des estimations et sur le temps de calcul, mais aucune règle générale concernant le choix de $S1$ et $S2$ peut être faite. Par la suite, il pourrait être intéressant de s'intéresser à d'autres algorithmes d'optimisation. Bien que Philipps et al. (2021) aient comparé l'algorithme de Marquardt-Levenberg à plusieurs autres algorithmes d'optimisations tels que EM, BFGS et L-BFGS-B, et démontrés la plus grande robustesse de Marquardt-Levenberg, l'approche Bayésienne développée dans le logiciel INLA pourrait être une alternative intréressante car beaucoup plus rapide (Rustand et al., 2023). Cependant les modèles LSJM n'entrent pas dans la catégories des modèles traités par INLA qui ne traite que des modèles qui peuvent s'exprimer comme des modèles à classes latentes Gaussiennes. Une combinaison d'approches serait donc nécessaire. Il n'est donc pas certain que la procédure globale reste plus rapide.

La deuxième limite des modèles proposés se rapporte à l'évaluation de l'estimation ainsi que la démonstration de sa supériorité par rapport à un modèle conjoint standard. En effet, l'évaluation globale du modèle longitudinal tel que proposée dans cette thèse ne fait pas apparaître la variabilité individuelle étant donné que les prédictions du marqueur sont moyennées sur l'ensemble des individus. Il n'y a donc que très peu de différences entre un modèle conjoint avec une variance résiduelle hétérogène et un modèle conjoint standard. Les graphiques individuels permettent de mieux appréhender l'impact de la modélisation individuelle de la variabilité dans l'ajustement du modèle aux données. Cependant, cela ne donne pas réellement accès à une évaluation de l'ajustement global du modèle. Il serait donc intéressant de développer de nouveaux outils d'évaluation du modèle.
Par ailleurs, il est nécessaire de distinguer les deux objectifs principaux pour lesquels ces modèles peuvent être employés. Il y a d'une part un objectif étiologique, afin de comprendre les mécanismes mis en jeux dans la survenue d'une maladie, et d'autre part un objectif de pré-

diction permettant de construire le modèle donnant les meilleures prédictions individuelles de survenue d'un événement. Pour ce qui est de ce deuxième objectif, les outils pour évaluer les capacités prédictives ont été développés cependant, bien qu'un effet soit trouvé pour l'impact de la variabilité sur le risque de CCVD, en terme de prédiction individuelle, l'amélioration reste très modeste par rapport à un modèle conjoint standard.

Ensuite, les applications présentent également quelques limites. En effet, dans le cadre de l'étude des CCVD grâce à l'essai clinique PROGRESS, la population était très particulière puisque les sujets ont déjà tous été victimes d'un AVC et dans le cadre d'un essai clinique, ils sont davantage suivis que la population générale. D'un point de vue santé publique, il serait donc pertinent de confirmer les résultats obtenus par rapport à l'impact de la variabilité de la PAS sur les CCVD dans une cohorte représentant la population générale. Concernant l'étude de la démence, nous avons effectivement considéré une cohorte en population générale mais en se focalisant sur les plus de 65 ans. Cependant, comme tendent à le souligner certaines études, la PAS à un âge de milieu de vie serait plus prédictive que la PAS au troisième âge (Livingston et al., 2020). Il pourrait donc être raisonnable de s'intéresser à l'impact de la variabilité de la PAS à l'âge moyen sur le risque de démence par rapport à celle au troisième âge.

Enfin, il est nécessaire de bien comprendre la quantité représentée par la variabilité résiduelle. En effet, les modèles *location-scale* prennent en compte l'ensemble de la variabilité résiduelle, que ce soit la variabilité liée à l'erreur de mesure ou biologique. Cependant, si le modèle pour modéliser la trajectoire individuelle du marqueur est très flexible, alors la variabilité résiduelle risque de diminuer, voir de ne capturer que la variabilité liée à l'erreur de mesure, qui ne nous intéresse pas directement. C'est pourquoi il est essentiel de bien définir l'objectif de l'étude, en fonction que l'on souhaite comprendre les mécanismes liés à la variabilité ou construire le meilleur modèle prédictif et de discuter avec les spécialistes du domaine d'application pour connaître l'évolution attendue de la variable longitudinale d'intérêt. Par ailleurs, il est intéressant de contraster cette définition de la variabilité avec celle de Wang et al. (2024), détaillée en section II.4.3. Ces auteurs définissent la variabilité comme l'intégrale de la dérivée seconde de la trajectoire du marqueur. Elle intègre donc la totalité des fluctuations du marqueur et non la variabilité résiduelle seule. L'interprétation est probablement assez proche de notre variabilité résiduelle dans le cas où la trajectoire individuelle du LSJM est supposée linéaire mais elle sera très différente avec un modèle de trajectoire individuelle plus flexible.

VI.3 Perspectives

Les modèles proposés tout au long de cette thèse peuvent être étendus de multiples manières, tant au point de vue méthodologique afin de traiter d'autres particularités des données longitudinales ou de survie, que d'un point de vue applicatif. Cette section présente certaines de ces extensions possibles, sans pour autant être exhaustive.

VI.3.1 Extensions du modèle de survie

VI.3.1.1 Événements récurrents

Les événements récurrents correspondent à un événement qui se produit plusieurs fois au cours du suivi. Par exemple, un patient peut subir plusieurs AVC au cours de sa vie. Les modèles à fragilités permettent d'analyser de tels événements (Rondeau et al., 2007). Le modèle à fragilité correspond à un modèle de survie cause-spécifique dans lequel un effet aléatoire est introduit pour prendre en compte la répétition des événements dans la modélisation du risque. Ce modèle a par ailleurs été étendu pour prendre en compte un événement terminal, par exemple le décès, en parallèle des événements récurrents (Rondeau et al., 2007). Afin de prendre en compte l'évolution d'un marqueur pour estimer les risques d'événements récurrents, (Król et al., 2016) ont proposé un modèle conjoint à fragilité tel que :

$$\begin{cases} Y_{ij} = X_{ij}^{\top}\beta + Z_{ij}^{\top}b_i + \epsilon_{ij} \\ r_i(t) = r_0(t)\exp\left(\nu_i + W_i^{r\top}\gamma_r + g_i\left(b_i, t\right)^{\top}\alpha_r\right) \\ \lambda_i(t) = \lambda_0(t)\exp\left(\eta\nu_i + W_i^{\lambda\top}\gamma_\lambda + h_i\left(b_i, t\right)^{\top}\alpha_\lambda\right) \end{cases} \tag{VI.1}$$

Les fonctions $r_i(t)$ et $\lambda_i(t)$ représentent les fonctions de risques pour les événements récurrents et l'événement terminal respectivement, avec $r_0(t)$ et $\lambda_0(t)$ les fonctions de risques de bases associées. Les fonctions g et h représentent les fonctions de dépendances entre les effets aléatoires du marqueur longitudinal et les fonctions de survies. ν_i est un effet aléatoire partagé afin de prendre en compte la corrélation individuelle entre les événements récurrents et l'événement terminal et le paramètre η représente la force de l'association entre les deux risques. Si $\eta = 0$ alors ils sont indépendants. Le package `frailtypack` (Rondeau et al., 2012) permet d'estimer ce modèle.

Les modèles conjoints *location-scale* tels que proposés dans cette thèse pourrait très bien être étendus aux événements récurrents en permettant de prendre également en compte la variabilité comme association entre le marqueur et les fonctions de risques, de la même

manière qu'explicitée dans les précédents chapitres. Cela permettrait ainsi d'étudier l'impact de la variabilité d'un marqueur, par exemple la PA, sur le risques d'événements récurrents, tel que la survenue de plusieurs AVC tout en prenant en compte le décès comme événement terminal.

VI.3.1.2 Flexibilité du lien avec le marqueur longitudinal

Nous avons considéré comme fonctions de dépendance entre notre marqueur longitudinal et la fonction de risque seulement la valeur courante, la pente et la variabilité du marqueur. Il pourrait être possible d'utiliser l'aire cumulée sous la courbe de la valeur courante ou tout autre transformation des effets aléatoires (Rizopoulos et al., 2023).

Par ailleurs, nous avons supposé jusqu'à présent un effet linéaire entre la valeur courante, la pente ou la variabilité et le risque. Cependant, ces associations n'impactent pas nécessairement un événement de façon linéaire et il serait donc cliniquement pertinent de permettre des fonctions non linéaire de ces associations. En particulier cela pourrait permettre d'analyser des effets en U ou logarithmiques (Lee et al., 2022; van Dalen et al., 2022).

VI.3.1.3 Outcome binaire

Certaines études ne considèrent pas la durée de survenue d'un événement et à des données de survie mais plutôt à une variable réponse binaire sans temps d'événements. Dans ce cadre il pourrait être utile d'étendre le modèle proposé par Elliott et al. (2012) afin de considérer une variabilité du marqueur plus flexible. Cette variabilité pourrait ainsi dépendre du temps ou distinguer différentes variabilités.

VI.3.2 Extensions du modèle longitudinal

VI.3.2.1 Modèle mixte non linéaire

Dans certains cas, par exemple en pharmacologie, la dynamique du marqueur est définie de façon non linéaire (Lindstrom and Bates, 1990). Il convient alors d'utiliser un modèle conjoint non linéaire. Pour cela, le modèle longitudinal s'écrit :

$$Y_{ij} = m(X_{ij}^\top, \beta, Z_{ij}^\top, b_i) + \epsilon_{ij}$$

avec m une fonction non linéaire des paramètres et effets aléatoires. On suppose toujours ici que ϵ_{ij} suit une distribution gaussienne centrée. Ce genre de modèle pourrait facilement être

étendu en considérant une variance résiduelle individuelle de façon analogue à celle proposée dans la thèse.

VI.3.2.2 Variable longitudinale non-gaussienne

Dans les modèles développés dans cette thèse, nous nous sommes restreints à la modélisation de données longitudinales distribuées selon une loi normale. Cependant, lorsque ces données sont continues mais non gaussiennes ou lorsqu'elles sont ordinales, il convient d'utiliser d'autres modèles.

Dans le premier cas, un modèle linéaire généralisé à effets mixtes peut être utilisé. Ces modèles sont définis par une distribution de la famille exponentielle pour la variable réponse Y et une fonction de lien (non linéaire en général) entre l'espérance de Y et le prédicteur linéaire :

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

Si la distribution inclut un paramètre de variance, on peut définir un modèle *location-scale* en supposant que le paramètre de variance dépend d'effets aléatoires spécifiques aux sujets, du temps et d'autres covariables. En particulier cela est envisageable pour une distribution Gamma. Ainsi, on pourrait considérer un modèle tel que $Y_i \sim \Gamma(\mu_i, \phi_i)$ avec ϕ_i spécifique au sujet et dépendant d'effets aléatoires. La distribution Gamma peut être utilisée pour étudier une variable aléatoire strictement positive avec une longue queue, par exemple pour la concentration de médicaments dans le sang.

Les modèles pour données ordinales peuvent être définis comme des modèles à processus latent (Hedeker and Gibbons, 1994). Cette approche consiste à séparer le modèle structurel qui décrit la quantité d'intérêt (un processus latent) en fonction du temps et des covariables, du modèle de mesure reliant la quantité d'intérêt aux observations. Le processus latent $\Gamma_i(t)$ est défini en temps continu selon un modèle mixte linéaire standard :

$$\Gamma_i(t_{ij}) = X_{ij}^\top \beta + Z_{ij}^\top b_i + \epsilon_{ij} = \eta_{ij} + \epsilon_{ij}$$

Supposons que $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_i^2(t)$ avec $\sigma_i^2(t))$ définit comme dans le chapitre III ou IV. Considérons le marqueur ordinal Y. Alors, pour $m \in \{1, ...M\}$,

$$Y_{ij} = m \iff \delta_{m-1} < \eta_{ij} + \epsilon_{ij} \leq \delta_m$$

Cela signifie que la variable observée prend la modalité m lorsque la quantité d'intérêt $\Gamma_i(t_{ij})$

est dans l'intervalle $]\delta_{m-1}, \delta_m]$. Ainsi, on obtient un modèle probit avec une variance résiduelle hétérogène entre les individus, où la probabilité d'observer la modalité $m \in, \{1, ...M\}$ est donnée par :

$$\mathbb{P}(Y_{ij} = m) = \mathbb{P}(\delta_{m-1} - \eta_{ij} < \epsilon_{ij} \leq \delta_m - \eta_{ij}) = F_{\epsilon i}(\delta_m - \eta_{ij}) - F_{\epsilon i}(\delta_{m-1} - \eta_{ij})$$

avec $F_{\epsilon i}$ la fonction de répartition de la loi normale centrée et de variance $\sigma_i^2(t)$. Notons que pour que le modèle soit identifiable, il est alors nécessaire de fixer un paramètre. En générale, la variance de l'intercept aléatoire est fixé à 1.

VI.3.2.3 Multi-marqueurs

Lorsque plusieurs marqueurs ou variables longitudinales sont disponibles, les modèles conjoints multi-marqueurs peuvent être utilisés afin d'étudier leur relation avec la survenue d'un événement de santé. Cela permet de prendre en compte une plus grande quantité d'informations afin de prédire la survenue d'un événement. En particulier, en combinant les informations issues de plusieurs marqueurs et en les reliant à un modèle de survie, les modèles conjoints multi-marqueurs offrent une meilleure capacité de prédiction du risque d'événements.

L'extension des modèles conjoints *location-scale* tels que présentés dans cette thèse à l'aspect multi-marqueurs pourrait permettre de prendre en compte plusieurs marqueurs tout en supposant que la variabilité résiduelle serait dépendante du temps pour certains marqueurs, homogène entre les individus pour d'autres ou encore composée d'une partie inter-visites et d'une autre intra-visite. Par exemple, dans le cadre de l'étude de la démence, on pourrait alors considérer tous les tests cognitifs réalisés à chaque visite étant donné que ces tests sont des marqueurs très importants de la démence. Notons toutefois que l'analyse multi-marqueurs est particulièrement intéressante dans un objectif de prédiction. Dans un objectif étiologique cela peut être plus compliqué. Par exemple, dans l'étude de la démence on pourrait ne plus trouver de lien entre la pression artérielle et le risque de démence ou de décès si l'on prends en compte les tests cognitifs, ceux-ci étant bien plus prédicteurs que la PA et pouvant jouer le rôle de médiateur dans la relation PA et démence.

Cependant, le développement de modèles conjoints multi-marqueurs est très lourd d'un point de vue computationnel, menant à des temps de calcul particulièrement long et potentiellement des problèmes de convergence. Pour réduire ce temps de calcul, il est possible de

considérer une approche en deux étapes, récemment proposée par Baghfalaki et al. (2024). Dans notre contexte, la première étape consisterait à estimer des modèles conjoints *location-scale* uni-marqueurs, pour chaque marqueur étudié, puis à inclure des fonctions des effets aléatoires prédits dans le modèle de survie. Bien que ce soit plus lourd computationnellement, il est préférable d'estimer des modèles conjoints uni-marqueurs lors de la première étape plutôt que des modèles linéaires mixtes car cela permet d'éviter les biais liés à la sortie d'étude informative.

VI.3.2.4 Régression quantile

Jusqu'à présent, nous avons exclusivement considéré la régression linéaire mixte pour le marqueur longitudinal. Cependant, en présence de distributions asymétriques et de valeurs extrêmes, la médiane est une mesure de tendance centrale plus pertinente que la moyenne et l'étude de l'association entre le risque d'un événement et les quartiles ou les quantiles plus extrêmes du marqueur peut être intéressante. Une telle étude est possible à l'aide des modèles conjoints combinant un modèle de survie et une régression quantile et ces modèles pourraient être rendus plus flexibles en supposant une variance hétérogène du marqueur. Il serait alors possible de prendre en compte l'erreur résiduelle autour de ce quantile dans la fonction de survie. Cela permettrait d'étudier l'impact de la variabilité résiduelle autour d'un quantile sur la survenue d'un événement.

Dans l'approche paramétrique de la régression quantile, le marqueur d'intérêt Y est supposé suivre une distribution de Laplace $\mathcal{AL}(\mu, \sigma, \tau)$, de paramètre de localisation $\mu \in \mathbb{R}$, d'échelle $\sigma \in \mathbb{R}_+^*$ et d'asymétrie $\tau \in]0, 1[$. Cette distribution de probabilité est intéressante dans sa définition car le paramètre de localisation μ correspond au quantile d'ordre τ de la distribution considérée. Ainsi, pour un quantile d'intérêt d'ordre τ, la régression linéaire quantile à effets mixtes (Geraci, 2014) se définit par :

$$Y_{ij} = q_{i,j,\tau} + \epsilon_{ij} = X_{ij}^\top \beta_\tau + Z_{ij}^\top b_{i,\tau} + \epsilon_{ij}$$

avec :

- β_τ : les effets fixes pour le quantile d'ordre τ, associés au vecteur de covariable X_{ij},
- $b_{i,\tau}$: les effets aléatoires pour le quantile d'ordre τ, associés au vecteur de covariable Z_{ij},
- $\epsilon_{i,j,\tau}$: l'erreur résiduelle telle que $\epsilon_{i,j,\tau} \sim \mathcal{AL}(0, \sigma, \tau)$,
- $q_{i,j,\tau}$: le quantile d'ordre $\tau \in]0, 1[$ de la distribution conditionnelle $Y|b$.

Afin de prendre une variance résiduelle hétérogène entre les individus et dépendante du temps ou d'autres covariables, il suffirait de considérer l'erreur résiduelle telle que $\epsilon_{i,j,\tau} \sim \mathcal{AL}(0, \sigma_{i,\tau}(t), \tau)$ avec $\log(\sigma_{i,\tau}(t)) = O_{ij}^{\top}\eta_\tau + W_{ij}^{\top}u_{i,\tau}$ avec η_τ des effets fixes et $u_{i,\tau}$ des effets aléatoires de la variance pour le quantile d'ordre τ.

VI.3.3 Perspectives cliniques et épidémiologiques

Dans ce travail, nous avons exclusivement étudié la variabilité de la pression artérielle systolique. Il pourrait cependant être intéressant de considérer l'impact de la variabilité de la pression artérielle diastolique ou de la pression artérielle moyenne sur les risques de CCVD ou de démence. En effet, bien que les études se soient penchées majoritairement sur l'effet de la PAS, il a été montré que la PAD pouvait avoir une utilité pronostique supplémentaire par rapport à la PAS pour prédire les CCVD, en particulier chez les individus de moins de 50 ans (Vishram-Nielsen et al., 2021). De plus, l'utilisation de ces modèles possède un spectre bien plus large et permet de considérer la variabilité de bien d'autres marqueurs pour d'une part comprendre l'étiologie d'une maladie et d'autre part améliorer la prédiction de certains événements de santé, à condition d'avoir collecté suffisamment de mesures du marqueur d'intérêt.
En effet, pour modéliser à la fois l'espérance individuelle et la variance individuelle, un nombre de mesures conséquent est nécessaire. Cela est maintenant possible grâce en particulier au développement de dispositifs portables permettant de collecter ces données fréquemment et facilement par les patients, même à domicile. Par exemple, la diffusion de questionnaires en lignes peut permettre de récupérer de façon fréquente l'état émotionnel de certaines personnes pour ensuite étudier l'impact de la variabilité émotionnelle sur les événements de santé mentale tels que le suicide ou la dépression.
Dans les centres hospitaliers et tout particulièrement aux urgences, certains biomarqueurs sont mesurés quasiment continuellement. Il peut donc être judicieux d'évaluer l'impact de la variabilité de certains biomarqueurs, tel que la pression artérielle sur le risque de certains événements, par exemple la survenue d'un vasospasme cérébral (De Courson, 2022) après une hémorragie sous-arachnoïdienne.
Cependant, il est important de souligner que bien qu'une plus grande quantité d'informations permet d'obtenir des estimations plus robustes et sans doute une meilleure prédiction, cela engendre un allongement significatif du temps de calcul.

Par ailleurs, le modèle distinguant la variabilité inter-visites de celle intra-visite peut être généralisé en considérant des périodes plutôt que des visites. Par exemple, on peut distinguer

une variabilité à court terme et une variabilité à long terme. Pour cela il convient de définir j non pas comme la visite mais comme une période et l comme des mesures collectées au court de cette période. Les covariables d'ajustement pour les effets fixes et aléatoires sont alors supposées constantes tout au long de la période j. Il pourrait également être possible de faire la distinction entre la variabilité diurne et la variabilité nocturne en considérant des mesures collectées pendant plusieurs nuits et plusieurs jours.

VI.4 Conclusion générale

Pour conclure, les applications menées dans cette thèse soulignent l'importance de la prévention concernant la pression artérielle et la nécessité de poursuivre les études pour évaluer l'impact de la variabilité de la pression artérielle sur divers événements de santé. En outre, Les modèles développés permettent de mieux comprendre l'impact de la dynamique d'un marqueur, et notamment sa variabilité, sur le risque de la survenue d'un événement. Ils sont robustes à certaines particularités importantes des données de survie : les risques compétitifs, la censure par intervalle avec un risque semi-compétitif et la troncature à gauche. Cependant, certaines limites computationnelles rendent ces modèles parfois longs à estimer et difficiles à évaluer. De nombreuses perspectives permettraient de prendre en compte des données de nature différentes afin de s'intéresser à d'autres applications, mais toutes ces extensions ont un coût computationnel non négligeable.

Bibliographie

3C Study Group (2003). Vascular factors and risk of dementia : design of the Three-City Study and baseline characteristics of the study population. *Neuroepidemiology*, 22(6) :316–325.

Aalen, O. (1976). Nonparametric Inference in Connection with Multiple Decrement Models. *Scandinavian Journal of Statistics*, 3(1) :15–27.

Aalen, O. O. and Johansen, S. (1978). An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. *Scandinavian Journal of Statistics*, 5(3) :141–150.

Abell, J. G., Kivimäki, M., Dugravot, A., Tabak, A. G., Fayosse, A., Shipley, M., Sabia, S., and Singh-Manoux, A. (2018). Association between systolic blood pressure and dementia in the Whitehall II cohort study : role of age, duration, and threshold used to define hypertension. *European Heart Journal*, 39(33) :3119–3125.

AD, R. (2022). 2022 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 18(4).

Aitkin, M. (1987). Modelling Variance Heterogeneity in Normal Regression Using GLIM. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3) :332–339.

Albert, P. S. and Shih, J. H. (2010). On estimating the relationship between longitudinal measurements and time-to-event data using a simple two-stage procedure. *Biometrics*, 66(3) :983–987 ; discussion 987–991.

Alioum, A. and Commenges, D. (1996). A Proportional Hazards Model for Arbitrarily Censored and Truncated Data. *Biometrics*, 52(2) :512–524.

Alpérovitch, A., Blachier, M., Soumaré, A., Ritchie, K., Dartigues, J.-F., Richard-Harston, S., and Tzourio, C. (2014). Blood pressure variability and risk of dementia in an elderly cohort, the Three-City Study. *Alzheimer's & Dementia : The Journal of the Alzheimer's Association*, 10(5 Suppl) :S330–337.

Andersen, P. K. and Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine*, 31(11-12) :1074–1088.

Andrinopoulou, E.-R., Rizopoulos, D., Takkenberg, J. J. M., and Lesaffre, E. (2014). Joint modeling of two longitudinal outcomes and competing risk data. *Statistics in Medicine*, 33(18) :3167–3178.

Baghfalaki, T., Hashemi, R., Helmer, C., and Jacqmin-Gadda, H. (2024). A two-stage joint modeling approach for multiple longitudinal markers and time-to-event data. *submitted*, pages 1–33.

Barrett, J. K., Huille, R., Parker, R., Yano, Y., and Griswold, M. (2019). Estimating the association between blood pressure variability and cardiovascular disease : An application using the ARIC Study. *Statistics in Medicine*, 38(10) :1855–1868.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1) :1–48.

Betensky, R. and Mandel, M. (2015). Recognizing the problem of delayed entry in time-to-event studies : Better late than never for clinical neuroscientists. *Ann Neurol.*, 78(6) :839–44.

Blanche, P., Dartigues, J.-F., and Jacqmin-Gadda, H. (2013). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*, 32(30) :5381–5397.

Blanche, P., Proust-Lima, C., Loubère, L., Berr, C., Dartigues, J.-F., and Jacqmin-Gadda, H. (2015). Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics*, 71(1) :102–113.

Brent, R. P. (1973). *Algorithms for minimization without derivatives*.

Brier, G. W. (1950). Verification of forecasts expressed un terms of probability. Section : Monthly Weather Review.

Brilleman, S., Crowther, M., Moreno-Betancur, M., Buros Novik, J., and Wolfe, R. (2018). Joint longitudinal and time-to-event models via Stan. StanCon 2018. 10-12 Jan 2018. Pacific Grove, CA, USA.

Chow, C. K., Teo, K. K., Rangarajan, S., Islam, S., Gupta, R., Avezum, A., Bahonar, A., Chifamba, J., Dagenais, G., Diaz, R., Kazmi, K., Lanas, F., Wei, L., Lopez-Jaramillo, P., Fanghong, L., Ismail, N. H., Puoane, T., Rosengren, A., Szuba, A., Temizhan, A., Wielgosz, A., Yusuf, R., Yusufali, A., McKee, M., Liu, L., Mony, P., Yusuf, S., and PURE (Prospective Urban Rural Epidemiology) Study investigators (2013). Prevalence, awareness, treatment, and control of hypertension in rural and urban communities in high-, middle-, and low-income countries. *JAMA*, 310(9) :959–968.

Clark, III, D., Colantonio, L. D., Min, Y.-I., Hall, M. E., Zhao, H., Mentz, R. J., Shimbo, D., Ogedegbe, G., Howard, G., Levitan, E. B., Jones, D. W., Correa, A., and Muntner, P. (2019). Population-Attributable Risk for Cardiovascular Disease Associated With Hypertension in Black Adults. *JAMA Cardiology*, 4(12) :1194–1202.

Commenges, D. and Jacqmin-Gadda, H. (2015). *Dynamical Biostatistical Models*. CRC Press.

Cook, R. D. and Weisberg, S. (1983). Diagnostics for Heteroscedasticity in Regression. *Biometrika*, 70(1) :1–10.

Courcoul, L., Barbieri, A., and Jacqmin-Gadda, H. (2023). *FlexVarJM : Estimate Joint Models with Subject-Specific Variance*. R package version 0.1.0.

Courcoul, L., Helmer, C., Barbieri, A., and Jacqmin-Gadda, H. (2024a). Joint model for interval-censored semi-competing events and longitudinal data with subject-specific within and between visits variabilities. *arXiv :2408.06769*.

Courcoul, L., Tzourio, C., Woodward, M., Barbieri, A., and Jacqmin-Gadda, H. (2024b). A location-scale joint model for studying the link between the time-dependent subject-specific variability of blood pressure and competing events. *arXiv :2306.16785*.

Cowles, M. K. (2004). Review of WinBUGS 1.4. *The American Statistician*, 58(4) :330–336.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2) :187–220.

Davidian, M. and Giltinan, D. M. (1993). Some Simple Methods for Estimating Intraindividual Variability in Nonlinear Mixed Effects Models. *Biometrics*, 49(1) :59–73.

De Courson, H. (2022). *Variabilité de la pression artérielle et risque cérébro-vasculaire : confirmation et nouvelles méthodes*. phdthesis, Université de Bordeaux.

de Courson, H., Ferrer, L., Barbieri, A., Tully, P. J., Woodward, M., Chalmers, J., Tzourio, C., and Leffondré, K. (2021). Impact of Model Choice When Studying the Relationship Between Blood Pressure Variability and Risk of Stroke Recurrence. *Hypertension*, 78(5) :1520–1526.

de Courson, H., Leffondré, K., and Tzourio, C. (2018). Blood pressure variability and risk of cardiovascular event : is it appropriate to use the future for predicting the present ? *European Heart Journal*, 39(47) :4220.

De Pouvourville, G. (2016). Coût de la prise en charge des accidents vasculaires cérébraux en France. *Archives of Cardiovascular Diseases Supplements*, 8(2) :161–168.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, 39(1) :1–22.

Ding, J., Davis-Plourde, K. L., Sedaghat, S., Tully, P. J., Wang, W., Phillips, C., Pase, M. P., Himali, J. J., Windham, B. G., Griswold, M., Gottesman, R., Mosley, T. H., White, L., Guðnason, V., Debette, S., Beiser, A. S., Seshadri, S., Ikram, M. A., Meirelles, O., Tzourio, C., and Launer, L. J. (2020). Antihypertensive medications and risk for incident dementia and Alzheimer's disease : a meta-analysis of individual participant data from prospective cohort studies. *The Lancet Neurology*, 19(1) :61–70.

Dzubur, E., Ponnada, A., Nordgren, R., Yang, C.-H., Intille, S., Dunton, G., and Hedeker, D. (2020). MixWILD : A program for examining the effects of variance and slope of time-varying variables in intensive longitudinal data. *Behavior Research Methods*, 52(4) :1403–1427.

eClinicalMedicine (2023). The rising global burden of stroke. *eClinicalMedicine*, 59.

Elashoff, R. M., Li, G., and Li, N. (2008). A Joint Model for Longitudinal Measurements and Survival Data in the Presence of Multiple Failure Types. *Biometrics*, 64(3) :762–771.

Elliott, M. R., Sammel, M. D., and Faul, J. (2012). Associations between Variability of Risk Factors and Health Outcomes in Longitudinal Studies. *Statistics in medicine*, 31(23) :2745–2756.

Fine, J. P. and Gray, R. J. (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, 94(446) :496–509.

Finkelstein, D. M. (1986). A Proportional Hazards Model for Interval-Censored Failure Time Data. *Biometrics*, 42(4) :845–854.

Forette, F., Seux, M.-L., Staessen, J. A., Thijs, L., Babarskiene, M.-R., Babeanu, S., Bossini, A., Fagard, R., Gil-Extremera, B., Laks, T., Kobalava, Z., Sarti, C., Tuomilehto, J., Vanhanen, H., Webster, J., Yodfat, Y., Birkenhäger, W. H., and for the Syst-Eur Investigators (2002). The Prevention of Dementia With Antihypertensive Treatment : New Evidence From the Systolic Hypertension in Europe (Syst-Eur) Study. *Archives of Internal Medicine*, 162(18) :2046–2052.

Foulley, J. L., San Cristobal, M., Gianola, D., and Im, S. (1992). Marginal likelihood and Bayesian approaches to the analysis of heterogeneous residual variances in mixed linear Gaussian models. *Computational Statistics & Data Analysis*, 13(3) :291–305.

Freitag, M. H., Peila, R., Masaki, K., Petrovitch, H., Ross, G. W., White, L. R., and Launer, L. J. (2006). Midlife pulse pressure and incidence of dementia : the Honolulu-Asia Aging Study. *Stroke*, 37(1) :33–37.

Gao, F., Miller, J. P., Xiong, C., Beiser, J. A., Gordon, M., and The Ocular Hypertension Treatment Study (OHTS) Group (2011). A joint-modeling approach to assess the impact of biomarker variability on the risk of developing clinical outcome. *Statistical Methods Applications*, 20(1) :83–100.

Geraci, M. (2014). Linear Quantile Mixed Models : The lqmm Package for Laplace Quantile Regression. *Journal of Statistical Software*, 57 :1–29.

Gerds, T. A., Cai, T., and Schumacher, M. (2008). The performance of risk prediction models. *Biometrical Journal. Biometrische Zeitschrift*, 50(4) :457–479.

Gerds, T. A. and Schumacher, M. (2006). Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. *Biometrical Journal*, 48(6) :1029–1040.

Gonnet, P. (2012). A Review of Error Estimation in Adaptive Quadrature. *ACM Computing Surveys*, 44(4) :22 :1–22 :36.

Gueorguieva, R., Rosenheck, R., and Lin, H. (2012). Joint modelling of longitudinal outcome and interval-censored competing risk dropout in a schizophrenia clinical trial. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, 175(2) :417–433.

Guo, X., Zhang, X., Guo, L., Li, Z., Zheng, L., Yu, S., Yang, H., Zhou, X., Zhang, X., Sun, Z., Li, J., and Sun, Y. (2013). Association Between Pre-hypertension and Cardiovascular Outcomes : A Systematic Review and Meta-analysis of Prospective Studies. *Current Hypertension Reports*, 15(6) :703–716.

Han, J., Slate, E. H., and Peña, E. A. (2007). Parametric latent class joint model for a longitudinal biomarker and recurrent events. *Statistics in medicine*, 26(29) :5285–5302.

Hankey, G. J. (2020). Population Impact of Potentially Modifiable Risk Factors for Stroke. *Stroke*, 51(3) :719–728.

Harville, D. A. (1974). Bayesian Inference for Variance Components Using Only Error Contrasts. *Biometrika*, 61(2) :383–385.

Hazzouri, A. Z. A., Haan, M. N., Kalbfleisch, J. D., Galea, S., Lisabeth, L. D., and Aiello, A. E. (2011). Life-Course Socioeconomic Position and Incidence of Dementia and Cognitive Impairment Without Dementia in Older Mexican Americans : Results From the Sacramento Area Latino Study on Aging. *American Journal of Epidemiology*, 173(10) :1148.

Hedeker, D. and Gibbons, R. D. (1994). A Random-Effects Ordinal Regression Model for Multilevel Analysis. *Biometrics*, 50(4) :933–944.

Hedeker, D., Mermelstein, R. J., and Demirtas, H. (2008). An Application of a Mixed-Effects Location Scale Model for Analysis of Ecological Momentary Assessment (EMA) Data. *Biometrics*, 64(2) :627–634.

Hedeker, D. and Nordgren, R. (2013). Mixregls : A program for mixed-effects location scale analysis. *Journal of Statistical Software*, 52(12) :1–38.

Helmer, C., Joly, P., Letenneur, L., Commenges, D., and Dartigues, J.-F. (2001). Mortality with Dementia : Results from a French Prospective Community-based Cohort. *American Journal of Epidemiology*, 154(7) :642–648.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4) :465–480.

Hickey, G. L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2018). joineRML : a joint model and software package for time-to-event and multivariate longitudinal outcomes. *BMC Medical Research Methodology*, 18(1) :50.

Houwelingen, H. v. and Putter, H. (2012). *Dynamic Prediction in Clinical Survival Analysis*. CRC Press, Boca Raton.

in't Veld, B. A., Ruitenberg, A., Hofman, A., Stricker, B. H. C., and Breteler, M. M. B. (2001). Antihypertensive drugs and incidence of dementia : the Rotterdam Study. *Neurobiology of Aging*, 22(3) :407–412.

Joly, P., Commenges, D., Helmer, C., and Letenneur, L. (2002). A penalized likelihood approach for an illness-death model with interval-censored data : application to age-specific incidence of dementia. *Biostatistics*, 3(3) :433–443.

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

Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282) :457–481.

Kim, J. (2003). Maximum Likelihood Estimation for the Proportional Hazards Model with Partly Interval-Censored Data. *Journal of the Royal Statistical Society Series B*, 65 :489–502.

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

Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, 38(4) :963–974.

Lane, C. A., Barnes, J., Nicholas, J. M., Sudre, C. H., Cash, D. M., Parker, T. D., Malone, I. B., Lu, K., James, S.-N., Keshavan, A., Murray-Smith, H., Wong, A., Buchanan, S. M., Keuss, S. E., Gordon, E., Coath, W., Barnes, A., Dickson, J., Modat, M., Thomas, D., Crutch, S. J., Hardy, R., Richards, M., Fox, N. C., and Schott, J. M. (2019). Associations between blood pressure across adulthood and late-life brain structure and pathology in the neuroscience substudy of the 1946 British birth cohort (Insight 46) : an epidemiological study. *The Lancet. Neurology*, 18(10) :942–952.

Law, C. G. and Brookmeyer, R. (1992). Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in Medicine*, 11(12) :1569–1578.

Lee, C. J., Lee, J.-Y., Han, K., Kim, D. H., Cho, H., Kim, K. J., Kang, E. S., Cha, B.-S., Lee, Y.-h., and Park, S. (2022). Blood Pressure Levels and Risks of Dementia : a Nationwide Study of 4.5 Million People. *Hypertension*, 79(1) :218–229.

Leffondré, K., Touraine, C., Helmer, C., and Joly, P. (2013). Interval-censored time-to-event and competing risk with death : is the illness-death model more accurate than the Cox model ? *International Journal of Epidemiology*, 42 :1177–1186.

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2(2) :164–168.

Lin, X., Raz, J., and Harlow, S. D. (1997). Linear mixed models with heterogeneous within-cluster variances. *Biometrics*, 53(3) :910–923.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics*, 46(3) :673–687.

Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., Brayne, C., Burns, A., Cohen-Mansfield, J., Cooper, C., Costafreda, S. G., Dias, A., Fox, N., Gitlin, L. N., Howard, R., Kales, H. C., Kivimäki, M., Larson, E. B., Ogunniyi, A., Orgeta, V., Ritchie, K., Rockwood, K., Sampson, E. L., Samus, Q., Schneider, L. S., Selbæk, G., Teri, L., and Mukadam, N. (2020). Dementia prevention, intervention, and care : 2020 report of the Lancet Commission. *Lancet*, 396(10248) :413–446.

Ma, Y., Tully, P. J., Hofman, A., and Tzourio, C. (2020). Blood Pressure Variability and Dementia : A State-of-the-Art Review. *American Journal of Hypertension*, 33(12) :1059–1066.

Ma, Y., Wolters, F. J., Chibnik, L. B., Licher, S., Ikram, M. A., Hofman, A., and Ikram, M. K. (2019). Variation in blood pressure and long-term risk of dementia : A population-based cohort study. *PLOS Medicine*, 16(11) :e1002933.

Mac Mahon, S., Neal, S., Tzourio, C., Rodgers, A., Woodward, M., Cutler, J., Anderson, C., and Chalmers, J. (2001). Randomised trial of a perindopril-based blood-pressure-lowering regimen among 6105 individuals with previous stroke or transient ischaemic attack. *The Lancet*, 358(9287) :1033–1041.

Marquardt, D. W. (1963). An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2) :431–441.

Martin, S. and Rast, P. (2024). *LMMELSM : Fit Latent Multivariate Mixed Effects Location Scale Models*. R package version 0.2.0.

McGrath, E. R., Beiser, A. S., DeCarli, C., Plourde, K. L., Vasan, R. S., Greenberg, S. M., and Seshadri, S. (2017). Blood pressure from mid- to late life and risk of incident dementia. *Neurology*, 89(24) :2447–2454.

Mehlum, M. H., Liestøl, K., Kjeldsen, S. E., Julius, S., Hua, T. A., Rothwell, P. M., Mancia, G., Parati, G., Weber, M. A., and Berge, E. (2018). Blood pressure variability and risk of cardiovascular events and death in patients with hypertension and different baseline risks. *European Heart Journal*, 39(24) :2243–2251.

Meyer, K., Houle, D., et al. (2013). Sampling based approximation of confidence intervals for functions of genetic covariance matrices. In *Proc. Assoc. Advmt. Anim. Breed. Genet*, volume 20, pages 523–526.

Molenberghs, G. and Verbeke, G. (2006). *Models for Discrete Longitudinal Data*.

Morgan-Wall, T. (2024). *spacefillr : Space-Filling Random and Quasi-Random Sequences*. R package version 0.3.3.

Nelson, W. (1969). Hazard Plotting for Incomplete Failure Data. *Journal of Quality Technology*, 1(1) :27–52.

Odell, P. M., Anderson, K. M., and D'Agostino, R. B. (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics*, 48(3) :951–959.

Pan, J. and Thompson, R. (2007). Quasi-Monte Carlo estimation in generalized linear mixed models. *Computational Statistics & Data Analysis*, 51(12) :5765–5775.

Parati, G., Di Rienzo, M., Ulian, L., Santucciu, C., Girard, A., Elghozi, J. L., and Mancia, G. (1998). Clinical relevance blood pressure variability. *Journal of Hypertension. Supplement : Official Journal of the International Society of Hypertension*, 16(3) :S25–33.

Pase, M. P., Beiser, A., Enserro, D., Xanthakis, V., Aparicio, H., Satizabal, C. L., Himali, J. J., Kase, C. S., Vasan, R. S., DeCarli, C., and Seshadri, S. (2016). Association of Ideal Cardiovascular Health With Vascular Brain Injury and Incident Dementia. *Stroke*, 47(5) :1201–1206.

Philipps, V., Hejblum, B. P., Prague, M., Commenges, D., and Proust-Lima, C. (2021). Robust and Efficient Optimization Using a Marquardt-Levenberg Algorithm with R Package marqLevAlg. *The R Journal*, 13 :273.

Philipson, P., Hickey, G. L., Crowther, M. J., and Kolamunnage-Dona, R. (2020). Faster Monte Carlo estimation of joint models for time-to-event and multivariate longitudinal data. *Computational Statistics & Data Analysis*, 151 :107010.

Philipson, P., Sousa, I., Diggle, P. J., Williamson, P., Kolamunnage-Dona, R., Henderson, R., and Hickey, G. L. (2018). *joineR : Joint Modelling of Repeated Measurements and Time-to-Event Data*. R package version 1.2.8.

Pinheiro, J., Bates, D., and R Core Team (2024). *nlme : Linear and Nonlinear Mixed Effects Models*. R package version 3.1-166.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69(2) :331–342.

Pringle, E., Phillips, C., Thijs, L., Davidson, C., Staessen, J. A., de Leeuw, P. W., Jaaskivi, M., Nachev, C., Parati, G., O'Brien, E. T., Tuomilehto, J., Webster, J., Bulpitt, C. J., Fagard, R. H., and Investigators, o. b. o. t. S.-E. (2003). Systolic blood pressure variability as a risk factor for stroke and cardiovascular mortality in the elderly hypertensive population. *Journal of Hypertension*, 21(12) :2251–2257.

Proust-Lima, C., Dartigues, J.-F., and Jacqmin-Gadda, H. (2016). Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death : a latent process and latent class approach. *Statistics in Medicine*, 35(3) :382–398.

Proust-Lima, C., Joly, P., Dartigues, J.-F., and Jacqmin-Gadda, H. (2009). Joint modelling of multivariate longitudinal outcomes and a time-to-event : A nonlinear latent class approach. *Computational Statistics & Data Analysis*, 53(4) :1142–1154.

Proust-Lima, C., Philipps, V., and Liquet, B. (2017). Estimation of Extended Mixed Models Using Latent Classes and Latent Processes : The R Package lcmm. *Journal of Statistical Software*, 78 :1–56.

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics : competing risks and multi-state models. *Statistics in Medicine*, 26(11) :2389–2430.

Rizopoulos, D. (2010). JM : An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. *Journal of Statistical Software*, 35 :1–33.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3) :819–829.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data : With Applications in R*. CRC Press.

Rizopoulos, D. (2016). The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC. *Journal of Statistical Software*, pages 1–46.

Rizopoulos, D., Papageorgiou, G., and Miranda Afonso, P. (2023). *JMbayes2 : Extended Joint Models for Longitudinal and Time-to-Event Data*. R package version 0.4-5.

Rondeau, V., Marzroui, Y., and Gonzalez, J. R. (2012). frailtypack : An R Package for the Analysis of Correlated Survival Data with Frailty Models Using Penalized Likelihood Estimation or Parametrical Estimation. *Journal of Statistical Software*, 47 :1–28.

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

Rothwell, P. M., Howard, S. C., Dolan, E., O'Brien, E., Dobson, J. E., Dahlöf, B., Sever, P. S., and Poulter, N. R. (2010). Prognostic significance of visit-to-visit variability, maximum systolic blood pressure, and episodic hypertension. *The Lancet*, 375(9718) :895–905.

Rouanet, A., Joly, P., Dartigues, J.-F., Proust-Lima, C., and Jacqmin-Gadda, H. (2016). Joint latent class model for longitudinal data and interval-censored semi-competing events : Application to dementia. *Biometrics*, 72(4) :1123–1135.

Rouch, L., Cestac, P., Sallerin, B., Piccoli, M., Benattar-Zibi, L., Bertin, P., Berrut, G., Corruble, E., Derumeaux, G., Falissard, B., Forette, F., Pasquier, F., Pinget, M., Ourabah, R., Danchin, N., Hanon, O., Vidal, J.-S., and for the S.AGES investigators (2020). Visit-to-Visit Blood Pressure Variability Is Associated With Cognitive Decline and Incident Dementia. *Hypertension*, 76(4) :1280–1288.

Rustand, D., van Niekerk, J., Krainski, E. T., Rue, H., and Proust-Lima, C. (2023). Fast and flexible inference for joint models of multivariate longitudinal and survival data using integrated nested laplace approximations. *Biostatistics*.

Salari, N., Morddarvanjoghi, F., Abdolmaleki, A., Rasoulpoor, S., Khaleghi, A. A., Hezarkhani, L. A., Shohaimi, S., and Mohammadi, M. (2023). The global prevalence of myocardial infarction : a systematic review and meta-analysis. *BMC cardiovascular disorders*, 23(1) :206.

Shimbo, D., Newman, J. D., Aragaki, A. K., LaMonte, M. J., Bavry, A. A., Allison, M., Manson, J. E., and Wassertheil-Smoller, S. (2012). Association Between Annual Visit-to-Visit Blood Pressure Variability and Stroke in Postmenopausal Women. *Hypertension*, 60(3) :625–630.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data : Extending the Cox Model*. Springer New York, NY.

Torp-Pedersen, C., Mortensen, R. N., Jeppesen, J., and Gerds, T. A. (2018). Blood pressure and the uncertainty of prediction using hazard ratio. *European Heart Journal*, 39(47) :4219.

Touraine, C., Gerds, T. A., and Joly, P. (2017). SmoothHazard : An R Package for Fitting Regression Models to Interval-Censored Observations of Illness-Death Models. *Journal of Statistical Software*, 79 :1–22.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data : an overview. *Statistica Sinica*, 14(3) :809–834.

Tsiatis, A. A., DeGruttola, V., and Wulfsohn, M. S. (1995). Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS. *Journal of the American Statistical Association*, 90(429) :27–37.

Vallée, A., Gabet, A., Grave, C., Sorbets, E., Blacher, J., and Olié, V. (2020). Patterns of hypertension management in France in 2015 : The ESTEBAN survey. *Journal of Clinical Hypertension (Greenwich, Conn.)*, 22(4) :663–672.

van Dalen, J. W., Brayne, C., Crane, P. K., Fratiglioni, L., Larson, E. B., Lobo, A., Lobo, E., Marcum, Z. A., Moll van Charante, E. P., Qiu, C., Riedel-Heller, S. G., Röhr, S., Rydén, L., Skoog, I., van Gool, W. A., and Richard, E. (2022). Association of Systolic Blood Pressure With Dementia Risk and the Role of Age, U-Shaped Associations, and Mortality. *JAMA internal medicine*, 182(2) :142–152.

van Middelaar, T., van Dalen, J. W., van Gool, W. A., van den Born, B.-J. H., van Vught, L. A., Moll van Charante, E. P., and Richard, E. (2018). Visit-To-Visit Blood Pressure Variability and the Risk of Dementia in Older People. *Journal of Alzheimer's disease : JAD*, 62(2) :727–735.

Vishram-Nielsen, J. K., Kristensen, A. M. D., Pareek, M., Laurent, S., Nilsson, P. M., Linneberg, A., Greve, S. V., Palmieri, L., Giampaoli, S., Donfrancesco, C., Kee, F., Mancia, G., Cesana, G., Veronesi, G., Grassi, G., Kuulasmaa, K., Salomaa, V., Palosaari, T., Sans, S., Ferrieres, J., Dallongeville, J., Söderberg, S., Moitry, M., Drygas, W., Tamosiunas, A., Peters, A., Brenner, H., Grimsgaard, S., Savallampi, M., Olsen, M. H., and On behalf of the MORGAM Project (2021). Predictive Importance of Blood Pressure Characteristics With Increasing Age in Healthy Men and Women. *Hypertension*, 77(4) :1076–1085.

Wang, C., Shen, J., Charalambous, C., and Pan, J. (2024). Modeling biomarker variability in joint analysis of longitudinal and time-to-event data. *Biostatistics*, 25(2) :577–596.

Wang, Z.-T., Xu, W., Wang, H.-F., Tan, L., Tan, C.-C., Li, J.-Q., Yu, J.-T., and Tan, L. (2018). Blood Pressure and the Risk of Dementia : A Dose-Response Meta-Analysis of Prospective Studies. *Current Neurovascular Research*, 15(4) :345–358.

WHO (2023). Global report on hypertension : the race against a silent killer. geneva : World health organization. *Licence : CC BY-NC-SA 3.0 IGO*.

Willey, J. Z., Moon, Y. P., Kahn, E., Rodriguez, C. J., Rundek, T., Cheung, K., Sacco, R. L., and Elkind, M. S. V. (2014). Population attributable risks of hypertension and diabetes for cardiovascular disease and stroke in the northern Manhattan study. *Journal of the American Heart Association*, 3(5) :e001106.

Williamson, P. R., Kolamunnage-Dona, R., Philipson, P., and Marson, A. G. (2008). Joint modelling of longitudinal and competing risks data. *Statistics in Medicine*, 27(30) :6426–6438.

Wimo, A., Seeher, K., Cataldi, R., Cyhlarova, E., Dielemann, J. L., Frisell, O., Guerchet, M., Jönsson, L., Malaha, A. K., Nichols, E., Pedroza, P., Prince, M., Knapp, M., and Dua, T. (2023). The worldwide costs of dementia in 2019. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, 19(7) :2865–2873.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1) :330–339.

Ye, W., Lin, X., and Taylor, J. M. G. (2008). Semiparametric modeling of longitudinal measurements and time-to-event data–a two-stage regression calibration approach. *Biometrics*, 64(4) :1238–1246.

Ypma, T. J. (1995). Historical Development of the Newton-Raphson Method. *SIAM Review*, 37(4) :531–551.

Yu, J.-m., Kong, Q.-y., Schoenhagen, P., Shen, T., He, Y.-s., Wang, J.-w., Zhao, Y.-p., Shi, D.-n., and Zhong, B.-l. (2014). The prognostic value of long-term visit-to-visit blood pressure variability on stroke in real-world practice : A dynamic cohort study in a large representative sample of Chinese hypertensive population. *International Journal of Cardiology*, 177(3) :995–1000.

Zheng, Y. and Heagerty, P. J. (2007). Prospective Accuracy for Longitudinal Markers. *Biometrics*, 63(2) :332–341.

Annexes
Package LSJM : Implémentation et exemples

The package allows to estimate the models presented in section V.3 and provide the output of post-fit computations detailed in section V.4. The two estimation functions, `lsmm()` and `lsjm()`, rely on programs written in `R` and `C++`. This section describes the data used for the examples and the calls of these functions with some illustrations based on fictive data.

Fictive data

To illustrate and showcase the full scope of the functions available in the package, we will use fictive data in the examples including the following variables:

- `Y`: a continuous variable corresponding to the longitudinal marker,
- `time`: a continuous variable corresponding to the time of the measurement,
- `ID`: the identification of individual,
- `X1`: a qualitative or quantitative covariate,
- `num.visit`: the identification of the visits,
- `W1, W2, W3`: three time-fixed qualitative or quantitative covariates.

Estimation function lsmm

The call of `lsmm()` is

Code R

```
lsmm(formFixed, formRandom, formGroup, timeVar,
```

```
formVar = "standard", formFixedVar = NULL, formRandomVar = NULL,
random_inter = F, random_intra = F, formGroupVisit = NULL,
correlated_re = F, data.long, S1 = 500, S2 = 1000,
nproc = 1, clustertype = "SOCK", maxiter = 100, print.info = F,
file = "", epsa = 1e-04, epsb = 1e-04, epsd = 1e-04, binit = NULL)
```

We consider the following arguments:

- `formFixed`: a two-sided formula for the regression model at the population level, with the response variable `Y` on the left-hand side of a $\sim$ operator, and the covariates from X associated with fixed effects, separated by $+$ operator, on the right-hand side.
- `formRandom`: a one-sided formula with the covariates from Z associated with random effects, separated by $+$ operator, on the right-hand side.
- `formGroup`: a one-sided formula with the identification variable of the random effects.
- `timeVar`: a character providing the name of time variable.
- `formVar`: a character defining which definition of variability is used (details below).
- `data`: a database containing the data in long format, meaning there are n_i rows per subject, or $\sum_{j=1}^{n_i} n_{ij}$ when there is an additional nested level in the individual data (e.g., multiple measurements for a single time point or a follow-up window).
- `formFixedVar`, `formRandomVar`, `random_inter`, `random_intra`, `formGroupVisit` and `correlated_re`: specified depending on the variability used. The different patterns are described below.

Regarding the arguments specific to the computational aspect, we have:

- `S1`: an integer proving the number of QMC draws for the fisrt step.
- `S2`: an integer proving the number of QMC draws for the second step. It should be provided only if the user want to do the precision improvement step. In this case, `S2`>`S1`.
- `nproc`: an integer providing the number of processors to used for parallel computing.
- `clustertype`: a character indicating a clustertype supported from `makeCluster`.
- `maxiter`: an integer providing the maximum number of iterations in the optimization algorithm.
- `print.info`: a boolean indicating if the outputs of each iteration should be written.
- `file`: a character giving the name of the file where the outputs of each iteration should be written.

- `epsa`, `epsb`, `epsd`: numbers providing the threshlods for convergence criteria on the parameters, the log-likelihood and the derivatives, respectively. These arguments are specific to the `MarqLevAlg` algorithm.
- `binit`: a vector providing the initial value.

Standard linear mixed model

To specify a linear mixed model with an homogeneous variability, the argument `formVar` should be equal to `"standard"`. Arguments `formFixedVar`, `formRandomVar`, `random_inter`, `random_intra`, `formGroupVisit` and `correlated_re` do not need to be specify.

Example of call

Code R

```
lsmmStandard <- lsmm(formFixed = Y ~ time + X1, formRandom = ~ time,
                     formGroup = ~ ID, timeVar = "time", formVar = "standard",
                     data.long = data_lsmm)
```

For dataset `data_lsmm`, the call `lsmmStandard` fits a standard linear mixed model in which the dependent variable `Y` is explained according to `time` and `X1`. Two correlated random effects are assumed for the `intercept` (mandatory) and `time`. These random effects are grouped by `ID`.

Location-scale mixed model with time-dependent residual variability

To specify a location-scale mixed model with covariate and time-dependent residual variability, the argument `formVar` should be equal to `"cov-dependent"`. Argument `formFixedVar` defines the one-sided formula with the covariates O, the fixed effects on the variability. Argument `formRandomVar` provides the one-sided formula with the covariates M, the random effects on the variability. Argument `correlated_re` indicates if the random effects of the mean b and the random effects of the variability τ should be correlated or not.

Example of call

Code R

```
lsmmCovDep <- lsmm(formFixed = Y ~ time + X1, formRandom = ~ time,
                   formGroup = ~ ID,  timeVar = "time",
                   formVar = "cov-dependent", formFixedVar = ~ time + X2,
```

```
                    formRandomVar = ~ time, correlated_re = F,
                    data.long = data_lsmm2)
```

For dataset `data_lsmm`, the call `lsmmCovDep` fits a location-scale mixed model in which the dependent variable `Y` is explained according to `time` and `X1`. On the mean, two correlated random effects are assumed for the `intercept` and `time`. The residual variability is modeled according to `time` and `X2` for the fixed part and two correlated random effects, `intercept` and `time`. Random effects of the mean are supposed to be independent from the random effects of the variability.

Location-scale mixed model with between and within visits variabilities

To specify a location-scale mixed model with between and within visits variabilities, the argument `formVar` should be equal to `"inter-intra"`. Argument `random_inter` and `random_intra` indicates if the between-visits variability and the within-visit variability respectively should be subject-specific or not. Argument `formGroupVisit` provides a one-sided formula with the visit indicator variable. Argument `correlated_re` indicates if the random effects of the mean b and the random effects of the variability τ should be correlated or not.

Example of call

Code R

```
lsmmInterIntra <- lsmm(formFixed = Y ~ time + X1, formRandom = ~ time,
                           formGroup = ~ ID, timeVar = "time",
                           formVar = "inter-intra", random_inter = T,
                           random_intra = T, formGroupVisit = ~ num.visit,
                           correlated_re = T, data.long = data_lsmm2)
```

For dataset `data_lsmm2`, the call `lsmmInterIntra` fits a location-scale mixed model in which the dependent variable `Y` is explained according to `time` and `X1`. On the mean, two correlated random effects are assumed for the `intercept` and `time`. The between-visits and within-visit variabilities are both subject-specific by the inclusion of a random effect on their intercept, and all random effects are correlated. Visits are indicated by the `num.visit` variable.

Function lsjm

The call of `lsjm()` is

Code R

```
lsjm(Objectlsmm,  survival_type = c('Single', 'CR', 'IDM'),
     formSurv_01, formSurv_02 = NULL, formSurv_12 = NULL,
     sharedtype_01, sharedtype_02 = NULL, sharedtype_12 = NULL,
     hazardBase_01, hazardBase_02 = NULL, hazardBase_12 = NULL,
     delta1, delta2 = NULL, Time_T, Time_L = NULL, Time_R = NULL,
     Time_T0 = NULL, formSlopeFixed = NULL, formSlopeRandom = NULL,
     index_beta_slope = NULL, index_b_slope = NULL,
     nb.knots.splines = c(1,1,1), nb_pointsGK = 15, S1 = 1000, S2 = NULL,
     binit = NULL, nproc = 1, clustertype = "SOCK", maxiter = 100,
     print.info = FALSE, file = NULL, epsa = 1e-03, eps 1e-03, epsd = 1e-03)
```

Using the following arguments:

- `Objectlsmm` an object of the `lsmm` function.
- `survival_type` a character defining which type of survival scheme to use (see details below).
- `formSurv_01` is a one-sided formula providing on the right-side the regression covariates from W^{01} associated with the proportional hazard model, separated by a + operator.
- `hazardBase_01` a character providing the baseline hazard function, which is in `c("Exponential", "Weibull", "Gompertz", "Splines")`.
  If `hazardBase_01 = "Spline"`, argument `nb.knots.splines` indicates the number of internal knots. The knots are placed at the quantiles of event times. If `nb.knots.splines = 1`, the knot is at the median of the event times.
- `sharedtype_01` a vector of character(s) indicating the form(s) of the dependence structure. If the longitudinal model estimated is a standard mixed model, `sharedtype_01` should be included in `c("value", "slope", "random effects")`. If it is a location-scale mixed model with a covariate and time-dependent variability one can add `"variability"` and if it is a model distinguishing within from between visits variabilities one can add `c("variability inter", "variability intra")` if there are subject-specifics. If `"slope"` is included in `sharedtype_01`, arguments `formSlopeFixed, formSlopeRandom, index_beta_slope, index_b_slope` should be provided.
- `formSlopeFixed` a one-sided formula corresponding to the derivative with respect to time of the function of fixed effects given by `formFixed`.
- `index_beta_slope` a vector of integer(s) providing the index of β parameters from the mixed model used in the slope.

- `formSlopeRandom` a one-sided formula corresponding to the derivative with respect to time of the function of random effects given by `formRandom`
- `index_b_slope` a vector of integer(s) indicating the index of b parameters from the mixed model used in the slope.
- `delta1` a one-sided formula with on the right-side the variable corresponding to the indicator δ_{1i} of event (1 for subject experimenting the event and 0 for others).
- `Time_T` a one-sided formula with on the right-side the variable corresponding to the time T_i of event.
- `Time_T0` a one-sided formula with on the right-side the variable corresponding to the time of entry T_{0i} which must be provided in case of delayed entry.

Arguments `formSurv_02, formSurv_12, sharedtype_02, shardetype_12, hazardBase_02, hazardBase_12, delta2, Time_L, Time_R` are specified similarly to previous arguments, depending on the chosen survival model as described below. Argument `nb_pointsGK` provides the number of points to use for the Gauss-Kronrod approximation (either 7 or 15). The other arguments, related to the computationel aspect are the same than for `lsmm()` function. Argument `S2` should be provided only if the user want to do the precision improvement step. In this case, `S2`>`S1`.

Joint models for a single event

To fit a model with only a single event, argument `survival_type` should be equal to `'Single'`. Arguments `formSurv_02, formSurv_12, sharedtype_02, shardetype_12, hazardBase_02, hazardBase_12, delta2, Time_L, Time_R` do not need to be specified. They are `NULL` by default.

Examples of call

Code R

```
lsjm1 <- lsjm(Objectlsmm = lsmmStandard, survival_type = 'Single',
    formSurv_01 = ~ W1, sharedtype_01 = c("value", "slope") ,
    formSlopeFixed = ~ 1, formSlopeRandom = ~ 1, index_beta_slope = c(2),
    index_b_slope = c(2), hazardBase_01 = "Splines", delta1 = ~ event,
    Time_T = ~ Time_event, Time_T0 = ~ Time_entry, nb.knots.splines = c(3))


lsjm2 <- lsjm(Objectlsmm = lsmmCovDep, survival_type = 'Single',
    formSurv_01 = ~ 1, sharedtype_01 = c("value", "variability"),
```

```
   hazardBase_01 = "Weibull", delta1 = ~ event, Time_T = ~ Time_event)
```

The call `lsjm1` fits a joint model with a longitudinal submodel with a standard mixed model defined previously and fitted in the object `lsmmStandard` and a survival submodel with only one event. The hazard baseline function is defined by splines with 3 internal knots. The survival submodel is adjusted on the covariate `W1` and on the current value and current slope of the marker. This call takes into account delayed entry with an entry time defined by `Time_entry`. For the slope, given that in the corresponding linear mixed model, the trajectory of the marker is defined by:

$$\widetilde{Y}_i(t) = \beta_0 + b_{0i} + (\beta_1 + b_{1i})t \tag{VI.2}$$

then, the slope is defined by:

$$\frac{\partial \widetilde{Y}_i(t)}{\partial t} = \tilde{Y}_i'(t) = \beta_1 + b_{1i} \tag{VI.3}$$

The call `lsjm2` fits a joint model combining a LSMM with covariate and time-dependent variability, defined previously and fitted in the object `lsmmCovDep`, and a proportional hazard submodel for only one event. The hazard baseline function is defined by a Weibull function. The survival submodel depends on the current value and variability of the marker without any other covariates.

Joint models for competing events

In the case of competing events, there are two events so `survival_type` should be equal to `'CR'` and `formSurv_02`, `sharedtype_02`, `hazardBase_02` and `delta2` must be defined. In this case, `Time_T` corresponds to the date of first event (or censoring). `delta1` and `delta2` can not be both equal to 1 for one subject, only the first event is recorded.

Examples of call

Code R

```
lsjm3 <- lsjm(Objectlsmm = lsmmCovDep, survival_type = 'CR',
   formSurv_01 = ~ W1, formSurv_02 = ~ W2,
   sharedtype_01 = c("value", "slope", "variability") ,
   sharedtype_02 = c("random effects", "variability") ,
   hazardBase_01 = "Splines",   hazardBase_02 = "Splines",
   formSlopeFixed = ~ 1, formSlopeRandom = ~ 1, index_beta_slope = c(2),
   index_b_slope = c(2),  delta1 = ~ event1, delta2 = ~ event2,
```

```
  Time_T = ~ Time_event, Time_T0 = ~ Time_entry, nb.knots.splines = c(3,1))

lsjm4 <- lsjm(Objectlsmm = lsmmInterIntra, survival_type = 'CR',
   formSurv_01 = ~ 1, formSurv_02 = ~ W1,
   sharedtype_01 = c("value", "variability inter", "variability intra"),
   sharedtype_02 = c("value", "variability inter"),
   hazardBase_01 = "Weibull", hazardBase_02 = "Gompertz",
   delta1 = ~ event1, delta2 = ~ event2,
   Time_T = ~ Time_event)
```

The call `lsjm3` fits a joint model combining a LSMM with covariate and time-dependent variability, (defined and fitted in `lsmmCovDep`) and two cause-specific proportional hazard models for two competing events. Both risk functions are defined with baseline hazard functions on splines with respectively 3 and 1 internal knots for the first and the second risk. The first risk function is adjusted on covariate `W1`, on the current value, the slope and variability of the marker. The second risk function is adjusted on covariate `W2` on the random effects and the variability of the marker. Subjects are left truncated at their time of inclusion `Time_entry`.

The call `lsjm4` fits a joint model combining a LSMM distinguishing within and between visits residual variabilities, (defined and fitted in `lsmmInterIntra`) and two proportional hazard models for competing events. The first risk function is adjusted on the current value, the inter visits variability and the intra visit variability of the marker. The second risk function is adjusted on the covariate `W1`, the current value and the inter visits variability of the marker. For the baseline risk functions we consider a Weibull function for the first risk function and a Gompertz for the second one.

Joint models for illness-death events

To fit a joint model including an illness-death model, `survival_type` should be equal to `"IDM"`, and `formSurv_12, sharedtype_12, hazardBase_12, Time_L` and `Time_R` must be defined. In this case, `Time_T` provides a one-side formula with a variable corresponding to the minimum time between time-to-transition to state (2) and censoring time, `Time_L` is a one-side formula with a variable giving the last observed time in state (0) and `Time_R` is a one-side formula with a variable providing the first time known in state (1). If the subject i is never seen in state (1) then it is necessary, for computation reasons, to fix $R_i = T_i$.

Examples of call

Code R

```
lsjm5 <- lsjm(Objectlsmm = lsmmCovDep, survival_type = 'IDM',
  formSurv_01 = ~ W1, formSurv_02 = ~ W2, formSurv_12 = ~ W1,
  sharedtype_01 = c("value", "variability"),
  sharedtype_02 = c("value","slope", "variability"),
  sharedtype_12 = c("value"),
  hazardBase_01 = "Splines", hazardBase_02 = "Exponential",
  hazardBase_12 = "Splines",
  formSlopeFixed = ~ 1, formSlopeRandom = ~ 1, index_beta_slope = c(2),
  index_b_slope = c(2),  delta1 = ~ event1, delta2 = ~ event2,
  Time_T = ~ Time_final, Time_L = ~ Time_left, Time_R = ~ Time_right,
  Time_T0 = ~ Time_entry, nb.knots.splines = c(3,0,2))

lsjm6 <- lsjm(Objectlsmm = lsmmInterIntra, survival_type = 'IDM',
  formSurv_01 = ~ W1+W2, formSurv_02 = ~ 1, formSurv_12 = ~ W3,
  sharedtype_01 = c("value", "variability inter"),
  sharedtype_02 = c("value","slope", "variability intra"),
  sharedtype_12 = c("value", "variability inter", "variability intra"),
  hazardBase_01 = "Weibull", hazardBase_02 = "Exponential",
  hazardBase_12 = "Gompertz",
  formSlopeFixed = ~ 1, formSlopeRandom = ~ 1, index_beta_slope = c(2),
  index_b_slope = c(2),  delta1 = ~ event1, delta2 = ~ event2,
  Time_T = ~ Time_final, Time_L = ~ Time_left, Time_R = ~ Time_right)
```

The call `lsjm5` fits a joint model combining a LSMM with covariate and time-dependent variability, (defined and fitted in `lsmmCovDep`) and an illness-death model with 3 transitions. For transitions (01) and (12), the baseline hazard functions are defined on splines with respectively 3 and 2 internal knots. These two transitions intensities are adjusted on the same covariate `W1` and on the current value of the marker. Transition (01) is also adjusted on the current residual variability. Transition (02) is defined using an exponential baseline hazard function and is adjusted on covariate `W2`, the current value, the slope and the current residual variability of the marker. Subjects are left truncated (`Time_T0 = Time_entry`. The last visit seeing healthy is defined by covariate `Time_left` and the first visit seeing in state (1) by `Time_right`. The last time of information is `Time_final`.

The call `lsjm6` fits a joint model combining a LSMM with between-visits and within-

visit residual variabilities, defined and fitted in `lsmmInterIntra` and an illness-death model with 3 transitions. The baseline hazard functions are respectively a Weibull, an Exponential and a Gompertz function. Transition (01) is adjusted on `W1`, `W2`, the current value and the between-visits variability of the marker. Transition (02) is adjusted on the current value, the slope and the between-visits variability of the marker. Finally, transition (12) is adjusted on `W3`, the current value, and on both between-visits and within-visit variabilities of the marker.

Maximum likelihood estimates: generic function summary

Whatever the estimated model (lsmm or lsjm), the `summary()` method return maximum likelihood estimates, estimated standard errors, test statistics and p-values associated with Wald's test. The estimated upper triangular variance-covariance matrix of the maximum likelihood estimates is also provided. All estimated parameters and their standard errors can be also obtained in the `x$table.res` object with `x` a `lsmm` or `lsjm` object.

Random effects prediction: generic functions ranef and predict

Functions `ranef()` and `predict()` return the predicted random effects for all subjects and `predict()` also provides the longitudinal predictions of the marker or the predictive cumulative hazard function(s) (see details below).

Examples of call

Code R

```
re <- ranef(m1)
```

The `m1` object is either a `lsmm` object or a `lsjm` object.

Longitudinal and survival predictions: generic function predict

The `predict` function provides several predictions. Its call is

Code R

```
predict(object, which = c('RE','Y','Cum'), ranefObject = NULL,
                                                    data.long = NULL)
```

where:

- `object` provides either a `lsmm` object or a `lsjm` object,

- `which` indicates which predictions is computed. `'RE'` corresponds to the random effects, `'Y'` to the marker and `'Cum'` to the cumulative risk functions,
- `data.long` contains the longitudinal dataset used for making predictions. If `data.long` is not provided, the predictions are performed on the data used to estimate the model.

First, with the option `which = "RE"` it provides the prediction of the random effects as the `ranef` function. Secondly, with the option `which = "Y"` it computes the prediction of the marker `Y`. This function returns a table of individual predicted values of marker `Y` for visit times for each subject considered in the study. The option `which = c("RE", "Y")` indicates to compute both the predictions of random effects and of the marker. In this case, the function will returns two tables, one for the random effects and one for the marker. Finally, with the option `which = "Cum"` it computes the predicted cumulative risk function(s). The function returns as many tables as transitions with in each table the predicted value for the cumulative risk function corresponding to the transition.

Examples of call

Code R

```
predY.1 <- predict(m1, which = c("RE", "Y"))
predY.2 <- predict(m1, which = c("Y"), ranefObject = re)
predCum <- predict(m1, which = c("Cum"), ranefObject = re)
predTot <- predict(m1, which = c("RE", "Y","Cum"))
```

The `m1` object is either a `lsmm` object or a `lsjm` object. The call `predY.1` returns two tables, one for the predicted random effects and one for the predicted values of the marker. The call `predY.2` uses the previously predicted random effects `re` which is an object of the `ranef` function. The call `predTot` computes all possible predictions. For calls `predCum` and `predTot`, `m1` must be a `lsjm` object.

Goodness-of-fit: generic function plot

The `plot()` function provides several graphics. The call of the function is

Code R

```
plot(object, which = c('long.fit', 'survival.fit', 'traj.ind'),
                              predictObject, break.times, ID.ind, ...)
```

with the following arguments:

- `object` either a `lsmm` object or a `lsjm` object,

- `which` a character indicating which plot must be display,
- `predictObject` an object of the `predict` function,
- `break.times` a vector of breaking times to create windows if `which = 'long.fit'`,
- `ID.ind` a vector providing the id of subjects for whom we we wish to trace the individual trajectory if `which = 'traj.ind'`.

First, if argument `which = "long.fit"` `plot` function allows to assess the fit of the longitudinal submodel comparing the mean of marker predictions collected in some windows of times (defined by `break.times` or using percentiles) to the mean of the observed measurements and its 95% confidence interval.

Then, for a `lsjm` object only, if argument `which = "survival.fit"`, `plot()` function allows to assess the fit of the survival submodel. For each transition, the predicted cumulative hazard function at each event time is computed given the predicted random effects. Then the mean of the predicted cumulative hazard functions are compared with there Nelson-Aalen estimator in the case of a single event or with competing risks. For an illness-death model, the predicted cumulative hazard function for each transition is compared to an illness-death model estimated by penalized likelihood accounting for interval censoring with the `SmoothHazard` package (Touraine et al., 2017).

Lastly, with the option `which = "traj.ind"` the `plot` function represents the individual trajectory with its prediction interval.

Examples of call

Code R

```
plot(m1, which = 'long.fit', predictObject = predTot, break.times = vec_time)

plot(m2, which = 'survival.fit', predictObject = predTot,
                                        smoothHazObject = fit.smooth)


plot(m1, which = 'traj.ind', predictObject = predTot, ID.ind = vec_ind)
```

The `m1` object is either a `lsmm` object or a `lsjm` object, whereas the `m2` object must be a `lsjm` object. Argument `predictObject` provides the predicted random effects, marker values and cumulative hazard functions obtained with `predict` function. The argument `vec_time` provides the vector of breaking times to create windows. If this argument is not provided, the windows are created taking the percentiles of the visit times. The argument `smoothHazObject` provides the estimation of the illness-death model obtained with

SmoothHazard package. The argument vec_ind is a vector providing the id of subjects for whom we wish to trace the individual trajectory.

Dynamic prediction of the event: function dynpred

The dynpred() function allows to compute individual dynamic prediction of an event for any subject, included or not in the dataset used for estimating the model. The call of dynpred is

Code R

```
dynpred(newdata, lsjmObject, s, horizon, event, CI = 95, nb.draws = 500,
                                                                graph = F)
```

with the following arguments:

- newdata a dataset containing the individuals for whom predictions are expected.
- s the landmark time (a single value).
- horizon the horizon time of prediction: the probability of having the event between s and horizon is computed. It could be a vector or a single value.
- event an integer indicating for which event the prediction is computed. In the case of a competing risk model, then it could be equal to 1 or 2 whereas for both other models it could be equal to 1 only.
- CI an integer between 0 and 100 indicating the level for computing the confidence interval. By default, CI = 95 for a 95% confidence interval. If CI = NULL, then the confidence interval is not computed.
- nb.draws an integer providing the number of draws used to compute the confidence interval, 500 by default.
- graph a Boolean indicating if the graphic of dynamic prediction should be plotted or not.

Example of call

Code R

```
predyn <- dynpred(data2, m2, s = 3, horizon = c(4,5), event = 1,
    IC = 95, nb.draws = 1000)
```

In this example, m2 is a lsjm model. The predictions are computed for all subjects in data2 for the first event between times 3 and 4 and between times 3 and 5. The 95% confidence interval is computed with 1000 draws.

Modèles conjoints avec variance résiduelle hétéroscédastique : application à l'étude de l'impact de la variabilité de la pression artérielle sur des événements de santé compétitifs

Résumé : Ce travail a pour objectif de développer des modèles conjoints permettant de prendre en compte rigoureusement la variabilité individuelle d'un marqueur longitudinal comme facteur de risque d'événements de santé. Ces modèles conjoints combinent un modèle linéaire mixte avec une variance résiduelle hétéroscédastique pour décrire l'évolution dans le temps d'un biomarqueur, et un modèle de survie pour un ou deux événements (semi-)compétitifs dans lequel le(s) risque(s) d'événement est(sont) ajusté(s) sur la valeur courante, la pente et la variabilité du biomarqueur. Le modèle linéaire mixte avec variance hétéroscédastique est un modèle linéaire mixte dans lequel la variabilité résiduelle est définie en fonction d'effets aléatoires individuels et/ou de covariables. Dans ces travaux, cela se traduit par une variabilité résiduelle individuelle pouvant dépendre du temps et de covariables, ou permettant de différencier les variances inter-visites et intra-visite lorsque plusieurs mesures sont collectées à chaque visite. Dans la première partie, nous proposons un modèle conjoint pour risques compétitifs et une variance résiduelle dépendante du temps. Appliqué à l'essai clinique PROGRESS, ce modèle permet d'étudier l'impact de la variabilité courante de la pression artérielle sur les risques d'événements cardio et cérébrovasculaires tout en prenant en compte le risque compétitif de décès. Dans une seconde partie, nous présentons un modèle conjoint permettant de distinguer les variabilités résiduelles inter-visites et intra-visite et pour un événement censuré par intervalle, en compétition avec un événement terminal, par exemple le décès. Ce modèle est appliqué à la cohorte des Trois Cités afin d'évaluer l'impact des variabilités inter-visites et intra-visite de la pression artérielle sur le risque de démence et de décès. La troisième partie présente un package `R` développé pour permettre aux utilisateurs d'estimer différents modèles conjoints issus de ce travail. Ils diffèrent selon la façon dont la variabilité résiduelle est modélisée et selon le type de modèle de survie utilisé. Ce dernier peut gérer un ou deux événements (semi-)compétitifs, ainsi que la censure par intervalle ou encore l'entrée retardée. Ce package propose aussi des outils pour l'évaluation graphique de l'ajustement du modèle aux données et la prédiction dynamique des risques d'événements de santé, basées sur les estimations du modèle.

Mots clés : Modèle conjoint, Pression artérielle, Variabilité

Joint models with heteroscedastic residual variance: application to the study of the impact of blood pressure variability on competing health events.

Abstract: This work aims to develop joint models that rigorously account for individual variability of a longitudinal marker as a risk factor for health events. These joint models combine a linear mixed-effects model with a heteroscedastic residual variance to describe the evolution over time of a biomarker, and a survival model for one or two (semi-)competitive events in which the event risk(s) is/are adjusted for the current value, slope, and variability of the biomarker. The linear mixed-effects model with heteroscedastic variance is a mixed-effects model where the residual variability is defined based on individual random effects and/or covariates. In this work, this translates to individual residual variability that can depend on time and covariates or allows differentiating between inter-visit and intra-visit variances when multiple measurements are collected at each visit. In the first part, we propose a joint model for competing risks with time-dependent residual variance. Applied to the PROGRESS clinical trial, this model allows studying the impact of current blood pressure variability on the risks of cardio and cerebrovascular events while accounting for the competing risk of death. In the second part, we present an illness-death joint model that distinguishes between inter-visit and intra-visit residual variability and accounts for an interval-censored event in competition with a terminal event, such as death. This model is applied to the Three-City cohort to assess the impact of inter-visit and intra-visit blood pressure variability on the risk of dementia and death.The third part describes an `R` package developed to allow users to estimate various joint models from this work. These models differ depending on how residual variability is modeled and the type of survival model used. The latter can handle one or two (semi-)competitive events, interval censoring, or delayed entry. This package also offers tools for graphical evaluation of model fit to the data and dynamic prediction of health event risks based on model estimates.

Keywords: Blood Pressure, Joint Model, Variability

Unité de recherche: Inserm U1219, *Bordeaux Population Health*, Université de Bordeaux