Modélisation de l'effet de facteurs de risque sur la probabilité de devenir dément et d'autres indicateurs de santé

Camille Sabathé

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE SOCIÉTÉS, POLITIQUE, SANTÉ PUBLIQUE

SPÉCIALITÉ Santé Publique option Biostatistique

Par Camille SABATHÉ

Modélisation du risque de démence vie-entière et autres indicateurs de santé

Sous la direction de Pierre Joly

Soutenue le 15 Novembre 2019

Membres du jury :

M. ALIOUM, Ahmadou	Pr, Université de Bordeaux	Président
Mme GUILLOUX, Agathe	Pr, Université d'Évry Val d'Essonne	Rapporteur
M. LATOUCHE, Aurélien	Pr, CNAM	Rapporteur
M. FOUCHER, Yohann	MCF, Université de Nantes	Examinateur
M. JOLY, Pierre	Pr, Université de Bordeaux	Directeur

Modélisation du risque de démence vie-entière et autres indicateurs de santé

Résumé : Les indicateurs épidémiologiques de la démence tels que l'espérance de vie sans démence pour un âge donné ou le risque absolu sont des quantités utiles en santé publique. L'observation de la démence en temps discret entraine une censure par intervalle du temps d'apparition de la pathologie. De plus, certains individus peuvent développer une démence et décéder entre deux visites de suivi. Un modèle *illness-death* pour données censurées par intervalle est une solution pour modéliser simultanément les risques de démence et de décès et pour éviter la sous-estimation de l'incidence de la démence. Ces indicateurs dépendent à la fois du risque de démence mais aussi du risque de décès, contrairement à l'intensité de transition de la démence. Les modèles de régression disponibles ne prennent pas en compte la censure par intervalle ou ne sont pas adaptés à ces indicateurs. L'objectif de ce travail est de quantifier l'effet de facteurs de risque sur ces indicateurs épidémiologiques par des modèles de régression. La première partie de cette thèse est consacrée à l'extension de l'approche par pseudo-valeurs aux données censurées par intervalle. Les pseudo-valeurs sont calculées à partir d'estimateurs paramétriques ou d'estimateurs du maximum de vraisemblance pénalisée. Elles sont utilisées comme variable d'intérêt dans des modèles linéaires généralisés ou des modèles additifs généralisés pour permettre un effet non-linéaire des variables explicatives quantitatives. La seconde partie de cette thèse porte sur le développement d'un modèle par linéarisation des indicateurs épidémiologiques. L'idée est de calculer l'indicateur conditionnellement aux variables explicatives à partir des intensités de transition d'un modèle *illness-death* avec censure par intervalle du temps d'apparition de la maladie. Ces deux approches sont appliquées aux données de la cohorte française PAQUID pour étudier par exemple l'effet d'un score psychométrique (le MMS) sur des indicateurs épidémiologiques de la démence.
Mots clés : pseudo-valeurs, censure par intervalle, démence, risque absolu, espérance de vie.

Modelization of life-time risk of dementia and others health indicators

Abstract: Dementia epidemiological indicators as the life expectancy without dementia at a specific age or the absolute risk are quantities meaningful for public health. Dementia is observed on discrete-time in cohort studies which leads to interval censoring of the time-to-onset. Moreover, some subjects can develop dementia and die between two follow-up visits. Illness-death model for interval-censored data is a solution to model simultaneously dementia risk and death risk and to avoid under-estimation of dementia incidence. These indicators depend on both dementia and death risks as opposed to dementia transition intensity. Available regression models do not take into account interval censoring or are not suitable for these indicators. The aim of this work is to propose regression models to quantify impact of risk factors on these indicators. Firstly, the pseudo-values approach is extended to interval-censored data. Pseudo-values are computed by parametric estimators or by maximum penalized likelihood estimators. Then pseudo-values are used as outcome in a generalized linear models or in a generalized additive models in case of non-linear effect of quantitative covariates. Secondly, the effect of covariates are summarized by linearization of the maximum likelihood estimator. In this part, the idea is to compute indicators conditionally on the covariates values from transition intensities of an illness-death model. These two approaches are applied to the French cohort PAQUID to study effect of a psychometric test (the MMS) on these indicators for example.
Key words: pseudo-values, interval censoring, dementia, absolute risk, life expectancy.

INSERM U1219 Bordeaux Population Health Research Center Equipe Biostatistique - 146 rue Léo Saignat - 33076 Bordeaux Cedex

Cette thèse a été préparée au sein de l'équipe Biostatistique du centre de recherche INSERM U1219 *Bordeaux Population Health*. Elle a été rendue possible grâce au financement de l'Agence Nationale de la Recherche dans le cadre du projet SMALA.

Table des matières

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué, à leur façon, à ce travail de thèse.

Mes premiers remerciements vont à Pierre Joly, mon directeur de thèse. Vous m'avez beaucoup appris pendant ces années et je vous dois beaucoup. Merci pour tout le temps que vous m'avez consacré malgré vos nombreuses responsabilités, je sais que la porte de votre bureau m'a toujours été grande ouverte. Je vous remercie aussi pour la patience dont vous avez fait preuve et surtout d'avoir su répondre à mes questionnements avec beaucoup de pédagogie. J'espère que le chemin parcouru à vos côtés ne s'arrête pas maintenant et que nous aurons l'occasion de collaborer ensemble par la suite.

Je veux ensuite adresser mes remerciements aux membres de mon jury. Je remercie sincèrement les Professeurs Agathe Guilloux et Aurélien Latouche d'avoir pris le temps de lire ce manuscrit et d'avoir accepté d'être rapporteurs de cette thèse. Mes remerciements vont aussi à M. Yohann Foucher, qui a accepté d'examiner ce travail de thèse. Je remercie enfin le Professeur Ahmadou Alioum de bien vouloir présider ce jury.

Cette thèse a été réalisée au sein de l'équipe Biostatistique et c'est tout naturellement que je tiens à remercier Hélène Jacqmin-Gadda. Vos compétences scientifiques et vos conseils m'ont permis d'avancer aux bons moments. Mes remerciements vont aussi à Daniel Commenges. Votre expérience est une richesse et une partie de cette thèse n'aurait pas pu être réalisée sans vos idées et vos remarques. Un très grand merci à Sandrine. Je te remercie de ne pas avoir eu à gérer le côté administratif de la thèse.

J'ai aussi une pensé pour Angéline Galvin, Leslie Grasset et Matthieu Frasca. Merci pour la confiance que vous m'avez accordée durant nos différents projets. J'ai pris beaucoup de plaisir à échanger avec vous et à comprendre qu'en pratique, ce n'est jamais aussi simple qu'en simulation.

Je tiens à remercier les personnes qui ont partagé mon bureau. Merci pour les échanges, les coups de mains et le soutien que l'on s'apporte mutuellement. Rémi et Bachirou, grâce à vous, j'ai pu voir les différentes étapes d'une thèse. Merci d'avoir montré l'exemple. Merci aux autres personnes ayant passé plus ou moins de temps dans ce bureau ; je pense en particulier à Mathilde, Denis, Julien et Céline. Merci de n'avoir rien dit lors des (nombreux) moments où je râle et je souffle ! Jocelyn et Anaïs, vous êtes arrivés au bureau au moment le plus critique de ma thèse. Merci de ne pas m'en avoir tenu rigueur. Mes trois années de thèse n'auraient pas été les mêmes sans Corentin.

Un grand merci à toi en particulier pour avoir partagé les moments de joie de cette thèse et les moments où j'ai eu envie de jeter mon ordinateur par la fenêtre. Merci de m'accompagner en pause (ou serait-ce l'inverse ?) et pour tous les autres moments. On est arrivé au même moment en 40 et cela sera un grand vide de ne plus te voir 5 jours sur 7 l'an prochain.

Je tiens aussi à remercier Maude et Casimir qui sont montés dans le bateau de la thèse en même temps que moi. Merci aussi à mes collègues du troisième, Sophie, Soufiane, Morgane, Perrine, Bénédicte, Aline, Virginie et Leslie pour ces déjeuners quotidiens, les petits dej et goûters ainsi que ces moments (plus ou moins) studieux que l'on a passé ensemble. Je remercie les autres doctorants, ingénieurs et chercheurs qui ont croisé ma route au BPH, durant mon passage par l'asso ou ailleurs.

Mes années ispédiennes n'auraient pas été les mêmes sans mes camarades de Master. Je tiens à remercier en particulier Élodie, Gwendoline, Caroline, Dhikra et Andréa ainsi que Noémie et Marie pour les moments partagés ensemble. C'est toujours avec un grand plaisir que je pars vous retrouver pour ces week-ends loin de Bordeaux.

Je ne peux écrire ces remerciements sans citer mes amies présentes depuis le lycée. Aurore, merci pour ta franchise et ton obstination, deux qualités que j'apprécie beaucoup chez toi. Merci à Coline d'avoir toujours cette joie de vivre. Merci pour tous ces traquenards du vendredi soir. Merci à vous d'avoir été là quand tout allait bien mais aussi dans les moments plus difficiles. Je remercie aussi les bordelaises exilées à Paris, Éva et Tifany, pour les moments mémorables que nous avons passé ensemble. Merci à toutes de m'avoir aérer l'esprit et de me rappeler qu'il n'y a pas que le travail dans la vie. Merci pour tous les souvenirs que j'ai créé avec vous, je suis sure qu'ils seront encore nombreux.

Un grand merci à ma famille, qui s'intéresse toujours à mon parcours. Je sais la chance que j'ai de vous avoir et je suis toujours très contente de vous retrouver, autour d'un bon repas ou d'une autre activité entre cousins. Je tiens à remercier spécialement ma sœur.

Un immense merci à Thomas. Merci de me soutenir et de me supporter tous les jours. Merci d'être à mes côtés, même à distance, je sais que je peux compter sur toi. Continue de me rendre heureuse, c'est le plus beau remerciement que je puisse te faire.

Je finis ces lignes en remerciant mes parents. Si j'en suis là aujourd'hui, c'est surtout grâce à vous. Vous m'avez toujours soutenue. Maman, merci encore pour tout ce que tu fais. Papa, je ne pensais pas écrire ce manuscrit un jour. Maintenant qu'il existe, je sais que tu es fier de moi. Merci infiniment à tous les deux.

Valorisation scientifique

Articles

— **Sabathé C**, Andersen PK, Helmer C, Gerds TA, Jacqmin-Gadda H et Joly P. Regression analysis in an illness-death model with interval-censored data : a pseudo-value approach. *Statistical Method in Medical Research*. 2019.

— **Sabathé C**, Commenges D, Helmer C, Jacqmin-Gadda H et Joly P. Linearization of covariates effects on complex quantities in an illness-death model dealing with left-truncation and interval-censoring. En préparation pour *Statistics in Medicine*.

Communications orales

— **Sabathé C**, Joly P. Modelling of the effect of explanatory variables on the probability of becoming demented : a pseudo-values approach. *49ème Journées de la Statistique de la SFdS*. 2017, Avignon, France

— **Sabathé C**, Joly P. Modelling of the effect of explanatory variables on health indicators : a pseudo-values approach. *Journées conjointes du groupe de recherche « Statistique et santé » et de la Société Française de Biométrie*. 2017, Bordeaux, France

— **Sabathé C**, Joly P. Modelization of the effect of covariates on dementia health indicators : approach by pseudo values. *Survival Analysis for Junior Researchers 2018*. 2018, Leiden, Pays-Bas

— **Sabathé C**, Joly P. A pseudo-values approach to model covariates effects on dementia health indicators. *XXIX International Biometric Conference*. 2018, Barcelone, Espagne

— **Sabathé C**, Commenges D, Joly P. Linearization of effect of covariates on complex quantities in an illness-death model dealing with left-truncation and interval-censoring. *Survival Analysis for Junior Researchers 2019*. 2019, Copenhague, Danemark

— **Sabathé C**, Commenges D, Joly P. A regression model to evaluate the effect of covariates on complex quantities in an illness-death model dealing with left-truncation and interval-censoring. *7th Channel Network Conference.* 2019, Harpenden, Royaume-Uni

Liste des figures

Liste des tableaux

Abréviations et notations

Abréviations

ADI *Alzheimer's Disease International*

ADL *Activity of Daily Living* - Activité de la vie quotidienne

APA *American Psychiatric Association*

BC Bande de confiance

CEP certificat d'étude primaire

CPI censure par intervalle

GAM Modèle Additif Généralisé

GEE Équations d' Estimation Généralisée

GLM Modèle Linéaire Généralisé

IADL *Instrumental Activity of Daily Living* - Activité instrumentale de la vie quotidienne

IC Intervalle de confiance

INSEE Institut National de la Statistique et des Études Économiques

MA Maladie d'Alzheimer

MMS *Mini Mental State*

OMS Organisation Mondiale de la Santé

PAQUID Personnes Agées QUID

RMSE Racine carrée de l'erreur quadratique moyenne

SMALA *Statistical Models for ALzheimer's disease and Aging* - Modèles statistiques pour la maladie d'Alzheimer et le vieillissement

WHO *World Health Organization*

Notations

$\alpha_{kl}(t)$ Intensité de transition de l'état k à l'état l au temps t

$\alpha_{kl}(t, Z_{kl})$ Intensité de transition entre l'état k et l'état l au temps t conditionnellement aux variables explicatives Z_{kl}

$\alpha_{kl}(t \mid Z_{kl})$ intensité de transition entre l'état k et l'état l au temps t à partir d'un modèle à intensités proportionnelles

β_{kl} Effets des variables explicatives sur une intensité de transition proportionnelles

Γ_j Effet de la variable j estimé par l'approche par linéarisation

γ_j Effet de la variable j estimé par l'approche par pseudo-valeurs

κ Paramètres de lissage de la vraisemblance pénalisée

$\theta(.)$ Estimateur d'une fonction des intensités de transition (p.exp. le risque absolu de la maladie)

$A_{kl}(s, t)$ Intensité de transition cumulée de l'état k à l'état l entre le temps s et t

$A_{kl}(s, t, Z_{kl})$ Intensité de transition cumulée entre l'état k et l'état l au temps t conditionnellement aux variables explicatives Z_{kl}

d_1 Fonction indicatrice de la démence

d_2 Fonction indicatrice du décès

$F_{01}(s,t)$ Risque absolu de la maladie entre les temps jusqu'en t d'un sujet non-malade au temps s

$LE_{0.}(s)$ Espérance de vie totale d'un sujet non malade à l'âge s

$LE_{00}(s)$ Espérance de vie sans démence d'un sujet non malade à l'âge s

$LTR(s)$ Risque de démence vie-entière d'un sujet non malade à l'âge s

$P_{00}(s,t)$ Probabilité d'être vivant non-malade au temps t pour un sujet non-malade au temps s

$RM(s,t)$ Moyenne restreinte des temps de survie en bonne santé jusqu'au temps t d'un individu non-malade au temps s

T Temps d'entré dans l'état 2 d'un modèle *illness-death*

T_0 Temps de sortie de l'état 0 d'un modèle *illness-death*

T_l Dernier temps où le sujet a été vu sain (borne droite de la CPI)

T_r Temps du diagnostic (borne droite de la CPI)

$Y_i(t)$ Pseudo-valeur du sujet i au temps t

Chapitre I

Introduction

I.1 Définition et épidémiologie de la démence

La démence est une neuro-pathologie qui touche principalement les personnes âgées de plus de 65 ans. Cette pathologie chronique entraine des altérations multidimensionnelles. La démence touche en effet les fonctions biologiques (avec une augmentation des protéines tau et β-amyloïde), les fonctions cérébrales (avec une atrophie cérébrale) et les fonctions cognitives (avec une dégradation des performances de la mémoire par exemple). La démence est un processus continu : il existe une première phase asymptomatique suivi d'une phase symptomatique où la dégradation des fonctions cognitives est plus marquée que lors de la première phase. Les symptômes sont variés et vont des pertes de mémoire aux sautes d'humeur en passant par des difficultés à accomplir les tâches de la vie quotidienne (WHO, 2017).

La démence regroupe différentes étiologies suivant les différentes altérations retrouvées chez le malade. La maladie d'Alzheimer (MA) est la principale cause de démence et représente au moins la moitié des cas de démence (Lobo *et al.*, 2000; Wu *et al.*, 2018). Les autres causes les plus répandues sont les démences vasculaires, les démences à corps de Lewy ou bien les démences fronto-temporales. Une étude *post-mortem* a montré que la plupart des cas de démences sont mixtes, c'est-à-dire que les atteintes cérébrales sont de type MA et de type démence vasculaire par exemple (Jelliger, 2006).

La prévalence de la démence est d'environ 50 millions de cas dans le monde en 2018 et l'incidence est d'environ 10 millions de nouveaux cas chaque année (ADI, 2015; ADI, 2018). Le coût financier mondial est estimé à 818 milliards de dollars américain (ADI, 2015). En France, les derniers chiffres de 2015 estiment la prévalence à 1,2 millions (ADI, 2015). Le coût financier de la MA est de 9,9 milliards d'euros pour la prise en charge médicale et médico-sociale (Association France Alzheimer, 2019).

D'après l'OMS, l'espérance de vie à la naissance est de 72 ans (74 ans pour les femmes et 69 ans pour les hommes) en 2019. En France, les dernières espérances de vie étaient de 85 ans pour les femmes et de 79 ans pour les hommes en 2018 (INSEE, 2018). L'espérance de vie à 65 ans

est en augmentation depuis les années 2000 et la part des personnes âgées de plus de 65 ans en France est d'environ 25% de la population totale. Cette augmentation de l'espérance de vie entraine par la même occasion une augmentation du nombre de cas de démence : les projections estiment que plus de 1,5 millions de personnes seront atteintes de démence en 2040 (suivant les scénarios démographiques et les variations potentielles de l'incidence, voir Helmer *et al.* (2006); Mura *et al.* (2010)).
Le vieillissement générale de la population est aussi visible mondialement. Par conséquent, les projections estiment à 75 millions en 2030 et 132 millions en 2050 le nombre de cas prévalents de démence (WHO, 2017). L'OMS a fait de la démence une priorité de santé publique en mettant en place un plan entre 2017 et 2025. Ce plan vise à améliorer la santé et le bien être des personnes démentes, aussi bien pour les personnes malades actuellement que pour les générations futures.

I.2 Diagnostic de la démence

Les différentes étiologies de la démence ne simplifient pas le diagnostic de cette pathologie. Un consensus établi repose sur les critères du manuel diagnostique et statistique des troubles mentaux (DSM) publié par l'*American Psychiatric Association* (APA, 1987; APA, 1994) ou sur les critères de la Classification Internationale des Maladies (CIM-10) de l'OMS (WHO, 1992).

Le diagnostic de la démence n'est donc pas facile à établir et repose sur plusieurs techniques comme un bilan neuro-psychologique ou de l'imagerie cérébrale. En médecine de ville, le praticien dispose de plusieurs tests psychométriques tels que l'examen du *Mini Mental State* (MMS, voir Folstein *et al.* (1975)) ou le test de la montre ainsi que l'évaluation des activités (instrumentales) de la vie quotidienne (Lawton et Brody, 1969; Katz , 1983) pour dépister une démence. Ces seuls tests ne se suffisent pas à eux-mêmes pour un diagnostic certain. En complément, le praticien peut se baser sur un questionnaire aux informants (comme les proches du patient) et peut effectuer des examens sanguins pour un diagnostic différentiel.

Cependant, une étude a montré qu'un peu moins de 550 000 personnes sont démentes en France d'après les données de l'assurance maladie en 2014. Ce nombre correspond au personnes ayant reçu un diagnostic (Carcaillon *et al.*, 2016), bien loin des estimations issues des données de cohorte. Le diagnostic de la démence n'est donc pas systématique.

I.3 Données d'étude : la cohorte PAQUID

En population générale, le diagnostic de démence arrive tardivement voire n'est pas établi pour certaines personnes. La tolérance de la pathologie existe chez les personnes âgées qui supposent que leurs symptômes sont dus à un vieillissement normal. De plus, les études cas-témoins rétrospectives sont biaisées à cause des troubles de la mémoire des malades. Les études de cohorte limitent ces deux points quand le suivi des participants est régulier. Ils en existent plusieurs à travers le monde :

aux États-Unis avec l'étude Framingham et le projet *Rush Memory and Ageing* (Farmer *et al.*, 1987; Bennett *et al.*, 2005), au Royaume Uni avec la cohorte établie à Cambridge (Brayne *et al.*, 1992) ou la cohorte prospective initiée aux Pays-Bas à la fin des années 1980 (Hofman *et al.*, 1991), avec une 4ème phase d'inclusion toujours en cours (Ikram *et al.*, 2017).
En France, plusieurs cohortes ont été mises en place à partir de la fin des années 1980 avec la cohorte PAQUID (Dartigues *et al.*, 1992), suivie de la cohorte des Trois Cités une décennie plus tard (3C Study Group , 2003) ou bien la cohorte AMI qui s'est intéressée en particulier à la démence dans le milieu agricole (Pérès *et al.*, 2012).

Les modèles de régression développés dans cette thèse ont été appliqués à la cohorte PAQUID. La richesse des données, tant d'un point de vue qualitatif que par la durée du suivi a motivé ce travail.

La cohorte PAQUID, pour « Personnes Âgées Quid » a inclus 3777 personnes en 1988 et 1989. L'inclusion a été faite par tirage au sort sur les listes électorales des personnes âgées d'au moins 65 ans et vivant à domicile en Gironde ou en Dordogne. Le taux d'acceptation initial était de 68 % et les participants étaient représentatif en terme de genre et d'âge de la population de ces deux départements (Dartigues *et al.*, 1991).
La mortalité des participants tout au long du suivi est similaire à celle de la population française de cette tranche d'âge. À l'inclusion, des données socio-démographiques comme l'âge, le genre, le niveau d'étude, le salaire ou bien les conditions de vie ont été recueilles auprès des participants ou bien des proches si ces derniers n'étaient pas en capacité de le faire.

L'objectif initial de PAQUID était d'étudier le vieillissement normal et pathologique du sujet âgé. En particulier, les objectifs spécifiques se sont concentrés sur l'étude de l'incidence, de la prévalence, des facteurs de risque de la démence et ses manifestations précliniques grâce à des mesures répétées des fonctions cognitives (Dartigues *et al.*, 1992). Des psychologues ont suivi les participants depuis maintenant plus de 27 ans avec une visite tous les deux à trois ans (à 1 an, 3 ans, 5 ans, 8 ans, 10 ans, 13 ans, 15 ans, 17 ans, 20 ans, 22 ans, 25 ans et 27 ans après l'inclusion). Une recherche active des cas de démence a été effectuée à l'inclusion et lors de chaque suivi. Les participants ont passé une série de tests psychométriques pour évaluer leurs fonctions cognitives. Le test du MMS a été utilisé pour évaluer les fonctions cognitives dans leur globalité et des tests spécifiques à une fonction cognitive particulière ont été menés comme la mémoire visuelle avec le test de Benton, la fluence verbale via le test d'Isaacs, l'attention par le test des barrages de Zazzo ou le test des codes qui mesure le raisonnement logique simple. La symptomatologie dépressive ainsi que la dépendance ont aussi été déterminées grâce à l'échelle CES-D et aux examens des *ADL* et des *IADL*. Les tests psychométriques évalués chez les participants de la cohorte PAQUID sont donc semblables à ceux utilisés en médecine de ville.

Une recherche active des cas de démence a été effectuée pour chaque suivi en se basant sur les critères du DSM (dans sa version III puis IV) avec la même procédure pour chaque visite. Un comité de suivi a examiné les cas suspectés par les psychologues pour confirmer ou infirmer la présence d'une démence. De plus, la date de décès des participants a été systématiquement

recherchée grâce aux données de l'état civil si cette information n'a pu être fournie par la famille ou le médecin traitant.

I.4 Facteurs de risque et facteurs protecteurs

L'étude des facteurs de risque ou protecteurs de la démence s'est fait depuis de nombreuses années. Certains facteurs sont controversés, d'autres sont maintenant reconnus et validés par la communauté scientifique. Les facteurs de risque modifiables sont des pistes de prévention de la démence car aucun traitement curatif n'est disponible à l'heure actuelle.

I.4.1 Facteurs non modifiables

Le principal facteur de risque de la démence est l'âge. En effet, l'incidence de la démence augmente de façon exponentielle avec celui-ci : l'incidence double par tranche de 5 ans avec une incidence de 2 pour 1000 personnes-années chez les 65-69 ans pour atteindre 74 nouveaux cas pour 1000 personnes-années chez les plus de 90 ans (Letenneur *et al.*, 1994; Fratiglioni *et al.*, 2000).

Le genre influe aussi sur le risque de démence : avant 80 ans, les hommes sont plus à risque de démence que les femmes. Cette tendance s'inverse après 80 ans (Letenneur *et al.*, 1999).

Le niveau d'éducation est aussi un élément à prendre en compte lors de l'analyse des facteurs de risque de démence. Un faible niveau d'étude (en général inférieur au certificat d'étude primaire ou CEP) est délétère pour l'incidence de la démence (Letenneur *et al.*, 1999).

La génétique joue un rôle important dans la démence : elle augmente le risque de maladie chez les plus jeunes et chez les plus âgés. Par exemple, la présence d'une allèle de l'apolipoprotéine $\epsilon 4$ augmente le risque de démence et est associé à une démence plus précoce (Prasher *et al.*, 2008).

I.4.2 Facteurs modifiables

Les facteurs modifiables comme la nutrition ouvrent un champ d'action considérable. La consommation de poissons par exemple serait protectrice du risque de démence (Kalmijn *et al.*, 1997; Lemeshow *et al.*, 1998; Ruitenberg *et al.*, 2002). De même, l'amélioration de l'état de santé général d'une population peut être bénéfique. Les maladies cardiovasculaires comme l'hypertension (Hofman *et al.*, 1997), le diabète (Ott *et al.*, 1996) ou la dépression (Jorm , 2000) sont des facteurs de risque de démence. La consommation de tabac est nocive pour le risque de démence (Merchant *et al.*, 1999).

I.5 Problèmes statistiques

Il existe différentes méthodes statistiques adaptées à l'étude étiologique de la démence c'est-à-dire à l'étude des facteurs de risque de la pathologie. Les schémas d'observation des données issues de cohortes peuvent être pris en compte par ces outils statistiques, comme le montrera la section II.1. D'autres indicateurs épidémiologiques, par exemple les espérances de vie ou le risque de démence vie-entière sont intéressants à étudier dans un contexte de santé publique ou dans le cadre de prédictions à une échelle plus individuelle. Cependant, ces indicateurs épidémiologiques dépendent à la fois du risque de la démence mais aussi du risque de décès (voir section II.2 pour plus de détails). De plus, le diagnostic de la démence n'est posé que lors des visites de suivi des cohortes, ce qui entraîne d'autres mécanismes d'observation à prendre en compte.

I.6 Objectifs de la thèse

L'objectif principal de cette thèse est de quantifier l'impact des déterminants de la démence sur des indicateurs épidémiologiques, par exemple le risque de démence vie-entière (ou sur toutes autres quantités d'intérêt). En d'autres termes, il s'agit de proposer des modèles de régression qui prennent en compte les mécanismes d'observation rencontrés dans les études de cohorte (comme la censure et la troncature ou le risque compétitif de décès). Ces objectifs sont les objectifs du *Work Package 2* du projet SMALA (*Statistical Models for ALzhiemer's disease and Aging*) financé par l'Agence Nationale de la Recherche (Investigateur Principal : Hélène Jacqmin-Gadda). En particulier, les modèles de régression proposés tiennent compte des particularités comme la censure par intervalle du début de la démence accompagnée du risque compétitif de décès. Ces deux points doivent être pris en compte simultanément car tous les cas de démence ne sont pas observés. Un autre objectif est de pouvoir tenir compte de l'entrée retardée dans les études de cohorte.

I.7 Plan

Cette thèse est organisée en différents chapitres. Le premier chapitre structure le cadre théorique de cette thèse en passant en revue les différentes méthodes de l'analyse des données de survie et en définissant des concepts clés utilisés tout au long de la thèse. La dernière partie de ce chapitre se concentrera sur le modèle *illness-death* pour temps de maladie censurée par intervalle ; c'est le modèle de régression servant de base au cœur de cette thèse. Ensuite, le deuxième chapitre présentera l'extension de l'approche par pseudo-valeurs pour temps de maladie censurées par intervalle dans un premier temps. Dans un second temps, cette approche sera appliquée pour étudier l'effet du score au MMS sur trois indicateurs épidémiologiques calculés pour un horizon fini. Puis, le troisième chapitre exposera la méthode de l'approche par linéarisation. L'approche par linéarisation sera employée sur trois indicateurs calculés pour un horizon infini. Enfin, une discussion et une conclusion seront apportées à ce travail. Des perspectives évoqueront des pistes

de développements futurs.

Chapitre II

État de l'art

L'état de l'art présenté dans ce chapitre pose le cadre théorique de ce travail. Cet état de l'art définit des concepts liés à la problématique statistique de la thèse et aux notions épidémiologiques.

II.1 Analyse des temps d'évènements

II.1.1 Observation des temps d'évènements

Lors de l'analyse de temps d'évènements, la modélisation se porte sur le temps de passage entre différents états. Ce temps peut être un délai depuis un diagnostic, le temps depuis l'inclusion dans une cohorte, l'âge au moment d'un évènement ou bien une durée. Différents mécanismes interviennent lors de l'analyse de temps d'évènements et ne permettent donc pas d'observer le temps de passage entre deux états pour tous les individus. Les schémas d'observation sont confrontés à différentes notions comme la troncature ou la censure. Notamment, la censure par intervalle est un schéma d'observation retrouvé dans l'analyse des données de cohorte étudiant la démence.

II.1.1.1 Censure à droite

L'analyse de données de temps jusqu'à un évènement suppose que si les individus sont suivis suffisamment longtemps, alors ils subiront tous l'évènement en question. En pratique, il est commun d'observer une censure à droite du temps d'évènement de certains sujets. La seule information disponible est de savoir que les sujets auront l'évènement après le temps de censure. D'un point de vue statistique, les variables aléatoires T et C sont respectivement définies comme le temps jusqu'à l'évènement et le temps de censure. Le temps jusqu'à l'évènement est en général appelé temps de survie. Soit δ l'indicatrice de l'évènement avec $\delta = \mathbb{1}\{T \leq C\}$. Ainsi, le couple de variables aléatoires $\left(\tilde{T}, \delta\right)$ correspond au temps d'évènement observé avec $\tilde{T} = \min(T, C)$.

Dans les études de cohortes, les individus peuvent être censurés à droite pour plusieurs raisons :

une censure intervient quand le sujet refuse les visites de suivi, lorsqu'il déménage et que le suivi ne peut avoir lieu ou bien lorsque l'étude se termine (cas de la censure administrative présente à 27 ans après l'inclusion pour les données disponibles dans la cohorte PAQUID).

II.1.1.2 Troncature à gauche

La troncature à gauche est une notion fréquente dans les études de cohortes. Elle arrive lorsque l'évènement d'intérêt ne peut survenir qu'après un temps précis noté T_e, c'est-à-dire que $T > T_e$ (avec $T_e \perp\!\!\!\perp T$ en général). Par exemple, lorsque l'on étudie l'âge jusqu'à un évènement et que les sujets ne sont pas suivis depuis leur naissance, le temps de survie est alors tronqué à gauche. C'est le cas dans la cohorte PAQUID où les individus étaient inclus à partir de 65 ans : l'âge jusqu'au décès est tronqué à gauche puisque seuls les sujets ayant survécu au moment de l'inclusion ont pu être observés, avec dans ce cas T_e leur âge à l'inclusion.

II.1.1.3 Censure par intervalle

L'apparition d'une démence n'est pas observée en continu dans les études de cohorte car le diagnostic de la démence n'est posé que lors d'une visite de suivi. De ce fait, le début de la pathologie est censurée par intervalle entre la dernière visite où le sujet a été vu non-dément et la visite de diagnostic. Si T représente le temps de démence, T_l le dernier temps où le sujet a été vu non-dément et T_r le temps du diagnostic, alors $T_l \leq T_{dem} \leq T_r$.

II.1.1.4 Risques compétitifs

L'analyse de données de type « temps d'évènements » suppose que chaque sujet est à risque de faire un évènement d'intérêt jusqu'à ce qu'il le subisse. Dans certaines applications, le sujet subit un autre évènement qui l'empêche de faire l'évènement étudié : le terme de risque compétitif est alors employé. Dans le cadre de cette thèse, le décès est un risque compétitif de la démence à partir du moment où les individus décédés non-déments ne peuvent pas développer une démence.

II.1.1.5 Censure par intervalle et risques compétitifs

La censure par intervalle du début de la démence et le risque compétitif de décès doivent être pris en compte simultanément dans l'analyse des données de cohorte. D'une part, les sujets de la cohorte PAQUID sont âgés ; le risque de décès augmente au fur et à mesure avec l'âge. D'autre part, le risque relatif de décès des personnes démentes est plus important que pour les personnes non démentes (Helmer *et al.*, 2001).

Une des solutions pour traiter des données censurées par intervalle est d'imputer une valeur du temps de maladie pour les individus diagnostiqués. Il vaut mieux éviter de procéder ainsi pour éviter

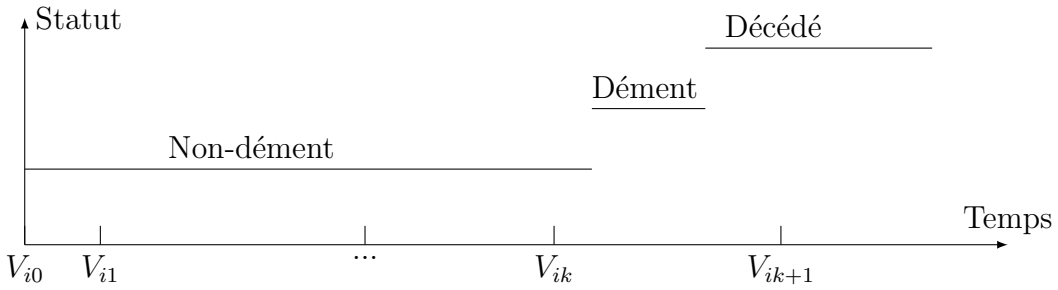

FIGURE II.1 – Exemple de suivi du sujet i

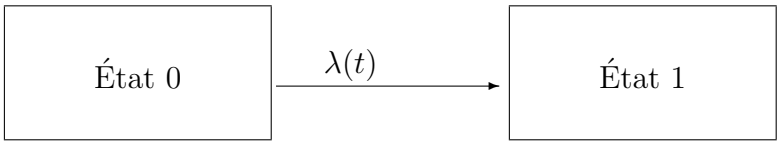

FIGURE II.2 – Représentation du modèle de survie

les biais d'estimation de l'incidence de la démence (Leffondré *et al.*, 2013; Binder et Schumacher, 2014). Cette censure par intervalle accompagnée du risque compétitif de décès entraine en fait un nombre de cas incidents observés plus faible à partir du moment où certains individus vont développer une démence et décéder entre deux visites de suivi. La figure figure II.1 est un exemple de ce cas de figure où le sujet i est vivant non-dément à sa k^{e} visite, développe une démence et décède avant la visite $k+1$ prévu initialement dans le protocole de l'étude.

II.1.2 Modèle multi-états

Les processus de comptage sont une solution pour modéliser les temps d'évènements grâce aux modèles multi-états (Andersen *et al.*, 1993). Cette section va définir trois cas particuliers des modèles multi-états.

II.1.2.1 Survie

L'analyse de survie est l'application la plus simple des modèles multi-états. Ici, il s'agit d'étudier le lien entre le passage d'un état initial (l'état sain en général) à un autre état (le décès ou la maladie par exemple) par une unique transition, comme le montre la figure II.2.

Plusieurs fonctions sont associées à la distribution des temps de survie T. La première fonction est la densité de probabilité, notée $f(t)$ et définie par :

$$f(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}\left(t \leq T < t + \Delta t\right)}{\Delta t}$$

La fonction de répartition $F(t)$ et la fonction de survie $S(t)$ sont égales à :

$$F(t) = \mathbb{P}(T \leq t) = \int_0^t f(u)\mathrm{d}u$$
$$S(t) = \mathbb{P}(T > t) = 1 - F(t)$$

Les deux autres fonctions utilisées (en particulier dans les modèles de régression, *cf.* section II.3) sont la fonction de risque $\lambda(t)$ et la fonction de risque cumulé $\Lambda(t)$ et sont définis par :

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}\left(t \leq T < t + \Delta t \mid T \geq t\right)}{\Delta t} = \frac{f(t)}{S(t)}$$
$$\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u$$

Le passage d'une fonction définie ci-dessus à une autre est aisée. Ainsi, la fonction de survie $S(t)$ est égale à

$$S(t) = \exp\left(-\int_0^t \lambda(u)\mathrm{d}u\right) = e^{-\Lambda(t)} \tag{II.1}$$

Des estimateurs non-paramétriques sont utilisés pour estimer les différentes fonctions liées à l'analyse de survie. Pour chaque temps d'évènement t_j, $j = 1, ..., k$, soient n_j l'effectif à risque et d_j le nombre d'évènement. La fonction de survie $S(t)$ peut être estimée à partir de l'estimateur de Kaplan-Meier (Kaplan et Meier, 1958) par :

$$\hat{S}(t) = \prod_{j:t_j \leq t} \frac{n_j - d_j}{n_j}$$

Le risque cumulée $\Lambda(t)$ peut être estimé par l'estimateur de Nelson-Aalen (Nelson , 1969)

$$\hat{\Lambda}(t) = \sum_{j:t_j \leq t} \frac{d_j}{n_j}$$

et la fonction de répartition par l'estimateur de Aalen-Johansen (voir section II.1.2.2).

Ces différents estimateurs sont asymptotiquement équivalents ; il est donc possible de passer d'une fonction à une autre très facilement (voir équation II.1). Par contre, ces estimateurs ne sont pas adaptés aux données censurées par intervalle car les estimateurs non-paramétriques requièrent un temps exact d'évènement pour être estimés. Il convient donc d'imputer un temps d'évènement ou de censurer à droite au dernier temps où le sujet est sain (T_l) si des estimateurs non-paramétriques sont appliqués à des données censurées par intervalle.

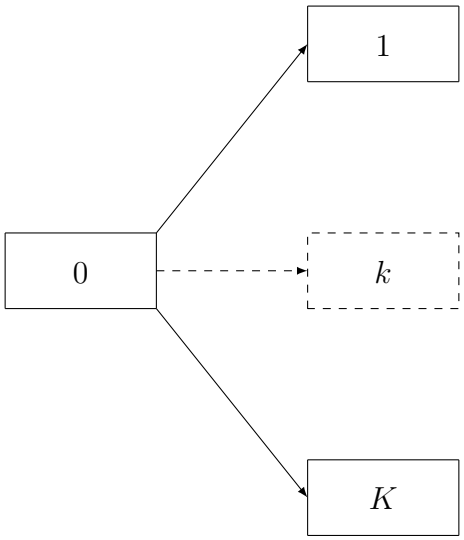

FIGURE II.3 – Représentation du modèle à risques compétitifs à K états absorbants

II.1.2.2 Risques compétitifs

Un autre type de modèle multi-états souvent utilisé lorsque plusieurs évènements rentrent en compétition à partir d'un même état est le modèle à risques compétitifs. Il sert par exemple à modéliser le temps jusqu'au décès dû à différentes causes, chaque cause étant un évènement en particulier. Ce modèle est présenté car il a servi de base à certains modèles de régression présentés dans la section II.3.

Dans un modèle à K risques compétitifs (voir figure II.3), le risque absolu de l'évènement k entre les temps s et t est défini par :

$$F_k(s,t) = \mathbb{P}(T \leq t, \delta = k \mid T \geq s) = \int_s^t \exp\left(-\sum_{k=1}^{K} \Lambda_k(s,u)\right) \lambda_k(u)\mathrm{d}u \tag{II.2}$$

avec λ_k et Λ_k l'intensité de transition et l'intensité de transition cumulée vers l'état k. Le risque absolu représente la probabilité d'avoir subi l'évènement k entre les temps s et t et peut être estimé non-paramétriquement par la formule d'Aalen-Johansen (Beyersmann et Scheike, 2014).

II.1.2.3 Modèle *illness-death*

Le modèle *illness-death* ou « sain-malade-mort » est un modèle multi-états qui tient son nom de l'application qui en est faite. Il est utilisé généralement pour modéliser simultanément le risque d'une maladie, le risque de décès sans maladie et le risque de décès avec maladie (Andersen *et al.*, 1993). Le modèle *illness-death* est défini à partir d'un processus stochastique $X = \{X(t), t \geq 0\}$ à valeurs dans $\{0, 1, 2\}$. Dans cette thèse, l'état 0 correspond à être non-dément, l'état 1 correspond à la démence et l'état 2 au décès. Le modèle *illness-death* ne permet pas de guérison, c'est-à-dire que les transitions kl possibles sont $kl \in \{01, 02, 12\}$, comme le montre la figure II.4. Les n participants sont tous non-déments à l'inclusion, c'est-à-dire que $X_i(0) = 0, \forall i = 1, ..., n$.

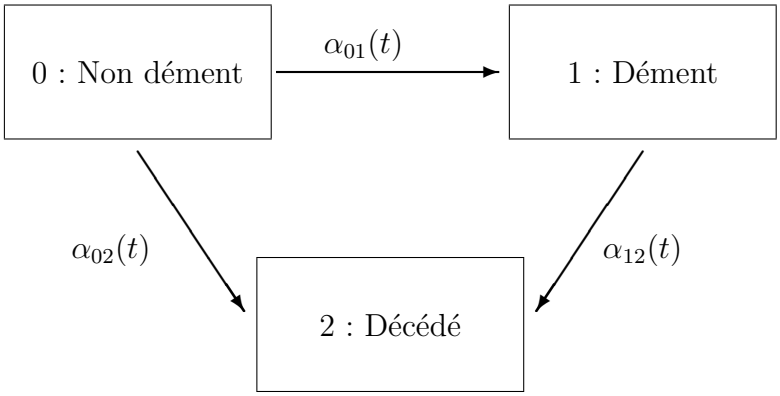

FIGURE II.4 – Représentation du modèle *illness-death*

Le modèle *illness-death* est caractérisé par ses intensités de transition $\alpha_{kl}(.)$. L'intensité de transition de l'état k à l'état l au temps t est définie par :

$$\alpha_{kl}(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}\left(X(t+\Delta t) = l \mid X(t) = k\right)}{\Delta t} \tag{II.3}$$

L'intensité de transition cumulée de l'état k à l'état l entre les temps s et t est définie par :

$$A_{kl}(s,t) = \int_s^t \alpha_{kl}(u)\mathrm{d}u \tag{II.4}$$

Remarque 1 *L'histoire d'un processus au temps s est notée par H_{s-} et la probabilité de transition d'un état k à un état l est défini par $P_{kl}(s,t) = \mathbb{P}\left(X(t) = l \mid X(s) = k, H_{s-}\right)$. Le modèle illness-death présenté ici est un modèle markovien non homogène : la probabilité $P_{kl}(s,t)$ ne tient pas compte de l'histoire du processus H_{s-} c'est-à-dire que le temps de transition ne dépend plus du passé mais seulement de s et de l'état au temps s, avec $P_{kl}(s,t) = \mathbb{P}\left(X(t) = l \mid X(s) = k\right)$. Cette hypothèse peut être assouplie par un modèle semi-Markov, qui suppose que le temps de transition dépend du temps présent et du temps de la dernière transition, avec dans ce cas $P_{kl}(s,t) = \mathbb{P}\left(X(t) = l \mid X(s) = k, t_k\right)$.*

II.2 Indicateurs épidémiologiques

Dans un but étiologique, les intensités de transition sont largement utilisées pour quantifier l'impact de variables sur une maladie. D'un point de vue de santé publique, d'autres quantités sont aussi intéressantes à étudier comme les différents indicateurs épidémiologiques développés dans cette section.

Les indicateurs épidémiologiques présentés dans cette thèse sont définis à partir des intensités de transition d'un modèle *illness-death* (Touraine *et al.* (2016) définissent les indicateurs suivants sauf la moyenne restreinte des temps de survie sans démence, adaptée de la moyenne restreinte des

temps de survie d'Andersen *et al.* (2004)).

II.2.1 Indicateurs épidémiologiques calculés pour un horizon fini

On entend par « horizon fini » tout indicateur calculé pour une plage de temps donnée, avec un temps de début noté s et un temps de fin noté t.

Le premier indicateur épidémiologique est le risque absolu de la démence $F_{01}(s,t)$, aussi appelé fonction d'incidence cumulée de la démence. Il est défini par :

$$F_{01}(s,t) = \int_s^t \exp\left[-A_{01}(s,u) - A_{02}(s,u)\right] \alpha_{01}(u)\mathrm{d}u \tag{II.5}$$

Le risque absolu de démence entre le temps s et le temps t représente la probabilité qu'un individu non-dément au temps s développe une démence avant le temps t.

Le deuxième indicateur est la probabilité d'être vivant non-dément (qui équivaut à la probabilité de rester dans l'état 0 du modèle *illness-death*), noté $P_{00}(s,t)$ avec :

$$P_{00}(s,t) = \exp\left[-A_{01}(s,t) - A_{02}(s,t)\right] \tag{II.6}$$

et représente la probabilité qu'un individu vivant non-dément au temps s soit toujours vivant non-dément au temps t.

Le dernier indicateur calculé pour un horizon fini est la moyenne restreinte des temps de survie sans démence, noté $RM(s,t)$ avec :

$$RM(s,t) = \int_s^t \exp\left[-A_{01}(s,u) - A_{02}(s,u)\right] \mathrm{d}u \tag{II.7}$$

La moyenne restreinte des temps de survie en bonne santé s'interprète de la manière suivante : pour un individu vivant non dément au temps s, c'est le temps qu'il peut espérer rester vivant non-dément entre s et t.

II.2.2 Indicateurs épidémiologiques pour un horizon infini

Ici, les indicateurs ne sont plus bornés entre deux temps s et t mais sont calculés à partir d'un temps s sans limite de période, c'est pourquoi la notation t disparait des notations de ces indicateurs. De plus, les indicateurs suivants ont une interprétation épidémiologique lorsque l'âge est utilisé comme temps de base.

Le risque de démence vie-entière d'un sujet non dément à l'âge s ($LTR(s)$) est défini par :

$$LTR(s) = \int_s^{+\infty} \exp\left[-A_{01}(s,u) - A_{02}(s,u)\right] \alpha_{01}(u)\mathrm{d}u \tag{II.8}$$

Cet indicateur s'interprète comme la probabilité de développer une démence jusqu'au décès pour un sujet non-dément et vivant d'âge s.

Les deux derniers indicateurs sur lesquels une attention a été portée sont deux espérances de vie. La première est l'espérance de vie sans démence d'un sujet non dément à l'âge s, notée par $LE_{00}(s)$ et se calcule par :

$$LE_{00}(s) = \int_s^{+\infty} \exp\left[-A_{01}(s,u) - A_{02}(s,u)\right] \mathrm{d}u \tag{II.9}$$

La seconde espérance de vie est l'espérance de vie (totale) d'un sujet non-dément à l'âge s, noté par $LE_{0.}(s)$ avec :

$$LE_{0.}(s) = \int_s^{+\infty} \Bigg(\exp\left[-A_{01}(s,u) - A_{02}(s,u)\right] + \int_s^u \exp\left[-A_{01}(s,v) - A_{02}(s,v)\right] \alpha_{01}(v) \exp\left[-A_{12}(v,u)\right] \mathrm{d}v \Bigg) \mathrm{d}u$$

L'espérance de vie (totale) $LE_{0.}$ s'interprètent comme le nombre d'années qu'un sujet non-dément à l'âge s peut espérer encore vivre et l'espérance de vie sans démence LE_{00} représente le nombre moyen d'années qui reste à un individu avant de développer une démence ou avant de décéder.

Remarque 2 *Il existe un parallèle entre les indicateurs pour un horizon fini et les indicateurs pour un horizon infini. En effet, $F_{01}(s,+\infty) = LTR(s)$ et $RM(s,+\infty) = LE_{00}(s)$. De plus, tous les indicateurs font intervenir $P_{00}()$ et cette probabilité dépend à la fois du risque de démence (avec A_{01}) et à la fois du risque de décès (avec A_{02}).*

Remarque 3 *Les indicateurs de santé détaillés ci-dessus sont des exemples d'application. Ils seront par la suite utilisés comme quantité sur lesquelles les modèles de régression s'appuieront. Il parait clair que d'autres indicateurs de santé ont aussi leur importance, comme par exemple la durée que peut espérer rester un sujet dans un état entre deux temps ou alors l'espérance de vie d'un sujet malade.*

II.3 Modèles de régression pour temps d'évènements

Les modèles de régression utilisés dans le cadre de l'analyse de temps d'évènement mesurent l'effet de variables explicatives sur une fonction des temps de transition entre états tout en tenant compte de la censure. L'intensité de transition (ou fonction de risque dans l'analyse de survie) est la variable à expliquer la plus utilisée dans les modèles de régression. D'autres modèles de régression utilisant la fonction d'incidence cumulée par exemple ou basés sur des approches moins conventionnelles seront aussi abordés dans cette partie.

II.3.1 Régression dans un modèle de survie

Les modèles de régression pour l'analyse de survie se basent essentiellement sur l'effet de variables explicatives sur la fonction de risque. Cette section détaillera en particulier les modèles à risques proportionnels et les modèles à risques additifs et évoquera d'autres exemples de modèles de régression.

II.3.1.1 Modèle à risques proportionnels

D. Cox a proposé en 1972 un modèle de régression pour la fonction de risque $\lambda(.)$. Il définit le modèle de régression par :

$$\lambda(t \mid Z_1, ..., Z_p) = \lambda_0(t) \exp(\beta_1 Z_1 + ... + \beta_p Z_p) \tag{II.10}$$

avec $\lambda_0(t)$ le risque de base (non estimé dans le modèle de Cox) et β le p-vecteur des coefficients de régression. Le rapport des fonctions de risques pour deux sujets i et j ayant les variables explicatives Z_i et Z_j est

$$\frac{\lambda(t \mid Z_i) = \lambda_0(t) \exp(\beta^\top Z_i)}{\lambda(t \mid Z_j) = \lambda_0(t) \exp(\beta^\top Z_j)} = \frac{\exp(\beta^\top Z_i)}{\exp(\beta^\top Z_j)} \tag{II.11}$$

La première hypothèse du modèle est donc la proportionnalité des risques car le rapport des fonctions relatives est constant au cours du temps. La seconde hypothèse est l'effet log-linéaire des variables continues, c'est-à-dire que $\exp(\beta)$ est le risque relatif pour une augmentation d'une unité de la variable quantitative. Ces deux hypothèses sont donc à prendre en compte lors de la modélisation des données. Nous verrons par la suite que d'autres modèles permettent de relâcher l'hypothèse de proportionnalité des risques.

II.3.1.2 Modèle à risques additifs

Une solution lorsque l'hypothèse de risque constant n'est pas réaliste consiste à l'utilisation d'un modèle de régression avec un effet dépendant du temps. Pour cela, le modèle additif d'Aalen peut être envisagé (Aalen , 1989). L'idée de base se rapproche du modèle de Cox mais en utilisant une forme additive pour modéliser l'effet des variables explicatives sur la fonction de risque.

$$\lambda(t \mid Z_1, ..., Z_p) = \beta_0(t) + \beta_1(t) Z_1(t) + ... + \beta_p(t) Z_p(t) \tag{II.12}$$

Lin et Ying ont proposé une restriction du modèle de Aalen pour des paramètres de régression qui ne dépendent pas du temps, excepté pour l'intercept (Lim et Zhang, 2009).

II.3.1.3 Extension à d'autres modèles

Il n'est pas présenté ici une liste exhaustive des modèles de régression pour l'analyse de survie. Dans des cas particuliers, d'autres modèles peuvent être utilisés. Les modèles de vie accélérée sont par exemple appropriés pour le domaine de la fiabilité, les modèles à fragilité ont été développés lorsqu'une corrélation entre individus existe. Des extensions du modèle ont aussi été proposées pour la prise en compte de variables dépendantes du temps. De même, pour certaines variables de ce type, il est souvent préférable d'utiliser un modèle conjoint.

II.3.1.4 Avantages et limites

L'avantage de la modélisation d'une des fonctions des temps d'évènements d'un modèle de survie est de pouvoir donner une estimation des paramètres de régression sur une autre fonction des temps d'évènements. Par exemple, si un modèle à risques proportionnels a été estimé, alors les paramètres de régression interviennent sur la fonction de survie telle que $S(t \mid Z) = S_0(t)^{\exp(\beta^\top Z)}$. Par contre, ce type de modélisation ne permet pas de prendre en compte un risque compétitif car la fonction de survie générale est définie en fonction de plusieurs transitions.

II.3.2 Modèle cause-spécifiques

Les différents modèles présentés en analyse de survie (c.à.d. dans un modèle à deux états et une transition) sont adaptables aux modèles multi-états avec plusieurs transitions. Les modèles sont dits « cause-spécifiques » lorsque la régression du modèle multi-états porte sur les intensités de transition, avec un modèle de régression par transition.

Par exemple, un modèle cause-spécifiques appliqué à un modèle *illness-death* intervient par :

$$\begin{cases} \alpha_{01}(t \mid Z) = \alpha_{01,0}(t) \exp(\beta_{01}^\top Z_{01}) \\ \alpha_{02}(t \mid Z) = \alpha_{02,0}(t) \exp(\beta_{02}^\top Z_{02}) \\ \alpha_{12}(t \mid Z) = \alpha_{12,0}(t) \exp(\beta_{12}^\top Z_{12}) \end{cases} \tag{II.13}$$

avec des variables explicatives qui peuvent être différentes sur chaque transition.

Les facteurs de risque ou les facteurs protecteurs de la démence sont maintenant connus et ont été largement étudiés au cours des dernières années (grâce notamment à l'estimation du vecteur β_{01} de l'équation (II.13) adaptée aux données censurées par intervalle, voir section II.4 pour plus de détails). Ces facteurs de risque représentent l'impact de variables sur les intensités de transition du modèle *illness-death*. D'un point de vue de santé publique, il est intéressant de regarder maintenant l'impact de ces déterminants sur d'autres quantités d'intérêt, appelés précédemment indicateurs épidémiologiques. Les différents modèles de régression présentés maintenant peuvent être une solution pour répondre à ce besoin de santé publique.

II.3.3 Modèle pour le risque absolu

Fine et Gray ont proposé en 1999 un modèle à risques proportionnels basé sur les risques de sous-répartition (aussi appelée risques de sous-distribution)(Fine et Gray, 1999). Le risque de sous-distribution $\lambda_k^*(t, Z)$ de l'évènement k ajusté sur les variables Z dans un modèle à risques compétitifs (comme la figure II.3) est défini par :

$$\lambda_k^*(t, Z) = -\frac{\mathrm{d}}{\mathrm{d}t} \log(1 - F_k(t, Z)) \tag{II.14}$$

avec $F_k(t, Z) = \mathbb{P}(X(\tau) = k, \tau \leq t \mid Z)$ le risque absolu de l'évènement k conditionnellement aux variables explicatives Z.

Ce modèle estime l'effet de variables explicatives sur le risque absolu d'un évènement par la fonction suivante :

$$\lambda_k^*(t, Z_i) = \lambda_{k0}^*(t) \exp(\beta_k^\top Z_i) \tag{II.15}$$

$$\log(-\log(1 - F_k(t, Z_i)) = \log(\lambda_{k0}^*(t)) + \beta_k^\top Z_i \tag{II.16}$$

L'inconvénient du modèle de Fine et Gray est son interprétation. En effet, le risque de base $\lambda_{k0}^*(.)$ considère les sujets ayant subi un évènement j tel que $j \neq k$ toujours à risque pour l'évènement k. L'interprétation des paramètres doit être vu d'un point de vue qualitatif (c'est-à-dire qu'une variable j augmente l'incidence d'un évènement si β_j est positif, mais on ne peut pas quantifier cette augmentation). A contrario, le modèle à risques proportionnels (de Cox) multiplie (ou divise) le risque de subir l'évènement par $\exp(\beta)$.

II.3.4 Régression binomiale

Une autre technique pouvant être envisagée pour regarder l'effet de variables sur un indicateur épidémiologique est l'utilisation des modèles de régression binomial pour l'analyse de temps d'évènement.

La régression binomiale s'applique pour étudier la variable réponse binaire à un temps précis. L'analyse de survie étudie le temps jusqu'à l'apparition d'un évènement. L'idée générale est alors de considérer l'analyse de survie comme une variable réponse à différents temps après l'origine comme une variable binaire ; qui vaut 1 si le sujet a subi l'évènement et 0 sinon (Grøn et Gerds, 2014).

Cette approche n'est pas plus détaillée car elle n'a pas été appliquée dans ce travail et reste peu utilisée.

II.3.5 Pseudo-valeurs

Andersen, Klein et Rosthøj ont proposé un modèle basé sur des estimateurs de type *jackknife* pour faire de l'inférence dans le cadre de l'analyse de temps d'évènements (Andersen *et al.*, 2003). La méthode est basée sur des pseudo-valeurs et permet d'estimer l'effet de variables explicatives sur des quantités d'intérêt différentes des intensités de transition.

L'approche par pseudo-valeurs est une méthode en deux étapes. À partir d'un processus stochastique $X(t)$, une quantité d'intérêt $\theta(t)$ est définie tel que $\theta(t) = \mathrm{E}\left[f(X(t)\right]$ avec $f(.)$ une fonction de transformation de $X(t)$. L'idée est d'estimer $\theta(t \mid Z) = \mathrm{E}\left[f(X(t)) \mid Z\right]$ pour des variables explicatives Z. Si les données sont totalement observées (c'est-à-dire qu'il n'y a pas de censure) alors il est possible d'estimer θ par la moyenne $\hat{\theta} = \frac{1}{n}\sum_i f(X_i(t))$, avec $i = 1, ..., n$.
Ici, la présence de la censure ne permet pas de calculer $\hat{\theta}$ de cette manière.Pour le sujet i, l'idée générale est de remplacer l'information incomplète de $f(X_i(t))$ par sa pseudo valeur Y_i au temps t :

$$Y_i(t) = n \times \hat{\theta}(t) - (n-1) \times \hat{\theta}^{-i}(t) \tag{II.17}$$

avec $\hat{\theta}(t)$ l'estimateur de la quantité d'intérêt calculé sur l'ensemble de l'échantillon et $\hat{\theta}^{-i}(t)$ l'estimateur calculé sur l'échantillon sans le sujet i. Ensuite, la seconde étape consiste à utiliser les différentes pseudo-valeurs comme variable réponse dans un modèle de régression :

$$g\left(Y_i(t)\right) = \gamma_0 + ... + \gamma_p Z_p + \varepsilon_i \tag{II.18}$$

avec $g(.)$ une fonction de lien et le vecteur γ estimé en général par GEE (Liang et Zeger, 1986).

Il est possible d'avoir un processus $Y_i(t)$ multivarié, c'est-à-dire que pour un individu, J différents temps sont choisis pour calculer les pseudo-valeurs $Y_i(t_j)$ avec $j = 1, ..., J$. Ici pour un sujet i, le J-vecteur des pseudo-valeurs est noté $Y_i = \left(Y_i(t_1), ...Y_i(t_J)\right)^\top$, le vecteur de taille $(p+1)$ des variables explicatives est noté $Z_i = \left(1, Z_{i1}, ..., Z_{ip}\right)^\top$. L'estimation des paramètres de régression par GEE se fait donc en résolvant les équations suivantes :

$$\begin{aligned} U(\gamma) &= \sum_i U_i(\gamma) \\ &= \sum_i \left(\frac{\mathrm{d}}{\mathrm{d}\gamma} g^{-1}(\gamma^\top Z_i)\right) V_i^{-1}\left(Y_i - g^{-1}(\gamma^\top Z_i)\right) = 0 \end{aligned} \tag{II.19}$$

avec V_i la matrice de variance-covariance de travail. La variance des estimations $\hat{\gamma}$, notée $\widehat{\mathrm{var}}(\hat{\gamma})$

est obtenue grâce à un estimateur sandwich tel que

$$
\begin{aligned}
I(\gamma) &= \sum_i \left(\frac{\mathrm{d}g^{-1}(\gamma^\top Z_i)}{\mathrm{d}\gamma}\right)^\top V_i^{-1} \left(\frac{\mathrm{d}g^{-1}(\gamma^\top Z_i)}{\mathrm{d}\gamma}\right) \\
\widehat{\mathrm{var}}\left(U(\hat{\gamma})\right) &= \sum_i U_i(\hat{\gamma})^\top U_i(\hat{\gamma}) \\
\widehat{\mathrm{var}}(\hat{\gamma}) &= I(\hat{\gamma})^{-1} \widehat{\mathrm{var}}\left(U(\hat{\gamma})\right) \left(I(\hat{\gamma})^{-1}\right)
\end{aligned} \tag{II.20}
$$

La littérature sur l'approche par pseudo-valeurs ne cesse d'augmenter. L'approche initiale d'Andersen *et al.* (2003) est basée sur des pseudo-valeurs calculées pour la probabilité d'être dans un état à partir d'un modèle *illness-death* (Andersen *et al.*, 2003). Ils utilisent pour cela l'estimateur non-paramétrique d'Aalen et Johansen des probabilités de transition. Les fonctions de liens utilisées pour $g(.)$ (voir équation II.18) sont la fonction logistique ($g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$), la fonction probit ($g(\pi)$ est le quantile d'ordre π de la loi normale centrée réduite) et la fonction complémentaire log-log (ou cloglog, $g(\pi) = \log\left[-\log(1-\pi)\right]$).

Un des avantages des pseudo-valeurs est de pouvoir proposer un modèle de régression qui estime l'effet de variables explicatives sur différentes quantités d'intérêt. D'autres auteurs se sont aussi intéressés aux probabilités de transition et aux probabilités d'être dans un état dans un modèle multi-états (Andersen et Klein, 2007; Andersen et Pohar Perme, 2008). L'approche par pseudo-valeurs a aussi été développée pour le risque absolu (Klein et Andersen, 2005), la survie attendue ajustée à la qualité de vie (Tunes-da-Silva et Klein, 2009), la survie pour un horizon fini (Klein *et al.*, 2007), l'espérance du temps dans un état (Grand et Putter, 2016) ou bien à la survie relative (Pavlič et Pohar Perme, 2019).

L'équation (II.18) permet d'estimer l'effet direct des variables explicatives sur la quantité d'intérêt θ. La fonction de lien détermine l'interprétation des effets γ. Ainsi, un lien complémentaire log-log, utilisé par exemple par Andersen *et al.* (2003); Klein et Andersen (2005); Andersen et Pohar Perme (2008) ou Zöller *et al.* (2016), permet une interprétation similaire à celle d'un modèle de Fine et Gray. L'utilisation d'une fonction de lien logistique, comme le font Andersen *et al.* (2003); Klein et Andersen (2005); Andersen et Pohar Perme (2008); Klein *et al.* (2007) permet une interprétation comme un rapport de côte. D'autres auteurs ont utilisé une fonction de lien log pour interpréter les effets d'un point de vue multiplicatif ou proportionnel (voir Andersen *et al.* (2004); Tunes-da-Silva et Klein (2009); Grand et Putter (2016)). La dernière fonction de lien utilisée dans la littérature est le lien identité, qui permet d'interpréter les paramètres γ comme une addition (ou une soustraction) sur l'échelle de la quantité d'intérêt (Andersen *et al.*, 2004; Klein , 2006; Tunes-da-Silva et Klein, 2009; Grand et Putter, 2016).

L'approche par pseudo-valeurs a été implémentée dans une macro SAS et un *package* R pour calculer des pseudo-valeurs pour le risque absolu, la moyenne restreinte des temps de survie et la fonction de survie (Klein *et al.*, 2008). Une fonction pour calculer les pseudo-valeurs du nombre d'années perdues a aussi été ajouté depuis 2012 (Pohar Perme et Gerster, 2012).

Les propriétés des pseudo-valeurs ont été étudiées comme la consistance et la normalité asymptotique des estimations des coefficients de régression pour le risque absolu (Graw *et al.*, 2009), les propriétés générales d'après un grand échantillon (Jacobsen et Martinussen, 2016) ainsi que les valeurs « attendues » des pseudo-valeurs (Andersen et Pohar Perme, 2010).
La censure est supposée indépendante de tout processus : cette hypothèse a d'abord été étudiée par Graw *et al.* (2009). Andersen et Pohar Perme (2010) ont proposé une alternative lorsque la censure dépend de variables discrètes. D'autres auteurs (Binder *et al.*, 2014; Overgaard *et al.*, 2019) ont eux aussi regardé le comportement des pseudo-valeurs dans le cadre de censure dépendante des variables explicatives. Ces deux articles ont proposé une correction du calcul des pseudo-valeurs pour éviter les biais des estimations (par exemple avec une technique de pondération par l'inverse de la probabilité de la censure). La qualité d'ajustement aux données a été évaluée par Pavlič *et al.* (2018). Les propriétés des variances issues de modèles s'appuyant sur des pseudo-valeurs ont été montrées par Overgaard *et al.* (2018).

Les derniers développements proposés pour l'approche par pseudo-valeurs ont été proposés pour les données groupées (Logan *et al.*, 2011), pour les variables explicatives dépendantes du temps (Nicolaie *et al.*, 2013) en se basant sur une approche par *landmark* ou bien pour des fractions de risques attribuables (von Cube *et al.*, 2019). Zöller *et al.* (2016) ont étendu les modèles de régression pour prendre en compte des effets dépendants du temps et Mogensen et Gerds (2013) se sont appuyés sur des forêts aléatoires pour calculer les pseudo-valeurs. Certains auteurs ont proposé des pseudo-valeurs pour le risque absolu en l'absence de la cause de l'évènement (c'est-à-dire que le temps d'évènement est connu, mais pas la cause, (Moreno-Betancur et Latouche, 2013; Do et Kim, 2017). Grand *et al.* (2018) ont étendu l'approche par pseudo-valeurs aux données tronquées à gauche. Les premiers travaux pour données censurées par intervalle ont été proposés par Kim et Kim (2016) et Do et Kim (2017), bien que dans leurs cas, tous les sujets malades ont été diagnostiqués (c'est-à-dire qu'il n'y avait pas de transition non-observée).

Pour résumer, l'approche par pseudo-valeurs est une technique qui prend en compte l'effet de variables explicatives sur différentes quantités d'intérêt (comme le risque absolu ou un temps passé dans un état) à partir du moment où un estimateur $\hat{\theta}$ est disponible pour la quantité d'intérêt. L'approche fournit aussi une estimation de la quantité d'intérêt de base (contrairement aux modèles à risques proportionnels cause-spécifiques ou au modèle de Fine et Gray, où le risque de base est en général non-spécifié). De plus, le calcul des pseudo-valeurs se fait marginalement. Ce point est un avantage car il limite les hypothèses de relation entre les variables explicatives et la quantité d'intérêt mais cela peut aussi être un inconvénient lorsque la censure dépend des variables explicatives. En effet, ce dernier point est une limite de l'approche car l'hypothèse d'indépendance de la censure n'est parfois pas réaliste. Un autre point positif est l'interprétation faite des modèles de régression, la fonction de lien choisie pour le modèle II.18 induit une interprétation plutôt additive ou multiplicative par exemple. Un des inconvénients est la non prise en compte de la censure par intervalle (avec non observation de toutes les transitions) pour le calcul des pseudo-valeurs.

II.4 Analyse des données censurées par intervalle

Les précédentes méthodes présentées dans les sections II.1, II.2 et II.3 ne tiennent pas compte de la censure par intervalle de manière optimale. Le cas le plus simple pour l'analyse de survie d'évènements censurés par intervalle est l'imputation d'un temps d'évènement. En général l'imputation se fait au milieu de l'intervalle ($\frac{T_l+T_r}{2}$) ou au moment du diagnostic de l'évènement (T_r) car un temps d'évènement exact est requis par les méthodes non-paramétriques. Cette option d'imputation entraine des biais (Law et Brookmeyer, 1992).

Dans cette partie, une présentation des méthodes pour gérer ce type de données est faite. Plusieurs auteurs ont proposé un modèle à risques proportionnels pour des données censurées par intervalles (Finkelstein , 1986; Alioum et Commenges, 1996; Joly *et al.*, 1998; Pan et Chappell, 2002). Mais Leffondré *et al.* (2013) ont montré qu'un modèle de survie pour données censurées par intervalle n'est pas approprié en présence d'évènements compétitifs tel que le décès. Il parait plus adapté d'utiliser un modèle *illness-death* pour ce type de données.

II.4.1 Modèle *illness-death* pour temps de maladie censurée par intervalle

Comme cela a été montré précédemment, lors de l'analyse de données de cohorte de la démence, le début de la pathologie n'est pas observé en temps continu. Le début de la maladie est donc censuré par intervalle entre le dernière date où le sujet a été vu sain (c.à.d. non-dément) et le temps de diagnostic de la démence. De plus, certains sujets décèdent sans diagnostic de démence. Les méthodes d'estimation d'un modèle *illness-death* avec censure par intervalle de la maladie doivent donc prendre en compte ces particularités mais surtout permettre de prendre en compte le fait que certains sujets ne soient pas diagnostiqués avant d'entrer dans un état absorbant, ici le décès (*cf.* figure II.1). Pour cela, deux méthodes d'estimation ont été utilisées et sont détaillées dans les paragraphes suivants.

Le modèle *illness-death* est caractérisé par ses intensités de transition. La méthode du maximum de vraisemblance permet d'estimer ce modèle. Ainsi, suivant si les sujets sont observés vivants ou décédés et malades/non malades, leur contribution au calcul de la vraisemblance n'est pas la même. Soient δ_1 et δ_2 les indicatrices d'évènements telles que $\delta_1 = 1$ si le sujet est diagnostiqué malade au cours du suivi et $\delta_2 = 1$ si le sujet est décédé avant la date de point, 0 sinon. La date de point correspond à la fin de l'étude ou à la dernière visite de suivi si la cohorte est toujours en cours. Soit $\tilde{T} = \min(T, C)$ avec T le temps de décès et C le temps de censure. Pour les sujets diagnostiqués déments, T_l et T_r correspondent respectivement aux bornes gauche et droite de la censure par intervalle. Pour un sujet non diagnostiqué dément, T_l correspond au dernier temps d'observation. De plus, le temps de début d'observation des sujets est noté T_e avec T_e qui peut être nul ou égale à un âge d'entrée dans la cohorte (pour une prise en compte de la troncature à gauche).

Pour rappel, la probabilité de transition de l'état k à l'état l entre les temps s et t est noté par

$P_{kl}(s,t)$.

II.4.1.1 Contribution des sujets aux calculs de la vraisemblance

La contribution des sujets au calcul de la vraisemblance dépend de l'observation ou non de la démence et du décès. Les différents schémas d'observation possibles sont représentés graphiquement par la figure II.5. La contribution du sujet i est notée $\mathcal{L}_i$ et le temps de dernière visite où le sujet est sain, le temps du diagnostic de la maladie, le temps de décès ou le temps de censure sont respectivement notés T_{li}, T_{ri}, T_i et C_i. De plus, soit $P_{kl}(s,t)$ la probabilité d'être dans l'état l au temps t sachant qu'on est dans l'état k au temps l. En particulier,

$$P_{00}(s,t) = \exp(-A_{01}(s,t) - A_{02}(s,t))$$

$$P_{11}(s,t) = \exp(-A_{12}(s,t))$$

avec $A_{kl}(s,t) = A_{kl}(0,t) - A_{kl}(0,s)$. Les différentes contributions s'écrivent ainsi (Joly *et al.*, 2002) :

cas 1 : si le sujet i est vivant en C_i (censuré à droite), non dément en T_{li}, avec statut de la démence non connu en C_i et une inclusion en T_{ei}

$$\mathcal{L}_i = P_{00}(T_{ei}, C_i) + P_{00}(T_{ei}, T_{li}) \int_{T_{li}}^{C_i} P_{00}(T_{li}, u)\alpha_{01}(u)P_{11}(u, C_i)\mathrm{d}u$$

cas 2 : si le sujet i est décédé en T_i, il était non dément en T_{li}, le statut de la démence en T_i n'est pas connu et il est inclus en T_{ei}

$$\mathcal{L}_i = P_{00}(T_{ei}, T_i)\alpha_{02}(T_i) + P_{00}(T_{ei}, T_{li}) \int_{T_{li}}^{T_i} P_{00}(T_{li}, u)\alpha_{01}(u)P_{11}(u, T_i)\alpha_{12}(T_i)\mathrm{d}u$$

cas 3 : si le sujet i a un début de démence censuré par intervalle (c'est-à-dire non-dément en T_{li} et diagnostiqué en T_{ri}) et est censuré à droite pour le décès C_i

$$\mathcal{L}_i = P_{00}(T_{ei}, T_{li}) \int_{T_{li}}^{T_{ri}} P_{00}(T_{li}, u)\alpha_{01}(u)P_{11}(u, C_i)\mathrm{d}u$$

cas 4 : si le sujet i a un début de démence censuré par intervalle (c.à.d. non-dément en T_{li} et diagnostiqué en T_{ri}) et est décédé en T_i

$$\mathcal{L}_i = P_{00}(T_{ei}, T_{li}) \int_{T_{li}}^{T_{ri}} P_{00}(T_{li}, u)\alpha_{01}(u)P_{11}(u, T_i)\alpha_{12}(T_i)\mathrm{d}u$$

II.4.1.2 Estimation par vraisemblance pénalisée

Une des solutions possibles pour estimer un modèle *illness-death* avec censure par intervalle de la maladie est d'utiliser une méthode semi-paramétrique par vraisemblance pénalisée.

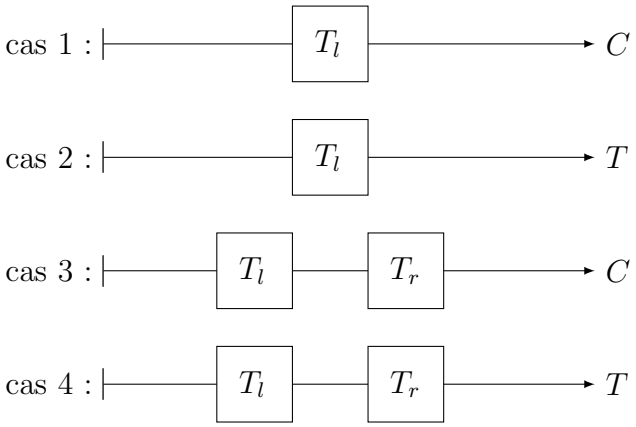

FIGURE II.5 – Représentation des différents types d'observation

Ainsi la vraisemblance pénalisée ($pl(\alpha_{01}, \alpha_{02}, \alpha_{12})$) du modèle s'écrit comme :

$$pl(\alpha_{01}, \alpha_{02}, \alpha_{12}) = l(\alpha_{01}, \alpha_{02}, \alpha_{12}) - \kappa_{01} \int \alpha_{01}^{''2}(u)\mathrm{d}u - \kappa_{02} \int \alpha_{02}^{''2}(u)\mathrm{d}u - \kappa_{12} \int \alpha_{12}^{''2}(u)\mathrm{d}u$$

avec κ_{01}, κ_{02} et κ_{12} les paramètres de lissage utilisés pour obtenir un compromis entre la régularité des fonctions et la fidélité aux données.

Le maximum de la vraisemblance pénalisée est défini à partir des intensités de transition du modèle. Ces intensités de transition sont approchées par des splines (voir Ramsay (1988) pour plus de détails). Des M-splines sont utilisées pour approcher les intensités de transition et des I-splines sont utilisées pour les intensités de transition cumulées. Les M-splines sont des B-splines normalisées et positives pour forcer les intensités de transition $\alpha(.)$ à être positives. Les I-splines sont des M-splines intégrées pour s'assurer de la monotonicité des intensités de transition cumulées $A(.)$. Les estimations se font par l'algorithme de Levenberg-Marquardt (Levenberg , 1944; Marquardt , 1963) implémenté dans le package R `SmoothHazard` (Touraine *et al.*, 2017).

II.4.1.3 Estimation paramétrique

Une autre solution pour estimer un modèle *illness-death* avec censure par intervalle de la maladie est d'utiliser une approche paramétrique. Par exemple, il est possible d'utiliser des lois de Weibull pur modéliser les intensités de transition. Le principal avantage de cette méthode est le nombre de paramètre à estimer : au total, 6 paramètres sont nécessaire pour estimer un modèle *illness-death*, avec un paramètre de forme et un paramètre d'échelle d'une loi de Weibull pour chacune des trois intensités de transition (et donc aucun paramètre de lissage n'a à être estimé/choisi).

Un estimation par maximum de vraisemblance permet d'obtenir des intensités de transition de la forme suivante :

$$\alpha_{kl}(t) = a_{kl} \left(\frac{1}{b_{kl}}\right)^{a_{kl}} t^{a_{kl}-1} \tag{II.21}$$

avec a_{kl} et b_{kl} les paramètres de forme et d'échelle pour la transition de l'état k à l'état l.

II.4.2 Inférence

Les intensités de transition du modèle *illness-death* peuvent aussi dépendre des variables explicatives. Trois propositions sont faites ici :

— un modèle à intensité de transition proportionnelle

$$\alpha_{kl}(t \mid Z_{kl}) = \alpha_{0,kl}(t) \times \exp\left[\beta_{kl}^{\top} Z_{kl}\right] \tag{II.22}$$

où l'intensité de transition kl est défini par rapport à un risque de base $\alpha_{0,kl}$ et l'effet des variables explicatives Z_{kl} est pris en compte par des paramètres de régression β_{kl}

— un modèle *illness-death* estimé avec stratification (dans le cas de variables explicatives qualitatives, avec dans ce cas là des intensités de transition différentes suivant les modalités des variables explicatives

$$\alpha_{kl}(t \mid Z_{kl}) = \alpha_{Z_{kl},kl}(t) \tag{II.23}$$

— un modèle *illness-death* utilisant à la fois de la stratification et à la fois la proportionnalité

$$\alpha_{kl}(t \mid Z_{kl}) = \alpha_{Z^{\bullet},0,kl}(t) \times \exp\left[\beta_{kl}^{\top} Z_{kl}^{\star}\right] \tag{II.24}$$

avec $Z^{\bullet}$ les variables utilisées pour stratifier (qui sont identiques pour les trois intensités du modèle *illness-death*) et $Z^{\star}$ les autres variables d'ajustement (qui peuvent être différentes suivant les transitions).

Les intensités de transition de base sont alors estimées par une méthode paramétrique ou une méthode semi-paramétrique.

Chapitre III

Approche par pseudo-valeurs pour temps de maladie censurés par intervalle

Ce chapitre présente l'extension de l'approche par pseudo-valeurs aux données censurées par intervalle à partir d'un modèle *illness-death*.

III.1 Méthode

La méthode pour l'approche par pseudo-valeurs a été développée pour estimer l'effet de facteurs de risque sur trois indicateurs épidémiologiques de la démence. Ces trois indicateurs sont le risque absolu de démence entre l'inclusion et le temps t ($F_{01}(0,t)$), la probabilité d'être toujours vivant non-dément au temps t ($P_{00}(0,t)$) et la moyenne restreinte des temps de survie sans démence jusqu'au temps t ($RM(0,t)$) à partir de l'inclusion. Ici, les indicateurs de santé ont été estimés en tenant compte de la censure par intervalle du début de la démence et du risque compétitif de décès. Le temps de base du modèle *illness-death* a été le délai depuis l'inclusion. La méthode s'est appuyée sur l'estimation d'un modèle *illness-death* par deux types d'estimation : une estimation semi-paramétrique (par maximisation d'une vraisemblance pénalisée et l'utilisation de splines pour approcher les intensités de transition, voir la partie II.4.1.2) et une estimation paramétrique (avec des distributions de Weibull, voir la partie II.4.1.3).

Pour rappel, la pseudo-valeur du sujet i au temps t est calculée par :

$$Y_i(t) = n \times \theta(t) - (n-1) \times \theta^{-i}(t) \tag{III.1}$$

avec $\theta^{-i}(t)$ l'estimateur calculé sur l'échantillon de taille $n-1$ auquel le sujet i a été ôté.

III.1.1 Estimation des intensités de transition

Remarque 4 *Dans ce chapitre, la notation des intensités de transition cumulées ou des indicateurs de santé ne fait plus apparaître la valeur de s car toutes ces quantités sont calculées depuis $s = 0$.*

Le vecteur des paramètres des estimations des intensités de transition d'un modèle *illness-death* :

— est composé des coefficients des splines lorsqu'une méthode semi-paramétrique est utilisée. Le vecteur a donc une taille de $(m_{01} + 2) + (m_{02} + 2) + (m_{12} + 2)$ avec m_{kl} le nombre de nœuds intérieurs pour la transition kl avec $kl \in \{01, 02, 12\}$.

— est composé de 6 paramètres lorsqu'une méthode semi-paramétrique est utilisée. Les intensités de transition suivent des distributions de Weibull avec un paramètre de forme et un paramètre d'échelle par transition.

L'étape préliminaire aux calculs des pseudo-valeurs consistent en l'estimation des trois intensités de transition du modèle *illness-death* à partir de l'échantillon avec les n sujets ainsi que les n modèles *illness-death* auquel un sujet différent est enlevé pour chaque sous-échantillon.

Les différentes pseudo-valeurs pour les trois quantités proposées ici peuvent s'exprimer en fonction des intensités de transition et des intensités de transition cumulées. Les différentes quantités définies précédemment et estimées à partir d'un sous-échantillon sans le sujet i sont notées : $\hat{F}_{01}^{-i}$, $\hat{P}_{00}^{-i}$, $\hat{RM}^{-i}$, $\hat{\alpha}_{01}^{-i}$, $\hat{A}_{01}^{-i}$ et $\hat{A}_{02}^{-i}$.

III.1.2 Pseudo-valeurs pour le risque absolu

Pour le risque absolu de démence t années après l'inclusion, la pseudo-valeur du sujet i se calcule suivant l'équation suivante :

$$\begin{aligned}
\hat{Y}_i(t) &= n \times \hat{F}_{01}(t) - (n-1) \times \hat{F}_{01}^{-i}(t) \\
&= n\left(\int_0^t \hat{P}_{00}(u)\hat{\alpha}_{01}(u)\mathrm{d}u\right) - (n-1)\left(\int_0^t \hat{P}_{00}^{-i}(u)\hat{\alpha}_{01}^{-i}(u)\mathrm{d}u\right) \\
&= n\left(\int_0^t \exp[-\hat{A}_{01}(u) - \hat{A}_{02}(u)]\hat{\alpha}_{01}(u)\mathrm{d}u\right) \\
&\quad -(n-1)\left(\int_0^t \exp[-\hat{A}_{01}^{-i}(u) - \hat{A}_{02}^{-i}(u)]\hat{\alpha}_{01}^{-i}(u)\mathrm{d}u\right)
\end{aligned}$$

III.1.3 Pseudo-valeurs pour la probabilité d'être vivant non-dément

La pseudo-valeur du sujet i calculée pour la probabilité d'être vivant non-dément au temps t est estimée par :

$$\begin{aligned}\hat{Y}_i(t) &= n \times \hat{P}_{00}(t) - (n-1) \times \hat{P}_{00}^{-i}(t) \\ &= n\left(\exp[-\hat{A}_{01}(t) - \hat{A}_{02}(t)]\right) - (n-1)\left(\exp[-\hat{A}_{01}^{-i}(t) - \hat{A}_{02}^{-i}(t)]\right)\end{aligned}$$

III.1.4 Pseudo-valeurs pour la moyenne restreinte des temps de survie en bonne santé

Le dernier indicateur pour lequel des pseudo-valeurs ont été calculées est la moyenne restreinte des temps de survie sans démence. La pseudo-valeur du sujet i au temps t est définie par :

$$\begin{aligned}\hat{Y}_i(t) &= n \times \widehat{RM}(t) - (n-1) \times \widehat{RM}^{-i}(t) \\ &= n\left(\int_0^t \hat{P}_{00}(u)\mathrm{d}u\right) - (n-1)\left(\int_0^t \hat{P}_{00}^{-i}(u)\mathrm{d}u\right) \\ &= n\left(\int_0^t \exp[-\hat{A}_{01}(u) - \hat{A}_{02}(u)]\mathrm{d}u\right) - (n-1)\left(\int_0^t \exp[-\hat{A}_{01}^{-i}(u) - \hat{A}_{02}^{-i}(u)])\mathrm{d}u\right)\end{aligned}$$

III.1.5 Modèles de régression

L'idée générale est d'estimer l'effet de variable explicative sur une quantité d'intérêt ajustée sur ces variables explicatives $\theta(t \mid Z)$ par un modèle linéaire généralisé (GLM). Les GLM estiment des effets linéaires des variables explicatives sur la variable d'intérêt.

Un GLM pour la quantité θ au temps t a été défini par l'équation :

$$g\left(\theta(t \mid Z_i)\right) \quad = \quad \gamma_{0,t} + \gamma_{1,t}Z_{1i} + \gamma_{2,t}Z_{2i} + ... + \gamma_{p,t}Z_{pi} \tag{III.2}$$

avec $g(.)$ une fonction de lien, $\gamma_{j,t}$ l'effet de la j^{e} variable explicative au temps t.

La présence de la censure ne permet pas d'estimer directement le modèle (III.2). L'approche remplace la quantité d'intérêt $\theta_i(t \mid Z_i)$ du sujet i par sa pseudo-valeur $Y_i(t)$.
Le modèle GLM utilisé pour estimer l'effet des variables explicatives sur les indicateurs épidémiologiques a été le suivant :

$$g\left(\hat{Y}_i(t)\right) \quad = \quad \gamma_{0,t} + \gamma_{1,t}Z_{1i} + \gamma_{2,t}Z_{2i} + ... + \gamma_{p,t}Z_{pi} \tag{III.3}$$

avec $g(.)$ des fonctions de lien identité ou log.

III.2 Étude de simulations

Les propriétés des pseudo-valeurs adaptées aux données censurées par intervalle d'un modèle *illness-death* ont été étudiées au travers d'une étude de simulations. Deux schémas de simulations ont été conduits et pour chaque schéma, trois scénarios ont été proposés.

III.2.1 Schéma A

III.2.1.1 Génération des données

Cinq cents jeux de données de 500 sujets chacun ont été générés. Une variable binaire Z a été simulée avec une distribution de Bernoulli telle que $\mathbb{P}(Z = 1) = 0,58$ pour mimer la répartition hommes/femmes dans la cohorte PAQUID.

Trois temps différents ont été créés pour simuler des observations d'un modèle *illness-death*. Les temps de démence, de décès sans démence et de décès avec démence ont été générés à partir de lois de Weibull avec un effet proportionnel de la variable Z sur les trois transitions tel que :

$$\alpha_{01}(t \mid Z) = \alpha_{01,0}(t)\exp(0,33 \times Z) \quad \text{(III.4)}$$

$$\alpha_{02}(t \mid Z) = \alpha_{02,0}(t)\exp(-0,58 \times Z), \quad \text{(III.5)}$$

$$\alpha_{12}(t \mid Z) = \alpha_{12,0}(t)\exp(-0,31 \times Z) \quad \text{(III.6)}$$

Les différentes valeurs pour les paramètres des lois de Weibull et les effets de Z sur chaque intensité de transition ont été choisies à partir de l'estimation (paramétrique) d'un modèle *illness-death* à risques proportionnels sur les données de la cohorte PAQUID. La variable Z avait donc un effet protecteur sur le décès mais délétère sur la démence. Les valeurs exactes des paramètres des distributions de Weibull sont présentées dans le tableau B.1 en page 104.

Trois scénarios différents ont été construits pour évaluer l'impact de la censure par intervalle :

— Scénario 1 : les données ont été générées en temps continu : T_0 et T ont été observés, avec une censure à droite à $t = 25$ et toutes les transitions ont été observées pour tous les individus.

— Scénario 2 : les données ont été observées en temps continu pour le décès et en temps discret pour la démence. Les temps de visite t_{vi} ont été générés aléatoirement pour chaque sujet i avec un temps minimum de 2 ans et un temps maximum de 3 ans entre deux visites consécutives. Ainsi, les temps de visites ont été définis par : $t_{vi+1} = t_{vi} + \mathcal{U}_i(2, 3)$ pour le sujet i. Les temps utilisés ont donc été $T_{li} = \max(t_{vi} \mid T_{0i} > t_{vi})$ et $T_{ri} = \min(t_{vi} \mid T_{0i} < t_{vi})$ avec les temps T_{0i} identiques dans les scénarios 1 et 2.

— Scénario 3 : les données ont été générées de la même manière que dans le scénario 2 en modifiant la longueur de l'intervalle entre deux visites consécutives. Dans ce cas, deux visites ont été espacées d'au moins 3 ans et d'au plus 6 ans (c.à.d. $t_{vi+1} = t_{vi} + \mathcal{U}_i(3, 6)$). Les temps

de visite utilisés ont suivi le même principe que pour le scénario 2 pour obtenir l'intervalle (T_l, T_r).

On rappelle que T_0 correspond au temps de sortie de l'état 0 et T au temps d'entrée dans l'état 2. Dans le cas où le sujet a transité directement de l'état 0 à l'état 2 (c'est-à-dire sans passé par l'état 1) alors $T_0 = T$, sinon $T_0 < T$. Pour les scénarios 2 et 3, une censure administrative à $t = 25$ a aussi été imposée de telle sorte que les temps de décès (ou de censure) soient identiques dans les trois scénarios. Le scénario 2 correspond aux suivis prévus dans les études de cohorte (avec une visite tous les deux ou trois ans) et le scénario 3 correspond à un scénario où un sujet serait amener à manquer une visite par exemple. La figure III.1 montre bien qu'entre les différents scénarios, certains individus ne sont pas toujours observés déments.

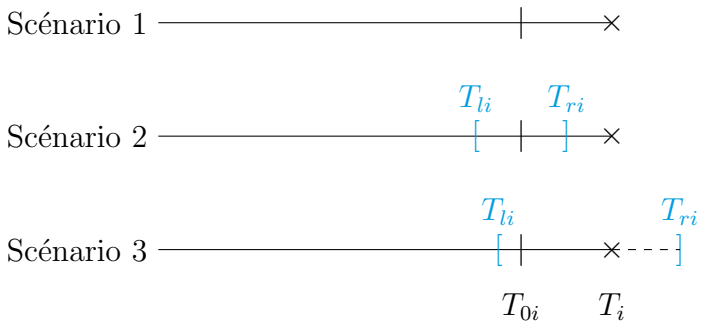

FIGURE III.1 – Exemple de simulation des temps d'évènements du sujet i suivant les trois scénarios du schéma A.

Le sujet i est dément au temps T_{0i} (représenté par |) et décède au temps T_i (représenté par $\times$). Les bornes gauche et droite de l'intervalle de censure sont décrites par les symboles [et]. Dans le scénario 2 le sujet i serait diagnostiqué dément au temps T_{ri} tout en sachant qu'il était sain au temps T_{li}. Par contre, dans le scénario 3, le sujet i a été vu vivant non-dément en T_{li} mais développe une démence et décède avant sa prochaine visite prévue en T_{ri}.

III.2.1.2 Procédure d'estimation

L'idée générale de ce schéma de simulations a été de comparer l'approche non-paramétrique d'Andersen *et al.* (2003) à une approche semi-paramétrique et une approche paramétrique. Les termes d'approche non-paramétrique, semi-paramétrique et paramétrique sont utilisés pour différencier la méthode utilisée pour estimer la quantité θ intervenant dans le calcul des pseudo-valeurs.

Pour le scénario 1 les approches paramétrique et semi-paramétrique ont été comparées à l'approche non-paramétrique, la référence, car les données ne sont pas censurées par intervalle.
Les résultats de l'approche semi-paramétrique entre les scénarios 2 et 3 ont été comparés aux résultats du scénario 1. De manière similaire, les résultats de l'approche paramétrique ont été comparés suivant les trois scénarios de générations des données.
Cette procédure permet de valider dans un premier temps les résultats des pseudo-valeurs en comparaison à la référence, l'approche non-paramétrique. Dans un second temps, cela permet de mesurer l'impact de la censure par intervalle sur le calcul des pseudo-valeurs, la censure par inter-

valle entraînant une perte d'information. Cette perte d'information est d'autant plus importante dans le scénario 3 que dans le scénario 2.

Pour l'approche semi-paramétrique, des splines ont été utilisés avec 5 nœuds intérieurs placés de manière équidistante sur le temps de suivi des sujets pour chaque transition. La recherche des paramètres de lissage κ est faite pour un échantillon de chaque scénario et les paramètres de lissage ont été conservés [1] pour les 499 autres jeux de données. Le choix des paramètres de lissage s'est fait en deux étapes : une première étape où κ est estimé par validation croisée puis un second temps où les valeurs des paramètres trouvées lors de l'étape 1 sont affinées manuellement grâce aux graphiques des intensités de transition.

L'effet de Z a été mesuré sur les trois indicateurs épidémiologiques (le risque absolu de démence, la probabilité d'être vivant non-dément et la moyenne restreinte des temps de survie sans démence) à différents temps ($t = 5$, $t = 10$, $t = 15$ et $t = 20$) par les équations suivantes :

$$\mathbb{E}\left[\hat{Y}_{i,\theta}(t)\right] = \gamma_{0,t,\theta} + \gamma_{1,t,\theta} Z_{1i} \tag{III.7}$$

$$\log\left(\mathbb{E}\left[\hat{Y}_{i,\theta}(t)\right]\right) = \gamma_{0,t,\theta} + \gamma_{1,t,\theta} Z_{1i} \tag{III.8}$$

avec θ correspondant à l'indicateur épidémiologique considéré.

Les paramètres de régression de l'équation (III.7) s'interprètent comme un ajout ou retrait de γ_j, t pour l'augmentation d'une unité de la variable j sur l'indicateur $\theta(t)$ alors que les paramètres de régression de l'équation (III.8) s'interprètent comme une multiplication de l'ordre de $\exp(\gamma_j)$ de $\theta(t)$ pour l'augmentation d'une unité de la variable j.

Les estimations $\hat{\gamma}$ obtenues à partir des différents scénarios et méthodes d'estimations ont été comparées en terme de moyenne des estimations, de biais, de RMSE (racine carrée de l'erreur quadratique moyenne) et du taux de couverture de l'intervalle de confiance à 95 %. De plus, les écart-types asymptotiques et empiriques sont présentées. Comme 500 jeux de données indépendants les uns des autres ont été générés, un taux de couverture a été considéré comme bon à partir du moment où il est compris dans l'intervalle $[93,1; 96,9]$.

III.2.1.3 Calcul des valeurs théoriques

Les paramètres choisis pour simuler les données interviennent tous sur les intensités de transition (voir équations (III.4) à (III.6)). En conséquence, l'influence de Z sur les indicateurs épidémiologiques définis en section II.2 n'est pas directement simulé. Par contre, il est possible de calculer les indicateurs épidémiologiques conditionnellement aux valeurs des variables explicatives (voir annexe A.1 pour le détail des calculs).

Soient $\theta(t \mid Z = 0)$ et $\theta(t \mid Z = 1)$ un indicateur épidémiologique pour deux sujets ayant

1. Les paramètres de lissages ont été conservés sauf si le modèle *illness-death* n'a pas convergé. Dans ce cas là, une recherche par validation croisée a été faite spécifiquement pour cet échantillon.

respectivement leur variable explicative égale à 0 et égale à 1. Ainsi, pour le modèle III.7, $\gamma_{0,t,\theta} = \theta(t \mid Z = 0)$ et $\gamma_{1,t,\theta} = \theta(t \mid Z = 1) - \theta(t \mid Z = 0)$. Les vraies valeurs pour le modèle III.8 sont $\gamma_{0,t,\theta} = \log(\theta(t \mid Z = 0))$ et $\gamma_{1,t,\theta} = \log\left(\frac{\theta(t|Z=1)}{\theta(t|Z=0)}\right)$ avec θ la fonction de risque absolu de la démence ou la probabilité d'être vivant non dément ou la moyenne restreinte des temps de survie sans démence.

III.2.1.4 Résultats

III.2.1.4.1 Descriptifs des échantillons En moyenne sur les 500 échantillons simulés, la variable explicative Z valait 1 pour 58% des sujets avec un écart type de 0,49 % pour la distribution de $\mathbb{P}(Z = 1)$ sur les 500 échantillons. Dans le scénario 1 (sans censure par intervalle du temps de démence), 50% des sujets sont devenus déments entre $t = 0$ et $t = 25$ (au minimum 42% et au maximum 61% sur les 500 échantillons) et les sujets sont décédés avant $t = 25$ pour 90% des sujets en moyenne (le minimum et le maximum sont respectivement de 85% et 93%.). Dans le scénario 2, à $t = 25$, en moyenne sur les 500 échantillons 24% des sujets ayant développé une démence n'ont pas été diagnostiqués. Pour le scénario 3, ce chiffre passe à 40% des individus déments qui n'ont pas été observés déments (c.à.d. qui ont développé la démence et sont décédés entre deux visites de suivi).

III.2.1.4.2 Commentaires des résultats

Remarque 5 *Par la suite, la méthode d'estimation des pseudo-valeurs sera un raccourci pour désigner la méthode d'estimation des quantités θ et ne change pas le calcul de la pseudo-valeur. De plus, une méthode dite « semi-paramétrique » sera équivalente à une méthode de maximisation du maximum de vraisemblance pénalisée, elle-même synonyme de méthode « splines ». Pareillement, les méthodes paramétriques feront référence à des estimateurs calculés à partir de lois de Weibull.*

Les tableaux III.1 à III.3 présentent respectivement l'ensemble des résultats des modèles (III.7) pour les trois indicateurs épidémiologiques ; le risque absolu de démence, la probabilité d'être vivant non-dément et la moyenne restreinte des temps de survie sans démence après 10 ans de suivi.

Pour le risque absolu de démence à 10 ans après l'inclusion dans la cohorte ($F_{01}(10)$), et pour le scénario 1 (celui sans données censurées par intervalle), les biais sont faibles pour les deux paramètres et les trois méthodes d'estimations des pseudo-valeurs. D'après le tableau III.1, la RMSE est plus faible pour la méthode paramétrique, suivie par la méthode semi-paramétrique. Les résultats par la méthode non-paramétrique sont les moins bons avec des taux de couverture de 91,6% pour les deux paramètres et des biais faibles.
Pour la méthode semi-paramétrique, si une comparaison de l'effet de la censure par intervalle est

Tableau III.1 – Résultats des simulations : comparaison de l'approche par pseudo-valeurs pour le risque absolu de démence à 10 ans de suivi($F_{01}(10)$) suivant trois méthodes d'estimation et trois scénarios. Schéma A : 500 échantillons de 500 sujets.

$\gamma_{.,10,F_{01}}$	Scénario et méthode	$\overline{\hat{\gamma}}$	Biais	ASE	ESE	RMSE × 1000	Taux de couverture
0,193	1- np	0,191	-0,002	0,027	0,028	29	91,6
	1- splines	0,192	-0,002	0,026	0,028	28	92,2
	1- Weibull	0,192	-0,001	0,024	0,026	26	94,0
	2- splines	0,193	0,000	0,034	0,033	33	94,2
	2- Weibull	0,191	-0,002	0,032	0,032	32	94,4
	3- splines	0,197	0,004	0,043	0,046	46	93,0
	3- Weibull	0,195	0,002	0,047	0,142	142	93,2
0,096	1- np	0,100	0,004	0,038	0,042	42	91,6
	1- splines	0,100	0,004	0,036	0,040	40	91,8
	1- Weibull	0,097	0,001	0,033	0,037	37	91,2
	2- splines	0,100	0,004	0,047	0,047	47	94,6
	2- Weibull	0,097	0,002	0,045	0,045	45	95,2
	3- splines	0,093	-0,003	0,059	0,062	62	94,2
	3- Weibull	0,093	-0,003	0,062	0,091	91	94,4

scénario 1 sans censure par intervalle

scénario 2 avec censure par intervalle et 2,5 ans entre deux visites en moyennes

scénario 3 avec censure par intervalle et 4,5 ans entre deux visites en moyennes

np = non-paramétrique

$\overline{\hat{\gamma}}$ = moyenne des estimations

ASE = écart-type asymptotique / ESE = écart-type empirique

RMSE = racine carrée de l'erreur quadratique moyenne

faite (c.à.d. une comparaison entre les scénarios 1, 2 et 3) alors le biais augmente lorsque la longueur de la censure par intervalle augment (aussi visible par une augmentation de la RMSE). Les biais restent faibles et les taux de couvertures sont compris dans l'intervalle $[93,1; 96,9]$ et sont donc bons.

L'impact de la censure par intervalle sur les pseudo-valeurs issues d'estimateurs paramétriques est la plus visible dans le scénario 3. En effet, les écarts-types asymptotiques et empiriques sont éloignées, ce qui peut s'expliquer par un biais d'estimation de certains échantillons (qui font donc augmenter l'ESE) pour l'intercept ($\gamma_{0,10,F_{01}}$). Ce résultat se retrouve aussi pour le deuxième paramètre. Les taux de couvertures sont bons pour les trois scénarios et les deux paramètres ($\gamma_{1,10,F_{01}}$ et $\gamma_{2,10,F_{01}}$) pour la méthode paramétrique.

Tableau III.2 – Résultats des simulations : comparaison de l'approche par pseudo-valeurs pour la probabilité d'être vivant non-dément 10 ans après l'inclusion ($P_{00}(10)$) suivant trois méthodes d'estimation et trois scénarios. Schéma A : 500 échantillons de 500 sujets.

$\gamma_{.,10,P_{00}}$	Scénario et méthode	$\overline{\hat{\gamma}}$	Biais	ASE	ESE	RMSE × 1000	Taux de couverture
0,434	1- np	0,433	-0,001	0,034	0,033	33	95,6
	1- splines	0,433	-0,000	0,032	0,032	32	96,0
	1- Weibull	0,432	-0,001	0,028	0,028	28	95,0
	2- splines	0,433	-0,001	0,032	0,032	32	95,4
	2- Weibull	0,433	-0,001	0,028	0,028	28	95,2
	3- splines	0,434	-0,000	0,032	0,032	32	95,6
	3- Weibull	0,433	-0,001	0,030	0,039	39	95,4
0,055	1- np	0,053	-0,002	0,045	0,049	49	93,0
	1- splines	0,052	-0,003	0,042	0,046	46	93,6
	1- Weibull	0,054	-0,001	0,037	0,039	39	93,6
	2- splines	0,053	-0,002	0,042	0,045	45	94,4
	2- Weibull	0,054	-0,001	0,037	0,040	40	93,0
	3- splines	0,053	-0,002	0,041	0,045	45	94,6
	3- Weibull	0,054	-0,001	0,039	0,045	45	94,4

scénario 1 sans censure par intervalle
scénario 2 avec censure par intervalle et 2,5 ans entre deux visites en moyennes
scénario 3 avec censure par intervalle et 4,5 ans entre deux visites en moyennes
np = non-paramétrique
$\overline{\hat{\gamma}}$ = moyenne des estimations
ASE = écart-type asymptotique / ESE = écart-type empirique
RMSE = racine carrée de l'erreur quadratique moyenne

Pour la probabilité d'être vivant non-dément 10 ans après le début du suivi, les écart-types empiriques et asymptotiques sont équivalents, sauf pour $\hat{\gamma}_{0,10,P_{00}}$ estimé par la méthode paramétrique sur les données du scénario 3. Lorsqu'il n'y a pas de censure par intervalle (c.à.d. pour le scénario 1), les résultats sont sensiblement les mêmes pour le biais de l'intercept et la méthode paramétrique a donné des biais plus faibles pour $\hat{\gamma}_{1,10,P_{00}}$. Pour les deux paramètres, la RMSE la plus faible est

obtenue par la méthode paramétrique et le meilleur taux de couverture (c.à.d. le plus proche de la valeur nominale) est donné par la méthode paramétrique.
Pour la comparaison des scénarios 1 à 3 par la méthode des splines et pour l'intercept, le biais est identique dans les scénarios 2 et 3, la RMSE est identique dans les trois scénarios et le taux de couverture le plus proche de la valeur nominale est trouvé pour le scénario 3 (voir tableau III.2). Pour le second paramètre ($\hat{\gamma}_{1,10,P}$), le biais diminue au fur et à mesure que la taille de l'intervalle de censure augmente, ce qui se traduit par des meilleurs taux de couvertures. Les RMSE sont identiques dans les scénarios 2 et 3 (c.à.d. les scénarios qui ont été générés avec de la censure par intervalle).
La méthode paramétrique donne des biais identiques sur les trois scénarios et les deux paramètres. Pour l'intercept, l'ESE est beaucoup plus grand pour le scénario 3, ce qui augmente la RMSE. Ceci peut être expliqué par une mauvaise estimation sur certains échantillons (bien que leur nombre reste mineur). Ce point a déjà été évoqué pour le risque absolu dans le tableau III.1. Le taux de couverture n'est pas affecté par cela. Les commentaires pour l'intercept ($\hat{\gamma}_{0,10,P_{00}}$) s'appliquent aussi au deuxième paramètre $\hat{\gamma}_{1,10,P_{00}}$ dans une moindre mesure. Le tableau III.2 montre des biais toujours négatifs.

Tableau III.3 – Résultats des simulations : comparaison de l'approche par pseudo-valeurs pour la moyenne restreinte des temps de survie 10 ans après l'inclusion ($RM(10)$) suivant trois méthodes d'estimation et trois scénarios. Schéma A : 500 échantillons de 500 sujets.

$\gamma_{.,10,RM}$	Scénario et méthode	$\overline{\hat{\gamma}}$	Biais	ASE	ESE	RMSE × 1000	Taux de couverture
7,255	1- np	7,251	-0,004	0,218	0,225	225	94,8
	1- splines	7,252	-0,003	0,264	0,293	293	91,6
	1- Weibull	7,249	-0,006	0,209	0,211	211	95,4
	2- splines	7,253	-0,002	0,219	0,224	224	94,4
	2- Weibull	7,253	-0,002	0,212	0,212	212	95,6
	3- splines	7,250	-0,005	0,225	0,235	235	94,6
	3- Weibull	7,246	-0,009	0,230	0,422	422	94,6
0,417	1- np	0,410	-0,007	0,279	0,299	299	93,2
	1- splines	0,378	-0,039	0,344	0,383	385	93,8
	1- Weibull	0,410	-0,006	0,268	0,280	280	94,2
	2- splines	0,406	-0,011	0,280	0,299	299	93,4
	2- Weibull	0,408	-0,009	0,272	0,285	285	94,6
	3- splines	0,406	-0,011	0,289	0,313	314	94,8
	3- Weibull	0,400	-0,017	0,295	0,363	363	93,6

scénario 1 sans censure par intervalle
scénario 2 avec censure par intervalle et 2,5 ans entre deux visites en moyennes
scénario 3 avec censure par intervalle et 4,5 ans entre deux visites en moyennes
np = non-paramétrique
$\overline{\hat{\gamma}}$ = moyenne des estimations
ASE = écart-type asymptotique / ESE = écart-type empirique
RMSE = racine carrée de l'erreur quadratique moyenne

Le tableau III.3 montre des RMSE qui sont bien plus grandes que dans les tableaux III.1 et III.2 : une explication possible réside dans l'échelle des valeurs des paramètres. La moyenne restreinte des temps de survie sans démence s'interprète comme des années alors que les deux autres indicateurs sont des probabilités. Pour cet indicateur épidémiologique les biais sont aussi tous négatifs (comme pour P_{00}).
Pour la moyenne restreinte des temps de survie sans démence, lorsqu'il n'y a pas de censure par intervalle (scénario 1), les méthodes non-paramétrique et paramétrique donnent des meilleurs résultats que l'approche par vraisemblance pénalisée. Ainsi, pour l'intercept, le biais est plus faible mais le taux de couverture n'est pas bon car les ASE sont trop faibles. Pour le deuxième paramètre $\hat{\gamma}_{1,10,R}$, le biais et la RMSE les plus élevés sont retrouvés pour la méthode semi-paramétrique et le taux de couverture est quant à lui bon. Lorsque la méthode d'estimation est semi-paramétrique, les résultats sont meilleurs en présence de censure par intervalle (c.à.d. scénarios 2 et 3) que sans censure par intervalle. La censure par intervalle n'affecte que les résultats du scénario 3 pour la méthode paramétrique, avec une augmentation des biais et de la RMSE.

Les résultats pour le lien logarithmique sont présentés dans tableau D.1 en annexe. Les résultats présentés sont similaires au lien identité : les biais sont légèrement plus importants pour le risque absolu de démence. Les écart-types empiriques sont à peu près égaux aux écart-types asymptotiques. Les taux de couverture sont bons dans leur ensemble.

III.2.2 Schéma B

III.2.2.1 Génération des données

Le schéma B est constitué de 50 réplications de 3200 sujets chacune. Les sujets avaient des temps d'évènements différents suivant la valeur d'une variable quantitative Z. Cette variable a été créée aléatoirement suivant une loi exponentielle de paramètre $0,1$ telle que $Z = \min(65 + \exp(0,1); 100)$ pour représenter un âge d'entrée dans une cohorte. La variable a ensuite été centrée et réduite avec $Z^* = \frac{Z-75}{5}$. La méthode de génération des trois scénarios du schéma A a aussi été appliquée aux trois scénarios du schéma B.

III.2.2.2 Procédure d'estimation

La méthode semi-paramétrique pour le calcul des pseudo-valeurs a été appliquée à chaque jeu de données pour les trois scénarios. Sur chaque intensité de transition, 5 nœuds internes ont été positionnés de manière équidistante sur la distribution des temps d'évènements. La recherche du paramètre de lissage s'est fait sur un seul échantillon par validation graphique pour chaque scénario. Les mêmes paramètres de lissage κ ont été repris pour les 49 autres réplications. Après le calcul des pseudo-valeurs, le modèle de régression qui estime l'effet de Z^* sur les trois indicateurs

épidémiologiques a été le modèle suivant :

$$\mathbb{E}\left[\hat{Y}_{i,\theta}(t)\right] = \gamma_{0,t,\theta} + \gamma_{1,t,\theta}Z^* + \gamma_{2,t,\theta}Z^{*2} + \gamma_{3,t,\theta}Z^{*3} \tag{III.9}$$

Ce modèle permet une modélisation plus souple de l'effet de Z^* sur l'indicateur épidémiologique.

III.2.2.3 Calculs des valeurs théoriques

Dans ce schéma, un modèle cubique a été utilisé pour estimer l'effet de Z^* sur les trois indicateurs épidémiologiques. Comme l'effet de Z^* dépend de la valeur de Z^*, il n'a pas été possible de calculer une valeur théorique comme dans le schéma A. Un jeu de données de 50 000 sujets a été généré **sans** censure (ni censure à droite et ni censure par intervalle). À partir de cet échantillon de grande taille, il a été possible de calculer les pseudo-valeurs correspondantes telles que :

— pour le risque absolu de la démence au temps t, la pseudo-valeur du sujet i vaut :

$$\begin{cases} 1 & \text{si } T_{0i} \leq t \text{ et } d_{1i} = 1 \\ 0 & \text{sinon} \end{cases}$$

— pour la probabilité d'être vivant non dément, la pseudo-valeur du sujet i vaut :

$$\begin{cases} 1 & \text{si } T_{0i} > t \\ 0 & \text{sinon} \end{cases}$$

— la pseudo-valeur de la moyenne restreinte des temps de survie en bonne santé vaut du sujet i est égale à :

$$\begin{cases} T_{0i} & \text{si } T_{0i} \leq t \\ t & \text{sinon} \end{cases}$$

Les « pseudo-valeurs » ont ensuite été utilisées comme variable d'intérêt d'un modèle cubique. La courbe d'effet obtenue par des splines a été utilisée comme référence.

III.2.2.4 Résultats

III.2.2.4.1 Descriptif des échantillons simulés La variable Z^* avait une valeur moyenne de 1,94 points (écart-type de 1,77) sur les 50 échantillons simulés. En ce qui concerne l'impact de la censure par intervalle

— 46% des sujets sont déments et 88% des sujets sont décédés à $t = 25$ dans le scénario 1 (les minimums et maximums vont de 43% à 48% pour la démence et de 87% à 89% pour le décès).

— 32% des sujets ont été observés déments en moyenne (31% à 34% pour les 50 échantillons). En fait, 24% de tous les sujets ayant développés une démence au cours du suivi n'ont pas pu être observés déments (car ils ont développés la maladie et décédés entre deux visites de suivi)

— 25% des sujets ont été diagnostiqués déments en moyenne sur les 50 échantillons (avec un minimum de 24% et un maximum de 27%). La perte d'information est encore plus importante ici : 41% des sujets déments n'ont pas été observés dans l'état 1 du modèle *illness-death*.

III.2.2.4.2 Commentaires des résultats

Les résultats du scénario B sont présentés par des graphiques (voir les figures III.2 à III.4 sur les pages 52 à 54). Les courbes noires en trait épais représentent la courbe de référence et les courbes de couleurs correspondent aux estimations sur chaque échantillon.

Pour le risque absolu de démence à $t = 10$, la référence montre une courbe qui augmente jusqu'à $Z^* = 2$ pour se stabiliser par la suite (avec une légère diminution à partir de $Z^* = 4$). En d'autres termes, si l'on compare un sujet $Z^* = 3$ par rapport à un sujet $Z^* = 0$, son risque absolu de démence à $t = 10$ est augmenté de 0,2. L'impact de la censure par intervalle est de plus en plus important pour cet indicateur épidémiologique : sans censure par intervalle (voir le graphique du haut de la figure III.2) les estimations sont proches des valeurs de référence. En présence de censure par intervalle avec un espace de deux à trois ans entre la dernière visite sans démence et le diagnostic de démence (scénario 2 - voir graphique central de la figure III.2), les estimations sont bonnes jusqu'à $Z^* = 2$ puis décrochent légèrement des valeurs théoriques.
Lorsque la censure par intervalle est plus marquée (avec le scénario 3), la majorité des courbes estimées ont une tendance similaire à la courbe théoriques pour $0 \leq Z^* \leq 3$. À partir de $Z^* = 3$, les estimations sont plus éloignées de la courbe théoriques.

Pour la probabilité d'être vivant non-dément 10 ans après le début du suivi, l'impact de la censure par intervalle n'est pas visible sur les différents scénarios de simulation, comme le montrent les trois graphiques de la III.3. Ces courbes montrent que plus Z^* augmente, plus la probabilité d'être vivant non-dément après 10 ans d'inclusion diminue. L'effet de Z^* est environ égale à $-0,4$ pour des valeurs de Z^* comprises entre 3 et 5 : l'effet semble se stabiliser.

D'après la figure III.4, les résultats pour la moyenne restreinte des temps de survie sans démence sont sensiblement les mêmes que pour la probabilité d'être vivant non-dément à 10 ans après l'inclusion. Dans le scénario 1 (celui sans données censurées par intervalle) pour 2 échantillons, l'effet de Z^* est mal estimé pour des valeurs de Z^* supérieures à 4.

III.3 Application

L'approche par pseudo-valeurs a été appliquée aux données de la cohorte PAQUID avec pour objectif d'étudier l'impact du score au test du MMS sur le risque absolu de démence à 10 ans après l'inclusion, la probabilité d'être vivant non-dément 10 ans après l'inclusion et la moyenne restreinte

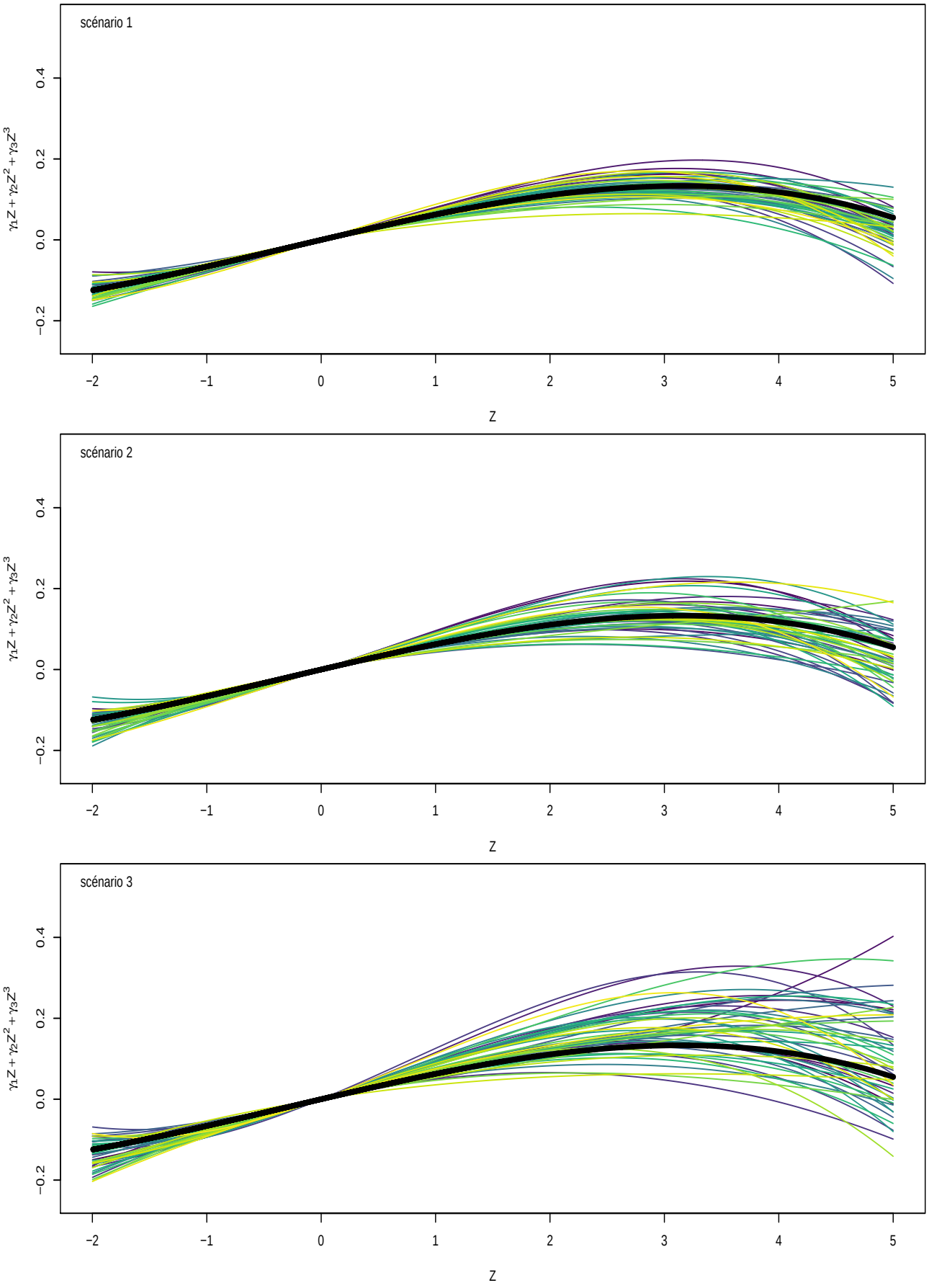

FIGURE III.2 – Résultats des simulations : comparaison de l'approche par pseudo-valeurs pour le risque absolu de démence 10 ans après une inclusion ($F_{01}(10)$) suivant trois scénarios (de haut en bas : scénario 1 sans censure par intervalle, scénario 2 avec censure par intervalle avec une visite tous les 2 à 3 ans et scénario 3 avec censure par intervalle avec une visite tous les 3 à 6 ans). Schéma B : 50 répliques de 3200 sujets.

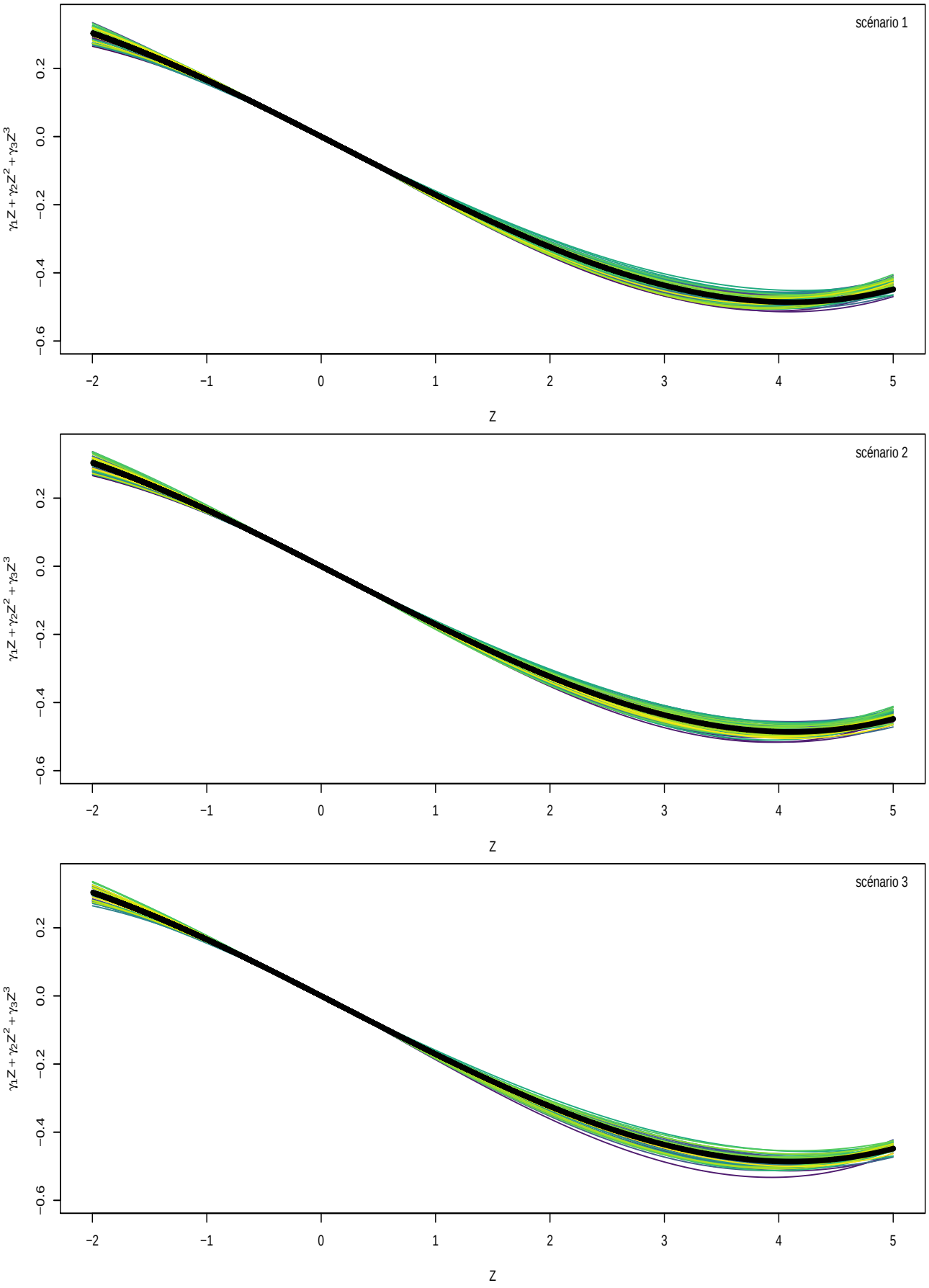

FIGURE III.3 – Résultats des simulations : comparaison de l'approche par pseudo-valeurs pour la probabilité d'être vivant non-dément 10 ans après une inclusion ($P_{00}(10)$) suivant trois scénarios (de haut en bas : scénario 1 sans censure par intervalle, scénario 2 avec censure par intervalle avec une visite tous les 2 à 3 ans et scénario 3 avec censure par intervalle avec une visite tous les 3 à 6 ans). Schéma B : 50 répliques de 3200 sujets.

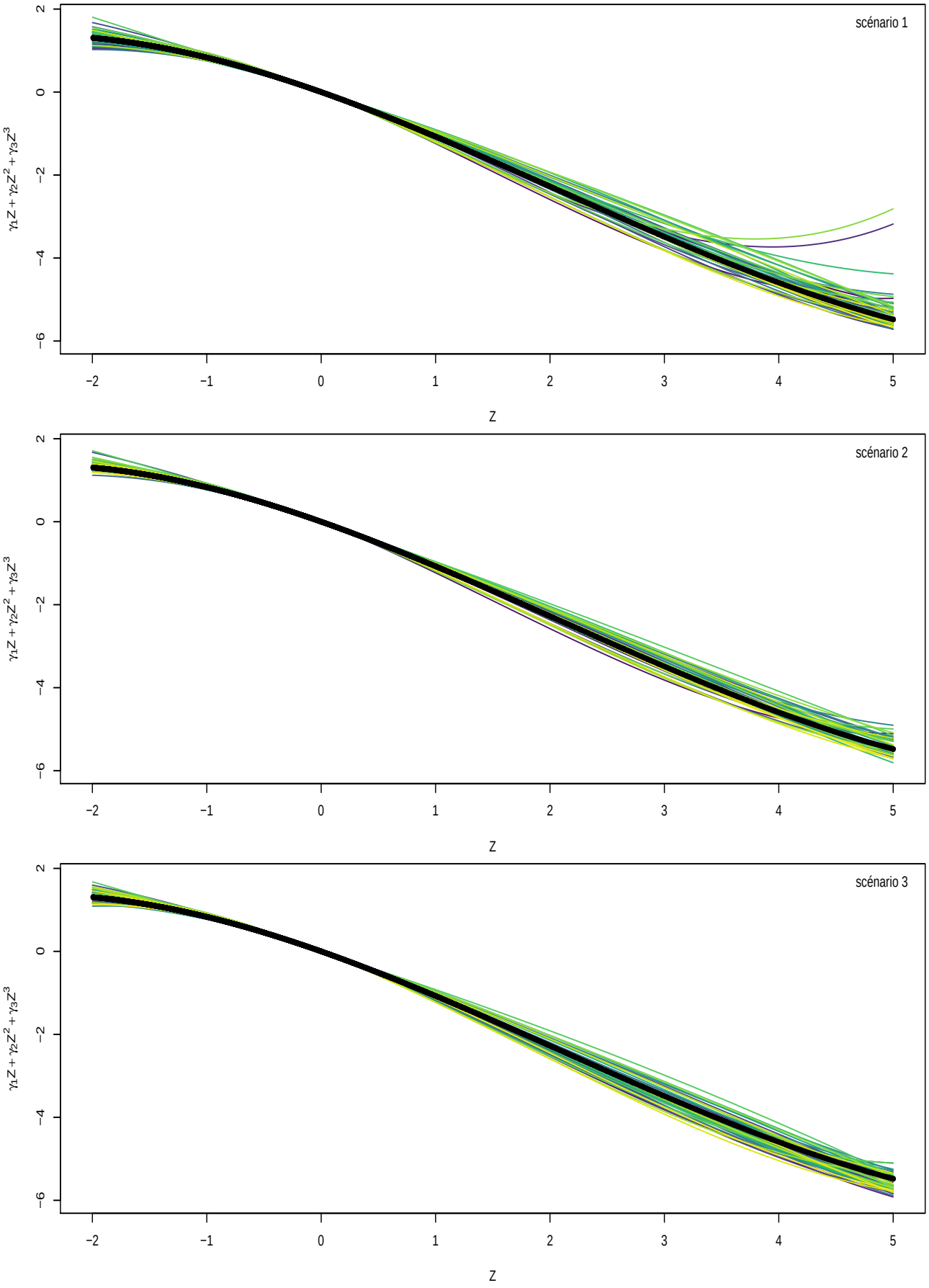

FIGURE III.4 – Résultats des simulations : comparaison de l'approche par pseudo-valeurs pour la moyenne restreinte des temps de survie sans démence 10 ans après une inclusion ($RM(10)$ suivant trois scénarios (de haut en bas : scénario 1 sans censure par intervalle, scénario 2 avec censure par intervalle avec une visite tous les 2 à 3 ans et scénario 3 avec censure par intervalle avec une visite tous les 3 à 6 ans). Schéma B : 50 répliques de 3200 sujets.

des temps de survie sans démence 10 ans après l'inclusion. L'effet du score au MMS a été ajusté sur le niveau d'éducation (la variable *educ* vaut 1 si le sujet a obtenu au moins le certificat d'étude primaire (CEP), 0 sinon), le sexe (qui vaut 0 pour les hommes et 1 pour les femmes) et l'âge (en année) à l'entrée dans la cohorte.

III.3.1 Modélisation

Un modèle linéaire a été utilisé pour chaque indicateur épidémiologique de la démence, avec la variable MMS^* , ajusté sur l'age^*, le sexe et le niveau d'éducation. Pour des raisons d'interprétation, les variables représentant le score au MMS (MMS) et l'âge à l'inclusion age ont été translatées, avec $MMS^* = MMS - 26$ et $age^* = age - 75$. L'individu de référence (c.à.d. avec toutes ces variables explicatives égales à 0) est un homme de 75 ans avec un MMS de 26 points à l'inclusion et qui n'a pas le certificat d'étude primaire. L'effet du MMS et de l'âge ont été modélisés avec une fonction cubique pour les trois indicateurs épidémiologiques.

Le modèle de régression pour le risque absolu de démence 10 ans après l'inclusion dans la cohorte PAQUID est défini par :

$$\mathbb{E}\left(\hat{Y}_{F_{01}i}(10)\right) = \gamma_{0,F} + \gamma_{1,F}MMS_i^* + \gamma_{2,F}MMS_i^{*2} + \gamma_{3,F}MMS_i^{*3} + \gamma_{4,F}age_i^* + \gamma_{5,F}age_i^{*2} + \gamma_{6,F}age_i^{*3} + \gamma_{7,F}sexe_i + \gamma_{8,F}educ_i \quad \text{(III.10)}$$

La probabilité d'être vivant non-dément 10 ans après l'inclusion dans la cohorte PAQUID, a été modélisée par :

$$\mathbb{E}\left(\hat{Y}_{P_{00}i}(10)\right) = \gamma_{0,P} + \gamma_{1,P}MMS_i^* + \gamma_{2,P}MMS_i^{*2} + \gamma_{3,P}MMS_i^{*3} + \gamma_{4,P}age_i^* + \gamma_{5,P}age_i^{*2} + \gamma_{6,P}age_i^{*3} + \gamma_{7,P}sexe_i + \gamma_{8,P}educ_i \quad \text{(III.11)}$$

Des pseudo-valeurs pour la moyenne restreinte des temps de survie sans démence ont été utilisées comme variable réponse dans le modèle de régression suivant :

$$\mathbb{E}\left(\hat{Y}_{RMi}(10)\right) = \gamma_{0,R} + \gamma_{1,R}MMS_i^* + \gamma_{2,R}MMS_i^{*2} + \gamma_{3,R}MMS_i^{*3} + \gamma_{4,R}age_i^* + \gamma_{5,R}age_i^{*2} + \gamma_{6,R}age_i^{*3} + \gamma_{7,R}sexe_i + \gamma_{8,R}educ_i \quad \text{(III.12)}$$

III.3.2 Descriptif de l'échantillon

L'échantillon est composé de 2641 sujets non-déments à l'inclusion ayant passé le test du MMS dans des conditions normales à T_0. Les conditions de passation du test sont dites « normales » lorsque le sujet a accepté de passer le test et qu'il n'y a pas eu d'arrêt au cours de la passation.

Le score moyen au MMS était de 25,8 points (avec un écart-type de 3,2, un minimum de 8 et un maximum de 30 points). Pour cet échantillon de la cohorte PAQUID, les individus ont été inclus en moyenne à 75,0 ans (avec un écart-type de 6,8 ans et au maximum à 101 ans). Environ 60% des individus sont des femmes et deux-tiers des sujets avaient au moins obtenus le certificat d'étude primaire. Un histogramme de la distribution du MMS à l'inclusion et un histogramme de l'âge à l'entrée dans la cohorte PAQUID pour cet échantillon sont présentés en figure C.1.

III.3.3 Résultats

Tableau III.4 – Modélisation du risque absolu de démence 10 ans après l'inclusion ($F_{01}(10)$) en fonction du score au MMS, de l'âge, du sexe et du niveau d'éducation. Estimation par pseudo-valeurs d'après 2641 sujets de la cohorte PAQUID.

	$\hat{\gamma}$	$\text{IC}_{95\%}(\hat{\gamma})$	p-valeur
Individu de référence	0,18964	[0,12242 ; 0,25686]	
MMS (en points)	-0,01768	[-0,02899 ;-0,00637]	$< 0,001$
MMM^2	0,00115	[-0,00211 ; 0,00442]	
MMM^3	-0,00001	[-0,00026 ; 0,00024]	
age (en années)	0,01622	[0,01000 ; 0,02244]	$< 0,001$
age^2	-0,00008	[-0,00085 ; 0,00068]	
age^3	-0,00004	[-0,00010 ; 0,00001]	
sexe (Femmes versus Hommes)	0,05596	[0,00509 ; 0,10684]	0,031
Certificat d'Étude Primaire (avec contre sans)	-0,02828	[-0,08836 ; 0,03179]	0,356

Individu de référence : homme sans CEP âgé de 75 ans à l'inclusion et un score au MMS de 26 points

Tableau III.5 – Modélisation de la probabilité d'être vivant non-dément à de démence 10 ans après l'inclusion ($P_{00}(10)$) en fonction du score au MMS, de l'âge, du sexe et du niveau d'éducation. Estimation par pseudo-valeurs d'après 2641 sujets de la cohorte PAQUID.

	$\hat{\gamma}$	$\text{IC}_{95\%}(\hat{\gamma})$	p-valeur
Individu de référence	0,39718	[0,35199 ; 0,44237]	
MMS (en points)	0,02194	[0,01433 ; 0,02954]	$< 0,001$
MMS^2	0,00117	[-0,00103 ; 0,00336]	
MMS^3	0,00003	[-0,00014 ; 0,00019]	
age (en années)	-0,03618	[-0,04036 ; -0,03200]	$< 0,001$
age^2	-0,00041	[-0,00093 ; 0,00010]	
age^3	0,00006	[0,00002 ; 0,00009]	
sexe (Femmes versus Hommes)	0,11364	[0,07944 ; 0,14784]	$< 0,001$
Certificat d'Étude Primaire (avec contre sans)	0,04326	[0,00287 ; 0,08364]	0,036

Individu de référence : homme sans CEP âgé de 75 ans à l'inclusion et un score au MMS de 26 points

Tableau III.6 – Modélisation de la moyenne restreinte des temps de survie sans démence jusqu'à 10 ans après l'inclusion ($RM(10)$) en fonction du score au MMS, de l'âge, du sexe et du niveau d'éducation. Estimation par pseudo-valeurs d'après 2641 sujets de la cohorte PAQUID.

	$\hat{\gamma}$	$\text{IC}_{95\%}(\hat{\gamma})$	p-valeur
Individu de référence	7,31799	[7,02583 ; 7,61014]	
MMS (en points)	0,15103	[0,10187 ; 0,20020]	$< 0,001$
MMM^2	-0,00730	[-0,02148 ; 0,00687]	
MMM^3	-0,00005	[-0,00113 ; 0,00104]	
age (en années)	-0,19317	[-0,22021 ;-0,16614]	$< 0,001$
age^2	-0,00599	[-0,00932 ;-0,00265]	
age^3	0,00030	[0,00006 ; 0,00055]	
sexe (Femmes versus Hommes)	0,88020	[0,65910 ; 1,10131]	$< 0,001$
Certificat d'Étude Primaire (avec contre sans)	0,07067	[-0,19043 ; 0,33177]	0,596

Individu de référence : homme sans CEP âgé de 75 ans à l'inclusion et un score au MMS de 26 points

Les équations (III.10) à (III.12) estiment des effets cubiques des variables quantitatives. Les estimations de l'ensemble des paramètres de régression sont respectivement retranscrites dans les tableaux III.4 à III.6. Des courbes de l'effet du MMS et de l'âge à l'inclusion sont présentées en figures III.5 à III.7.

Pour le risque absolu de démence, les résultats des effets linéaires sont présentés dans le tableau III.4. Ce dernier montre qu'un homme de 75 ans avec un MMS de 26 points à l'inclusion a une probabilité de 19% de devenir dément dans les 10 ans suivant l'inclusion. La probabilité de développer une démence est significativement plus élevée de 6% chez les femmes que chez les hommes, à âge, score au MMS et niveau d'éducation identiques. Le niveau d'éducation n'apparaît pas ici comme élément influençant le risque absolu de démence à 10 ans. La figure III.5 montre qu'entre 8 et 25 points de MMS, plus le score au MMS est faible à l'inclusion, plus la probabilité de devenir dément dans les 10 ans après l'inclusion augmente, par rapport à un MMS de 26 (et toutes choses égales par ailleurs). Par exemple, un sujet qui avait un MMS de 15 à l'inclusion a un risque absolu de démence de 35% de plus qu'un sujet qui avait un MMS de 26 ($\text{IC}_{95\%}=[0,18; 0,51]$). Pour l'âge, à partir de 89 ans à l'inclusion, les résultats ne sont plus significatifs par rapport à 75 ans d'âge à l'inclusion (l'intervalle de confiance contient la valeur 0 et seulement 55 sujets étaient âgés d'au moins 90 ans à l'inclusion). La tendance montre qu'avant 90 ans, plus les sujets ont été inclus vieux, plus la probabilité de développer une démence dans les 10 ans suivant l'inclusion augmente (voir figure de droite de la figure III.5). Par exemple, un sujet inclus à 65 ans a une probabilité diminuée de 0,13 par rapport à un sujet de 75 ans, pour deux sujets de mêmes sexes ayant le même score au MMS et le même niveau d'éducation ($\text{IC}_{95\%}=[-0,23; -0,02]$). La tendance de l'effet de l'âge à partir de 90 ans s'explique car les individus les plus âgés au moment de l'inclusion sont décédés plus rapidement et n'avaient donc pas le « temps » de développer une démence avant leur décès.

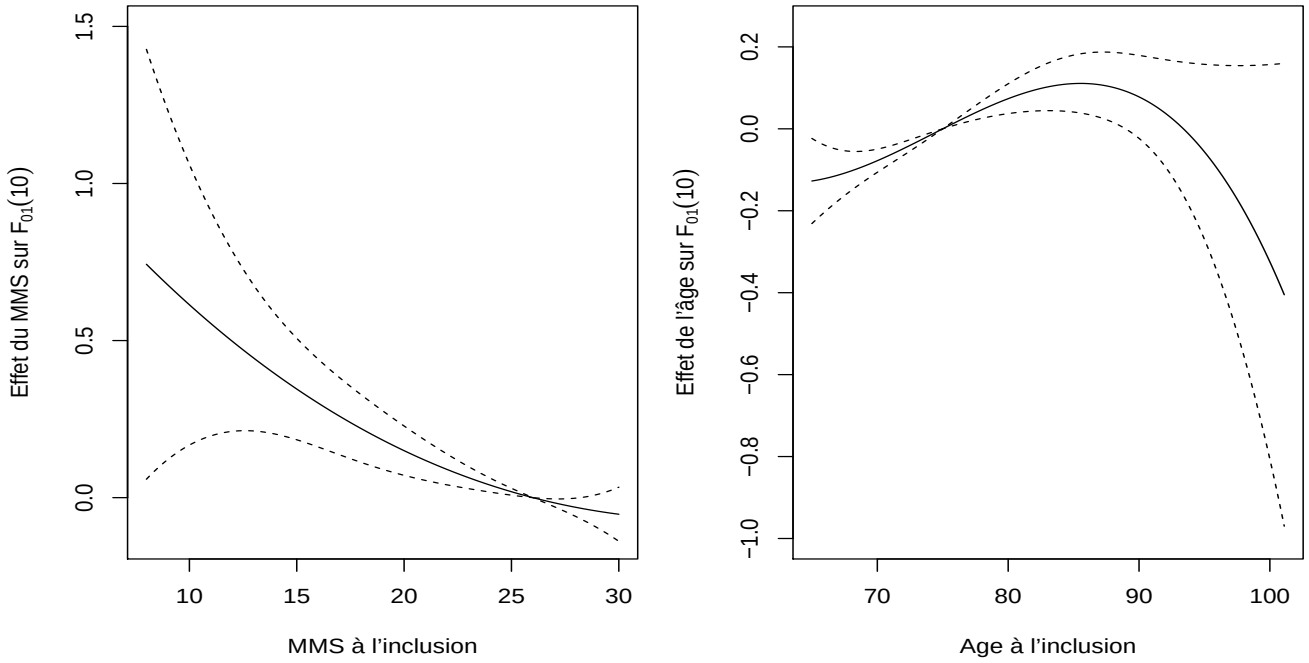

FIGURE III.5 – Effets du score au MMS (à gauche) et de l'âge (à droite), ajustés sur le sexe et le niveau d'éducation, sur la probabilité d'avoir développé une démence dans les 10 ans suivant l'inclusion ($F_{01}(10)$). Estimation par pseudo-valeurs d'après 2641 sujets de la cohorte PAQUID.

Les résultats de l'effet des variables explicatives sur la probabilité d'être vivant non-dément 10 ans après l'inclusion sont présentés dans le tableau III.5 et la figure III.6. La probabilité d'être vivant non-dément 10 ans après l'inclusion pour un homme sans certificat d'étude primaire âgé de 75 ans à l'inclusion avec un score au MMS de 26 points est de 40% environ. Cette probabilité augmente significativement de 0,11 pour les femmes par rapport aux hommes et augmente significativement de 0,04 pour les personnes ayant au moins le certificat d'étude primaire comme diplôme, par rapport à ceux qui ne l'ont pas. D'après la figure III.5, l'effet du MMS sur la probabilité de développer une démence dans les 10 ans suivant l'inclusion dans la cohorte PAQUID n'est pas significatif pour les personnes incluses avec un MMS inférieure à 14 points par rapport à une personne ayant eu un score de 26 points, ajusté sur l'âge à l'inclusion, le sexe et le niveau d'éducation. Ce résultat est dû au fait que seulement 11 sujets ont eu une mesure du MMS inférieure ou égale à 14 points. À partir de 14 points, plus le MMS est élevé à l'inclusion, plus la probabilité d'être vivant non-dément à $t = 10$ augmente, avec une augmentation plus importante pour les MMS les plus élevés. Par exemple, cette probabilité augmente de 0,11 ($IC_{95\%}=[0,05;0,17]$) pour un sujet qui avait un score au MMS de 30 à l'inclusion (le score le plus élevé) par rapport à un individu qui avait un score 26 (le score moyen sur l'échantillon). Pour l'âge, la courbe de droite de la figure III.6 montre une diminution de l'effet entre 65 et 90 ans puis à partir de 90 ans l'effet augmente de nouveau. À partir de 90 ans, la taille de l'intervalle de confiance augmente de plus en plus jusqu'à ne plus montrer d'effet de l'âge pour les sujets de plus de 99 ans à l'inclusion (une très grande minorité).

Les résultats des effets de la moyenne restreinte des temps de survie sans démence sont présentés

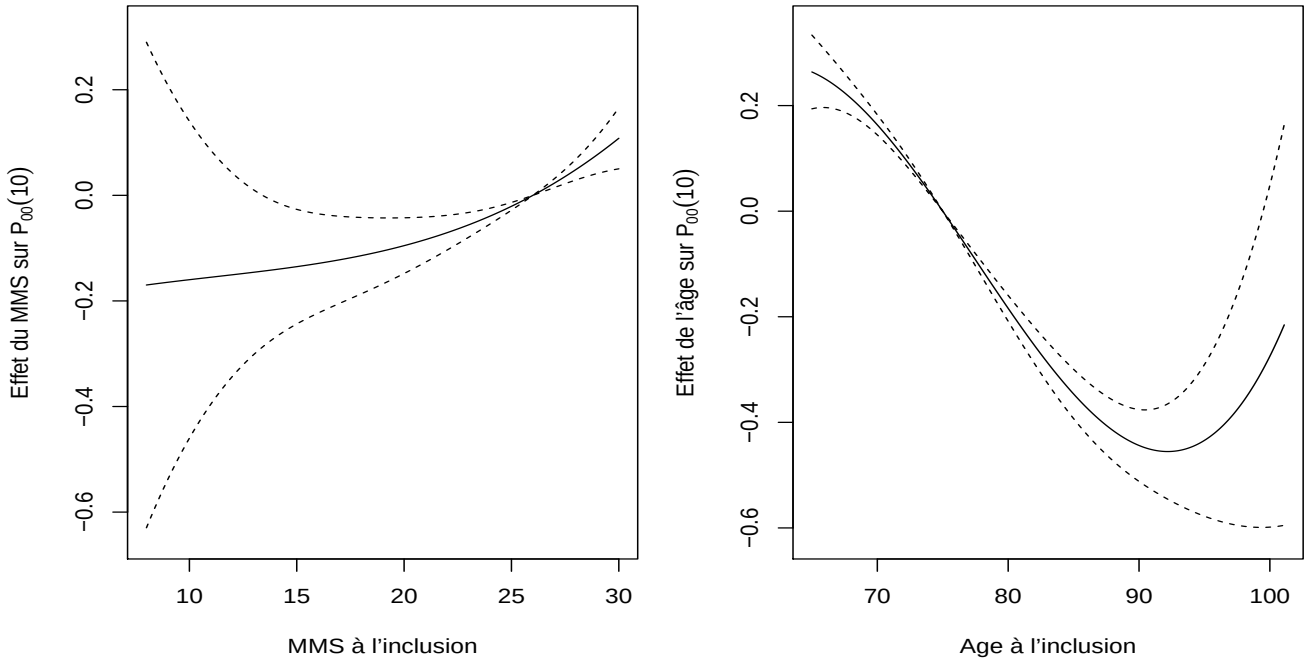

FIGURE III.6 – Effets du score au MMS (à gauche) et de l'âge (à droite), ajustés sur le sexe et le niveau d'éducation, sur la probabilité d'être encore vivant non-dément 10 ans après l'inclusion ($P_{00}(10)$). Estimation par pseudo-valeurs d'après 2641 sujets de la cohorte PAQUID.

dans le tableau III.6. Un homme sans CEP inclus à 75 ans et un score MMS de 26 points peut espérer rester 7 ans et trois mois vivant sans-démence dans les 10 ans suivant l'inclusion. Les femmes peuvent espérer rester un peu plus de 10 mois de plus que les hommes vivantes et non-démentes dans les 10 ans après l'inclusion, toutes choses égales par ailleurs. Les tendances de l'effet du MMS montre une augmentation de ce temps passé dans l'état vivant non-dément au fur et à mesure que le score au MMS était élevé à l'inclusion. Ainsi, un sujet qui avait un MMS de 10 à l'inclusion perd environ 4 ans de temps vivant sans démence par rapport à un sujet qui avait un score MMS de 26 ($\text{IC}_{95\%}=[-6,03;-2,16]$). Pour l'âge, l'effet est linéaire entre 75 et 95 ans : plus les sujets sont vieux à l'inclusion, plus le temps qu'ils peuvent espérer rester vivant sans-démence diminue. Pour les personnes incluses avant 75 ans et après 95 ans, l'effet se stabilise légèrement (voir la courbe de droite de la figure III.7). A titre informatif, un individu inclus à 85 ans passe 2 ans de moins dans l'état 0 qu'un individu de 75 ans à l'inclusion, toutes choses égales par ailleurs ($\text{IC}_{95\%}=[-2,53;-,1.93]$).

III.4 Conclusion et discussion

Ce chapitre a montré l'utilisation et le comportement de l'approche par pseudo-valeurs issues d'estimateurs d'un modèle *illness-death* pour temps de maladie censurés par intervalle. L'approche est proposée pour prendre en compte l'effet de variables explicatives sur trois indicateurs

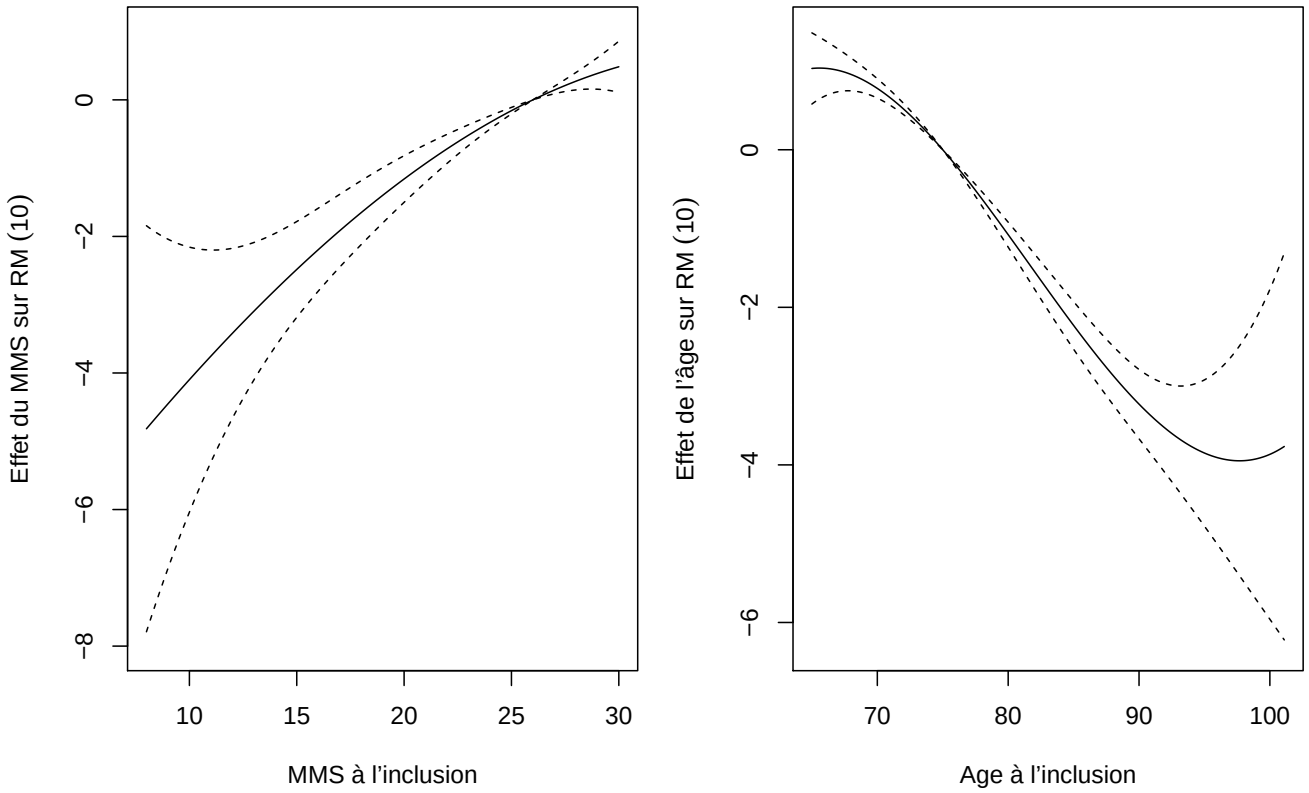

FIGURE III.7 – Effets du score au MMS (à gauche) et de l'âge (à droite), ajustés sur le sexe et le niveau d'éducation, sur la moyenne restreinte des temps de survie sans démence jusqu'à 10 ans suivant l'inclusion ($RM(10)$). Estimation par pseudo-valeurs d'après 2641 sujets de la cohorte PAQUID.

épidémiologiques de la démence : le risque absolu de démence, la probabilité d'être vivant non-dément et la moyenne restreinte des temps de survie sans démence, pour un horizon fixe t. Les résultats présentés dans ce chapitre sont basés sur des indicateurs épidémiologiques de la démence après 10 ans d'inclusion (c.à.d. $t = 10$). D'autres résultats sont présentés en annexe (pour d'autres temps et d'autres fonctions de liens).

Le plus gros avantage de cette approche est de pouvoir modéliser de manière souple l'effet de variables explicatives sur des indicateurs épidémiologiques. Cette souplesse intervient à deux niveaux : pour le calcul des estimateurs θ et θ^{-i} (avec $i = 1, ..., n$) et pour le choix du modèle de régression.

Deux méthodes ont été utilisées dans ce chapitre pour estimer les quantités θ ; la première basée sur une maximisation de la vraisemblance pénalisée (avec approche des intensités de transition par des splines) et la seconde est basée sur une maximisation paramétrique (en supposant une distribution de Weibull des intensités de transition). Les deux méthodes prennent en compte la censure par intervalle du temps de maladie et le fait que certaines transitions ne soient pas observées.

Pour un jeu de données de n sujets, il est nécessaire de calculer $(n + 1)$ estimations de θ, elles-mêmes estimées à partir de $(n + 1)$ estimations des modèles *illness-death*. Ce dernier point amène à un des inconvénients de l'approche par pseudo-valeurs. Les temps de calcul [2] pour l'approche par

2. Pour se donner une idée des temps de calcul, se référer à l'annexe B.2

pseudo-valeurs peuvent être un frein à l'utilisation de la méthode, en particulier pour l'approche par vraisemblance pénalisée où le nombre de paramètres à estimer est bien plus important que par l'approche paramétrique. De plus, plus le nombre de sujets augmente, plus le temps de calcul total augmente (car $(n+1)$ vecteurs de paramètres d'un modèle *illness-death* sont à estimer). Il parait donc évident qu'il ne faut pas utiliser les méthodes présentées dans ce chapitre si les données ne sont pas censurées par intervalle et leur préférer des approches non-paramétriques.
La remarque sur les temps de calcul est à relativiser : à partir du moment où les $n+1$ modèles *illness-death* ont été estimés, le calcul des pseudo-valeurs $\theta(t)$ est rapide. Il est donc possible de calculer $\theta(t)$ à différents temps (c.à.d. $t \in t_1, ...t_j$). Tout au long de ce chapitre, θ a fait référence à un estimateur d'une fonction des intensités de transition (cumulées). En particulier, θ a représenté ici le risque absolu de démence ($\theta(.) = F_{01}(.)$), la probabilité d'être vivant non-dément ($\theta(.) = P_{00}(.)$) ou la moyenne restreinte des temps de survie sans démence ($\theta(.) = RM(.)$). Il n'existe pas de raison de penser que cette approche ne peut pas être utilisée pour d'autres indicateurs épidémiologiques à partir du moment où ils peuvent être définis en fonction des intensités de transition d'un modèle *illness-death* (par exemple la probabilité d'être décédé sans démence, c'est-à-dire d'avoir fait la transition directement de l'état 0 à l'état 2).

La méthode par estimation paramétrique peut être une solution pour limiter les temps de calcul car le vecteur du modèle *illness-death* est composé de 6 paramètres par modèle *illness-death*. Ce point est un avantage pour le calcul des pseudo-valeurs qu'il faut nuancer car un des inconvénients reste que cette méthode paramétrique suppose une certaine distribution des intensités de transition.

Le choix de ne simuler que 50 jeux de données de 3200 sujets pour le schéma B est en partie expliqué par ce dernier point. En effet, pour réussir à capter un effet de la variable quantitative, n a du être augmenté entre les scénarios A et B. L'autre aspect à prendre en considération est la représentation graphique des simulations pour l'effet de la variable. Limiter le nombre total de réplications permet d'avoir des graphiques lisibles.

L'approche par pseudo-valeurs calcule des estimateurs θ et θ^{-i} marginalement. Aucune hypothèse n'est donc requise sur l'influence des variables explicatives sur ces estimateurs. Ce point est cependant à prendre en considération dans le cas où la censure n'est pas indépendante des variables explicatives. C'est l'hypothèse majeure que fait l'approche par pseudo-valeurs et que nous avons aussi supposée dans notre application. Si la censure ne respecte pas les hypothèses d'indépendance, alors il faut adapter des méthodes de pondérations (comme par l'inverse de la probabilité d'être observé) de l'estimateur pour corriger des biais (Binder *et al.*, 2014; Overgaard *et al.*, 2019).

Le second point évoqué comme un avantage de l'approche par pseudo-valeurs est la modélisation souple découlant des modèles de régression proposés dans ce chapitre. À partir des estimateurs θ d'une quantité d'intérêt, différents modèles peuvent être mis en œuvre pour mesurer l'effet de variables explicatives.

Un premier choix s'est porté ici sur un modèle linéaire généralisé estimé par GEE comme le fait la majorité des articles de la littérature. Différentes fonctions de lien ont été utilisées et

l'interprétation des résultats évoluent suivant le lien choisi. Un lien « identité » estime l'effet de variables explicatives en terme d'*ajout* (ou de *retrait*) pour une augmentation d'une unité de la variable explicative alors qu'un lien « logarithmique » estime des effets « multiplicatifs ».

Pour modéliser l'effet des variables quantitatives sur θ des modèles linéaires avec un effet cubique ont été estimés. L'avantage de ce type de modèle est de pouvoir estimer un effet non-linéaire des variables tout en gardant le contrôle de la modélisation. Une autre solution aurait été d'utiliser des modèles additifs généralisés (GAM). Les GAM ont l'avantage de modéliser d'une manière souple l'effet de variables explicatives quantitatives (en utilisant par exemple une base de splines). L'estimation de GAM a été peu retrouvée dans la littérature. Seuls quelques auteurs avaient utilisé des GAM dans le but de modéliser l'effet du temps dans leurs modèles de régression . Ceci s'explique car une grande partie des auteurs ont calculé pour un individu i plusieurs pseudo-valeurs $Y_i(t_j)$ pour différents temps $t_1 < ... < t_{\tau_j}$. Ils incluaient donc le temps comme variable explicative dans leur modèle (mixte). L'effet du temps était modélisé d'une manière qualitative ou d'une manière quantitative (avec par exemple des splines).
Dans ce chapitre, contrairement à la plupart des articles retrouvés dans la littérature, pour un modèle de régression, une seule pseudo-valeur par individu a été calculée car nous ne voulions pas faire d'hypothèses sur l'effet des variables explicatives au cours du temps.

L'approche par pseudo-valeurs permet aussi d'estimer directement un intercept c.à.d. un effet pour un individu avec toutes les variables explicatives égales à 0 (contrairement au modèle de Cox qui n'estime pas le risque de base par exemple). De plus, il est possible d'ajouter des interactions entre les variables explicatives, bien que cela n'ait pas été fait ici.

Par contre, il faut garder en tête que l'utilisation de pseudo-valeurs comme variable d'intérêt dans des GLM fait des hypothèses pas toujours vérifiées (par exemple une distribution gaussienne des erreurs de mesures). Dans ce chapitre, les pseudo-valeurs sont supposées suivre les hypothèses du modèle linéaire (comme la normalité de la variable réponse sachant les variables explicatives ou l'homoscédascité des variances) pour ne pas chercher à alourdir la modélisation si ces hypothèses n'étaient pas acceptables.

Les modèles proposés dans ce chapitre ont été appliqués aux données de la cohorte PAQUID. Pour résumer les résultats, l'effet du score au MMS est en cohérence avec les résultats de la littérature sur le risque instantané de démence. Plus le score est élevé à l'inclusion, plus le sujet a de chance d'être vivant et non-dément à 10 ans de suivi et donc inversement plus ils ont de chance de ne pas avoir développé une démence dans ces 10 années, ajusté sur l'âge, le sexe et le niveau d'éducation. Les effets de l'âge montrent qu'un sujet inclus entre 65 et 85 ans a tendance à avoir une probabilité développer une démence qui augmente, alors que les sujets de 85 ans et plus au moment de l'inclusion ont une probabilité de développer une démence qui diminue au fur et à mesure que l'âge augmente. Ceci est dû au risque compétitif de décès. Les sujets inclus les plus âgés sont décédés plus rapidement que les sujets inclus les plus jeunes : ils ont donc moins de chance de développer une démence au cours du suivi. Les femmes vivent plus longtemps que les hommes : ce résultat se traduit par une augmentation du risque absolu de démence également. Les résultats

concernant le niveau d'éducation, modélisé par le fait d'avoir au moins le certificat d'étude primaire ne sont pas déterminants. L'ajustement par l'âge et le score au MMS peut être une explication de ce résultat. En effet, à âge, sexe et score au MMS à l'inclusion équivalent, deux personnes qui n'ont pas le même niveau d'éducation auront en moyenne les mêmes chances de subir (ou non) un des deux évènements du modèle *illness-death*. Le score au MMS pourrait capter plus d'effets à lui tout seul que le niveau d'éducation catégorisé en deux modalités. Des valeurs faibles du score au MMS (inférieures à 15 points) ou des âges élevés au moment de l'inclusion (supérieures à 90 ans) ne montrent pas toujours un effet différent par rapport à l'individu de référence (c'est-à-dire comparé à un sujet avec un score de 26 points ou un sujet de 75 ans à l'inclusion). Pour capter un effet, il faut donc suffisamment d'informations.

Chapitre IV

Approche par linéarisation

La méthode proposée dans ce chapitre a le même objectif que l'approche par pseudo-valeurs : il s'agit de pouvoir quantifier l'effet de variables explicatives sur des indicateurs épidémiologiques de la démence, indicateurs qui dépendent à la fois du risque de démence mais aussi du risque de décès. Cette méthode s'appuie sur une idée de Daniel Commenges présentée durant un *workshop* sur les modèles multi-états à Berlin en Novembre 2016 et intitulée *Summarizing the effect of covariates on complex interesting quantities by linearisation of their maximum likelihood estimator.*

IV.1 Méthode

La méthode de l'approche par linéarisation est d'abord présentée dans un cadre général puis ce cadre est défini plus précisément à travers différentes méthodes d'estimations et exemples.

IV.1.1 Méthodologie Générale

L'idée générale de la méthode est de proposer un modèle de régression qui lie les variables explicatives Z à un indicateur épidémiologique $\theta(s,t)$ (ou $\theta(s)$ pour les indicateurs épidémiologiques calculés pour un horizon infini) défini en section II.2. Pour cela, l'effet de Z est résumé par le modèle linaire suivant :

$$\theta(s,t,Z) = \Gamma_{0,s,t} + \Gamma_{1,s,t}Z_1 + ... + \Gamma_{p,s,t}Z_p + \varepsilon$$

Les indicateurs épidémiologiques de la section II.2 sont des fonctions des intensités de transition et intensités de transition cumulées d'un modèle *illness-death.*

Soit $\alpha_{kl}(t, Z_{kl})$ l'intensité de transition entre l'état k et l'état l au temps t conditionnellement aux variables explicatives Z_{kl}. Soit $A_{kl}(s,t,Z_{kl}) = \int_s^t \alpha_{kl}(u, Z_{kl})\mathrm{d}u$ l'intensité de transition cumulée de l'état k vers l'état l entre les temps s et t conditionnellement aux variables explicatives. De plus, soit $\Phi(s,t,Z)$ un ensemble des fonctions du modèle *illness-death* comprenant les trois intensités de

transition et les trois intensités de transition cumulées.

1. La première partie de la méthode estime les trois intensités de transition d'un modèle *illness-death* $\hat{\alpha}_{kl}(t, Z_{kl})$, $kl \in \{01, 02, 12\}$ en fonction des variables explicatives Z_{kl} à partir d'un jeu de données n sujets. Notons par $\hat{\varphi}$ le vecteur des paramètres estimés dans cette étape et $\hat{V_\varphi}$ l'estimation de la matrice de variance-covariance associée au vecteur φ.

2. À partir de l'estimation de $\hat{\Phi}(s, t, Z)$, il est possible de calculer $\hat{\theta}(s, t, Z)$ pour n'importe quelle valeur de Z [1]. Un tirage au sort aléatoire de m sujets parmi n permet de réaliser la seconde étape : la méthode estime différentes valeurs de $\hat{\theta_i}(s, t, Z = z_i)$, avec $i = 1, ..., m$.

3. La troisième étape est l'estimation d'un modèle linéaire utilisant les m valeurs estimées $\hat{\theta_i}(s, t, Z = z_i)$ comme variable à expliquer :

$$\hat{\theta_i}(s, t, Z = z_i) = \hat{\Gamma}_{0,s,t} + \hat{\Gamma}_{1,s,t} Z_{1i} + ... + \hat{\Gamma}_{p,s,t} Z_{pi} + \varepsilon_i \tag{IV.1}$$

avec $i = 1, ..., m$.

Les paramètres $\hat{\Gamma}$ sont les estimations ponctuelles de l'effet des variables explicatives Z sur l'indicateur épidémiologique considéré entre les temps s et t.

Pour un temps s, un temps t et pour des valeurs z_i des variables explicatives du sujet i fixés, la variance de $\theta_i(s, t, Z = z_i)$ est nulle. Il n'est donc pas possible d'utiliser les variances estimées par l'équation (IV.1) pour calculer un intervalle de confiance du vecteur des estimations $\hat{\Gamma}$.

4. L'étape suivante est suggérée pour calculer une bande de confiance pour $\hat{\Gamma}$. Elle s'appuie sur la formalisation de Mandel (2013) pour le calcul d'une bande de confiance basée sur une technique de simulation. Soit φ l'ensemble des paramètres de toutes les fonctions de $\Phi(s, t, Z)$. En supposant que le vecteur des paramètres estimés $\hat{\varphi}$ suit asymptotiquement une loi normale multivariée, cette technique consiste à générer J vecteurs de paramètres $\varphi^{[1]}, \varphi^{[2]}, ..., \varphi^{[J]}$ à partir d'une loi normale multivariée de moyenne $\hat{\varphi}$ et de variance-covariance $\hat{V_\varphi}$.
Il s'agit maintenant d'estimer de nouveaux les paramètres $\theta_{i,j}(s, t, Z = z_i)$, $\forall i = 1, ..., m$ et $\forall j = 1, ..., J$ depuis les intensités de transitions $\Phi^{[j]}(s, t, Z)$ en utilisant les paramètres $\varphi^{[j]}$ à la place des paramètres φ. Les indicateurs épidémiologiques sont alors utilisés dans J modèles linéaires de régression :

$$\hat{\theta}_i^{[j]}(s, t, Z_i) = \hat{\Gamma}^{[j]}_{0,s,t} + \hat{\Gamma}^{[j]}_{1,s,t} Z_{1i} + ... + \hat{\Gamma}^{[j]}_{p,s,t,} Z_{pi} + \varepsilon_i^{[j]} \tag{IV.2}$$

avec $i = 1, ..., m$ et $j = 1, ..., J$. Cette étape combine les étapes 2 et 3 vu précédemment.

5. Une bande de confiance peut maintenant être estimée à partir du modèle IV.1 et des J modèles IV.2 estimés. Pour un paramètre $\Gamma_{q,s,t}$, une bande de confiance peut être donnée à partir des 2,5$^{\text{e}}$ et 97,5$^{\text{e}}$ percentiles de la distribution de $\left(\hat{\Gamma}_{q,s,t}, \hat{\Gamma}^{[j]}_{q,s,t}\right)$, avec $j = 1, ..., J$ et $q = 0, ..., p$.

1. Les valeurs de Z doivent tout de même être comprise sur l'étendue des valeurs observées sur le jeu de données initial.

IV.1.2 Notation du modèle de régression

La méthodologie générale proposée estime des paramètres de régression grâce à plusieurs modèles linéaires (un pour l'estimation ponctuelle et J modèles supplémentaires pour la bande de confiance de chaque paramètre). Pour simplifier, on notera la méthode d'estimation par :

$$\hat{\theta}(s,t,Z) \leftrightsquigarrow \hat{\Gamma}_0 + \hat{\Gamma}_1 Z_1 + ... + \hat{\Gamma}_p Z_p \tag{IV.3}$$

avec $\leftrightsquigarrow$ représentant les étapes 3 et 4, pour obtenir l'estimation ponctuelle et la bande de confiance des paramètres de régression $\hat{\Gamma}$ (par l'étape 5).

IV.1.3 Méthodes d'estimation

La méthodologie principale est proposée à partir de $\Phi(s,t,Z)$. Dans le cadre de données censurées par intervalle, nous proposons d'utiliser des estimations du maximum de vraisemblance pénalisée approchée par des splines (dite méthode semi-paramétrique, φ est composé des coefficients des splines) ou d'utiliser une méthode paramétrique par estimation de la vraisemblance par des distributions de Weibull (φ est alors composé d'un paramètre de forme et d'un paramètre d'échelle par transition).

Ces deux méthodes d'estimation prennent en compte la censure par intervalle du temps de maladie et aussi peuvent aussi prendre en compte la troncature à gauche (pour tenir compte de l'entrée retardée dans une cohorte).

IV.1.4 Modèle à intensités de transition proportionnelles

La modélisation de $\Phi(s,t,)$ peut se faire en utilisant un effet proportionnel des variables explicatives sur les intensités de transition. Les trois modèles cause-spécifiques sont alors définis par :

$$\begin{aligned}
\alpha_{01}(t \mid Z_{01}) &= \alpha_{01,0}(t)\exp(\beta_{01}^\top Z_{01}) \\
\alpha_{02}(t \mid Z_{02}) &= \alpha_{02,0}(t)\exp(\beta_{02}^\top Z_{02}) \\
\alpha_{12}(t \mid Z_{12}) &= \alpha_{12,0}(t)\exp(\beta_{12}^\top Z_{12})
\end{aligned}$$

Si ce choix de modélisation de l'effet de Z est choisi alors le vecteur $\hat{\varphi}$ correspond aux estimateurs des paramètres des intensités de transition de bases $(\alpha_{kl,0})$ et aux paramètres de régression (β_{kl}) pour les trois transitions. Donc $\hat{\varphi}$ est composé des coefficients de splines et des paramètres de régression si une méthode semi-paramétrique est utilisée ou alors $\hat{\varphi}$ est composé des 6 paramètres de lois de Weibull et des paramètres de régression si une méthode paramétrique est utilisée.

IV.1.5 Modèle stratifié

Une autre solution pour modéliser l'effet de variables explicatives sans supposer d'effet proportionnel est d'utiliser un modèle stratifié (lorsque les variables sont qualitatives). Dans ce cas, le modèle *illness-death* peut être défini par :

$$\begin{aligned}
\alpha_{01}(t \mid Z^\bullet, Z^\circ_{01}) &= \alpha_{01,0,Z^\bullet}(t)\exp(\beta^{\bullet\top}_{01} Z^\circ_{01}) \\
\alpha_{02}(t \mid Z^\bullet, Z^\circ_{02}) &= \alpha_{02,0,Z^\bullet}(t)\exp(\beta^{\bullet\top}_{02} Z^\circ_{02}) \\
\alpha_{12}(t \mid Z^\bullet, Z^\circ_{12}) &= \alpha_{12,0,Z^\bullet}(t)\exp(\beta^{\bullet\top}_{12} Z^\circ_{12})
\end{aligned}$$

avec $Z^\bullet$ les variables explicatives utilisées pour la stratification, Z°_{kl} les variables explicatives d'ajustement dont l'effet est supposé proportionnel sur l'intensité de transition de l'état k vers l'état l. Le vecteur φ est alors composé des différents sous-modèles estimés suivant les différentes modalités de $Z^\bullet$.

Le tirage au sort de m sujet parmi n dans la méthodologie générale doit respecter la proportion de sujets pour chaque modalité des variables utilisées pour stratifier.

Par exemple, si une variable binaire Z_1 sert de variable de stratification, alors $\alpha_{kl}(t \mid Z^\bullet_1, Z^\circ_{kl})$ est composé des deux estimations de bases pour $Z^\bullet_1 = 0$ et pour $Z^\bullet_1 = 1$ et des paramètres de régression $\beta^\bullet_{kl}$ pour $Z^\bullet_1 = 0$ et $Z^\bullet_1 = 1$, c.à.d. que $\Phi(s,t,Z)$ peut être vu comme deux sous-modèles $\Phi^{\bullet 0}(s,t,Z^\circ)$ et $\Phi^{\bullet 1}(s,t,Z^\circ)$. Le tirage au sort des m sujets est divisé en deux tirages aléatoires : le premier se fait avec $m_0 = m \times \mathbb{P}(Z^\bullet_1 = 0)$ parmi n_0, le nombre total de sujet ayant $Z^\bullet_1 = 0$ et $m_1 = m \times \mathbb{P}(Z^\bullet_1 = 1)$ parmi n_1, le nombre total de sujet ayant pour caractéristique $Z^\bullet_1 = 1$ (avec $m = m_0 + m_1$ et $n = n_0 + n_1$).

La quantité d'intérêt θ_i est alors calculé par

$$\theta_i(s,t,Z^\bullet_{1i}, Z^\circ_i) = \begin{cases} h\left(\Phi^{\bullet 0}(s,t,Z^\circ)\right) \text{ si } Z^\bullet_{1i} = 0 \\ h\left(\Phi^{\bullet 1}(s,t,Z^\circ)\right) \text{ si } Z^\bullet_{1i} = 1 \end{cases}$$

pour $i = 1, ..., m$ et $h(.)$ une fonction des intensités de transition.

IV.1.6 Cas particulier - modèle sans variable quantitative

Lorsque l'ajustement aux variables explicatives ne se fait que par des variables qualitatives, il n'est pas nécessaire de tirer au sort m sujets parmi n mais seulement d'estimer $\theta_i(s,t,Z_i)$ pour que l'ensemble des combinaisons possibles des variables explicatives soit représenté. Par exemple, avec deux variables binaires Z_1 et Z_2, il suffit de prendre m égal à 4 pour représenter les différentes combinaisons ($Z_1, Z_2 = \{00, 01, 10, 11\}$). Le modèle linéaire est alors un modèle linéaire pondéré par la distribution des variables qualitatives Z_1 et Z_2 (pour les étapes 3 et 4).

De plus, si l'objectif est de regarder l'effet d'une seule variable binaire Z sur l'indicateur

épidémiologique θ alors il n'y a pas besoin d'estimer l'effet Γ par un modèle linéaire (car $\Gamma_0 = \theta(s,t,Z=0)$ et $\Gamma_1 = \theta(s,t,Z=1) - \theta(s,t,Z=0)$).

IV.2 Étude de simulations

IV.2.1 Schéma A

Une première partie des simulations reprend les mêmes données que celles présentées dans le schéma A de l'approche par pseudo-valeur (voir section III.2.1 pour le détail de la génération des données).

IV.2.1.1 Procédure d'estimation

L'approche par linéarisation a été utilisée avec des modèles à intensités de transition proportionnelles suivant la variable binaire Z avec :

$$\alpha_{01}(t \mid Z) = \alpha_{01,0}(t)\exp[\beta_{01} \times Z]$$
$$\alpha_{02}(t \mid Z) = \alpha_{02,0}(t)\exp[\beta_{02} \times Z]$$
$$\alpha_{12}(t \mid Z) = \alpha_{12,0}(t)\exp[\beta_{12} \times Z]$$

L'estimation du modèle *illness-death* a été faite en utilisant le délai comme temps de base par deux méthodes d'estimation :

— un premier qui approche le maximum de vraisemblance pénalisée par des splines. Le vecteur φ est composé alors des coefficients de splines et des paramètres de régression β. Une base de M-splines cubiques avec 5 nœuds intérieurs placés de manière équidistante pour chaque transition a été utilisée. Les paramètres de lissage κ de la vraisemblance pénalisée ont été choisis par visualisation des intensités de transition de base du premier échantillon. Les paramètres ont été les mêmes pour les autres échantillons [2].

— un second type d'estimateur basé sur une approche paramétrique. Le vecteur φ est constitué des 6 paramètres pour les intensités de base (un paramètre de forme et un paramètre d'échelle d'une loi de Weibull par transition) et des paramètres de régression β.

Un retirage de 1000 vecteurs de paramètres φ d'une loi normale multivariée $\mathcal{NM}(\hat{\varphi}, \hat{V}_{\varphi})$ a été fait pour pouvoir calculer les bandes de confiances des valeurs estimées.

2. Une recherche par validation croisée des paramètres de lissage a été faite si il n'y a pas eu convergence du modèle *illness-death* avec les paramètres de lissage du premier échantillon

Les trois modèles estimés par l'approche par linéarisation ont été :

$$F_{01}(0, 10 \mid Z) \leftrightsquigarrow \Gamma_{0,F} + \Gamma_{1,F} Z_1 \tag{IV.4}$$

$$P_{00}(0, 10 \mid Z) \leftrightsquigarrow \Gamma_{0,P} + \Gamma_{1,P} Z_1 \tag{IV.5}$$

$$RM(10 \mid Z) \leftrightsquigarrow \Gamma_{0,R} + \Gamma_{1,R} Z_1 \tag{IV.6}$$

c'est-à-dire que l'effet de Z est estimé pour le risque absolu de démence 10 ans après l'inclusion, pour la probabilité d'être encore vivant non-dément 10 ans après l'inclusion et pour la moyenne restreinte des temps de survie sans démence dans un horizon de 10 ans. Comme la régression s'est basé sur une seule variable binaire, $m = 2$ et il n'y a pas eu besoin de faire un modèle de régression.

IV.2.1.2 Résultats

Les résultats sont présentés dans le tableau IV.1 pour les trois indicateurs épidémiologiques calculés à 10 ans de suivi.

Les résultats de l'approche par linéarisation sont sensiblement équivalents à ceux donnés par l'approche par pseudo-valeurs (voir tableaux III.1 à III.3) et les interprétations des paramètres de régression sont identiques à celles énoncées dans tableaux III.1 à III.3. Les biais sont quelques fois plus importants par linéarisation, surtout pour la moyenne restreinte des temps de survie sans démence estimés par des méthodes semi-paramétriques (pour les trois scénarios, voir en particulier $\Gamma_{0,R}$). Si l'on compare l'approche par linéarisation à l'approche par pseudo-valeurs, les RMSE sont toujours plus faible pour l'approche par linéarisation. Le tableau IV.1 donne des taux de couverture compris dans l'intervalle $(93,1; 96,9)$ pour presque tous les paramètres, toutes les méthodes d'estimation et tous les scénarios. Les exceptions sont pour $\Gamma_{1,F}$ pour le scénario 1 et le scénario 3 (seulement pour la méthode paramétrique), pour $\Gamma_{1,P}$ avec le scénario 2 et la méthode paramétrique et le scénario 3 estimé par la méthode semi-paramétrique. Le taux de couverture pour la méthode semi-paramétrique et le scénario 1, sans censure par intervalle, donne un mauvais taux de couverture pour l'intercept ($\Gamma_{0,R}$ avec un taux de couverture de 85,8), dû à un léger biais.

La censure par intervalle (comparaison entre les scénarios 1 (sans censure par intervalle) et les scénarios 2 et 3) ne semble pas influencer la qualité des résultats. Les biais ont tendance à légèrement augmenter mais ils restent faibles. Également, les RMSE augmentent au fur et à mesure que la perte d'information due à la censure par intervalle augmente.

Tableau IV.1 – Résultats des simulations : comparaison de l'approche par linéarisation ($m = 2$ et $J = 1000$) pour le risque absolu de démence ($F_{01}(10)$), la probabilité d'être vivant non-dément ($P_{00}(10)$) et la moyenne restreinte des temps de survie sans démence ($RM(10)$) 10 ans après l'inclusion suivant deux méthodes d'estimation et trois scénarios. Schéma A : 500 échantillons de 500 sujets.

Indicateur	Γ	Scénario et méthode	$\overline{\hat{\Gamma}}$	Biais	ASE	ESE	RMSE × 1000	Taux de couv.
F_{01}	$\Gamma_{0,F} = 0,193$	1- Splines	0,192	-0,001	0,023	0,023	23	95,2
		1- Weibull	0,193	-0,000	0,021	0,021	21	96,0
		2- Splines	0,196	0,003	0,034	0,028	28	94,8
		2- Weibull	0,195	0,002	0,028	0,027	27	96,6
		3- Splines	0,198	0,005	0,038	0,036	36	93,6
		3- Weibull	0,197	0,004	0,034	0,035	35	93,8
	$\Gamma_{1,F} = 0,096$	1- Splines	0,098	0,002	0,028	0,030	30	91,8
		1- Weibull	0,096	-0,000	0,027	0,030	30	91,2
		2- Splines	0,095	-0,001	0,037	0,036	36	95,0
		2- Weibull	0,092	-0,004	0,034	0,035	35	95,0
		3- Splines	0,093	-0,003	0,044	0,046	46	95,0
		3- Weibull	0,090	-0,006	0,041	0,044	44	90,6
P_{00}	$\Gamma_{0,P} = 0,434$	1- Splines	0,435	0,001	0,029	0,027	27	96,2
		1- Weibull	0,432	-0,002	0,027	0,026	26	96,0
		2- Splines	0,435	0,001	0,031	0,027	28	94,8
		2- Weibull	0,432	-0,002	0,027	0,027	27	96,2
		3- Splines	0,436	0,002	0,030	0,028	28	95,8
		3- Weibull	0,433	-0,001	0,027	0,027	27	96,2
	$\Gamma_{1,P} = 0,055$	1- Splines	0,052	-0,003	0,034	0,036	36	93,6
		1- Weibull	0,055	-0,000	0,034	0,036	36	94,4
		2- Splines	0,051	-0,004	0,035	0,037	37	93,6
		2- Weibull	0,055	0,000	0,034	0,036	36	93,0
		3- Splines	0,050	-0,005	0,036	0,038	38	93,6
		3- Weibull	0,054	-0,001	0,035	0,037	37	94,4
RM	$\Gamma_{0,R} = 7,255$	1- Splines	7,157	-0,098	0,195	0,190	214	85,8
		1- Weibull	7,247	-0,008	0,182	0,176	177	96,4
		2- Splines	7,267	0,012	0,235	0,180	180	94,8
		2- Weibull	7,249	-0,006	0,184	0,178	178	96,0
		3- Splines	7,270	0,015	0,218	0,184	184	96,6
		3- Weibull	7,254	-0,001	0,186	0,180	180	95,8
	$\Gamma_{1,R} = 0,417$	1- Splines	0,399	-0,018	0,204	0,206	207	94,0
		1- Weibull	0,417	0,000	0,197	0,206	206	94,4
		2- Splines	0,392	-0,025	0,222	0,217	218	93,0
		2- Weibull	0,416	-0,001	0,203	0,216	216	93,0
		3- Splines	0,375	-0,042	0,226	0,225	229	94,6
		3- Weibull	0,403	-0,014	0,210	0,220	221	93,2

scénario 1 sans censure par intervalle

scénario 2 avec censure par intervalle et 2,5 ans entre deux visites en moyennes

scénario 3 avec censure par intervalle et 4,5 ans entre deux visites en moyennes

$\overline{\hat{\Gamma}}$ = moyenne des estimations

ASE = écart-type asymptotique / ESE = écart-type empirique

RMSE = racine carrée de l'erreur quadratique moyenne

Taux de couv. = pourcentage des 500 échantillons ayant $\Gamma \in \mathrm{BC}_{95\%}(\hat{\Gamma})$

IV.2.2 Schéma C

IV.2.2.1 Génération des données

Ce schéma a simulé des données se rapprochant le plus des observations d'une cohorte avec entrée retardée, c'est-à-dire que les sujets ne sont pas tous inclus aux mêmes âges. Cinq cents jeux de données de 3500 sujets chacun ont été générés aléatoirement.

Pour chaque jeu de données, deux variables explicatives Z_1 et Z_2 ont été créées : Z_1 était une variable binaire qui suit une loi de Bernoulli avec $\mathbb{P}(Z_1 = 1) = 0,58$ qui mime la distribution du sexe dans la cohorte PAQUID et Z_2 était une variable quantitative définie tel que $Z_2 \sim \mathcal{N}(24; 3, 3^2)$ pour représenter l'Indice de Masse Corporelle (IMC) à l'inclusion dans la cohorte PAQUID. Les intensités de transition du modèle ont été simulées avec un effet proportionnel de Z_1 et $Z_2^* = Z_2 - 24$ avec :

$$\begin{aligned}
\alpha_{01}(t \mid Z) &= \alpha_{01,0}(t) \exp(0,192 \times Z_1) \exp(0,080 \times Z_2^*) \\
\alpha_{02}(t \mid Z) &= \alpha_{02,0}(t) \exp(-0,660 \times Z_1) \exp(0,006 \times Z_2^*) \\
\alpha_{12}(t \mid Z) &= \alpha_{12,0}(t) \exp(-0,405 \times Z_1) \exp(0,020 \times Z_2^*)
\end{aligned}$$

Les paramètres des intensités de transition de bases ont été choisis à partir des données de la cohorte PAQUID (voir tableau B.1 pour les valeurs choisies).

De plus, une entrée retardée a été simulée suivant une loi de Weibull. La génération de cette première loi, qui peut être interprétée comme le temps entre la naissance et l'inclusion, n'a pas supposé de valeur minimale. Les jeux de données simulées avaient donc une taille de base de 5000 sujets et les 3500 premiers sujets ayant un âge d'entrée supérieur à 65 ans ont été conservés pour l'analyse.

Trois scénarios de simulation ont été générés sur la même méthodologie que le schéma A pour la censure par intervalle (voir section III.2.1.1) :

— scénario 1 : le temps de maladie et de décès est observé en temps continu.

— scénario 2 : le temps de maladie est censuré par intervalle avec une taille de l'intervalle comprise entre 2 et 3 ans.

— scénario 3 : le temps de maladie est censuré par intervalle avec une taille de l'intervalle comprise entre 3 et 6 ans.

De plus, une censure administrative à 100 ans a été faite pour le décès.

Ce schéma permet de regarder le comportement de l'approche par linéarisation sur des données tronquées à gauche et censurées par intervalle. L'approche par linéarisation permet de prendre en compte ces deux phénomènes d'observation simultanément et de pouvoir regarder l'influence de facteurs de risques sur des indicateurs épidémiologiques dépendant de l'âge (comme une espérance de vie, voir en particulier les indicateurs épidémiologiques définis en section II.2.2).

IV.2.2.2 Modélisation

Les modèles *illness-death* ont été estimés sous une hypothèse d'intensités de transition proportionnelles des deux variables explicatives Z_1 et Z_2^*, avec une méthode d'estimation par vraisemblance pénalisée et avec une méthode paramétrique. Les intensités de transition de base sont approchées par des splines avec 5 nœuds intérieurs pour chaque transition et les paramètres de lissages κ ont été choisis pour le premier échantillon et conservés pour les autres échantillons. Une première étape de validation croisée générale a été faite pour choisir un premier jeu de paramètre κ puis les valeurs de ces paramètres ont été affinées manuellement en traçant les intensités de transition de base.

Les trois indicateurs épidémiologiques qui sont étudiés dans cette partie sont le risque de démence vie-entière pour un individu non-dément de 75 ans, l'espérance de vie sans démence à 75 ans d'un sujet non-dément et l'espérance de vie totale à 75 ans d'un sujet non-dément. Ces trois indicateurs épidémiologiques sont respectivement notés $LTR(75)$, $LE_{00}(75)$, $LE_{0\cdot}(75)$. L'effet de Z_1 et Z_2 est résumé par les trois modèles suivant :

$$LTR(75 \mid Z_1, Z_2^*) \rightsquigarrow \Gamma_0 + \Gamma_1 Z_1 + \Gamma_2 Z_2^* \tag{IV.7}$$

$$LE_{00}(75 \mid Z_1, Z_2^*) \rightsquigarrow \Gamma_3 + \Gamma_4 Z_1 + \Gamma_5 Z_2^* \tag{IV.8}$$

$$LE_{01}(75 \mid Z_1, Z_2^*) \rightsquigarrow \Gamma_6 + \Gamma_7 Z_1 + \Gamma_8 Z_2^* \tag{IV.9}$$

Pour chaque indicateur épidémiologique, une estimation a été faite à partir de $m = 100$ sujets tirés aléatoirement parmi les 3500 utilisés dans l'estimation du modèle *illness-death*. Pour calculer les bandes de confiance des estimations, $J = 1000$ nouvelles valeurs ont été retirées par rapport à une loi normale multivariée.

IV.2.2.3 Calculs des valeurs théoriques

Comme l'effet des variables explicatives est simulé sur les intensités de transition (avec une hypothèse de proportionnalité des risques), l'effet des variables sur des indicateurs épidémiologiques n'est pas connu directement. Les indicateurs épidémiologiques sont des fonctions complexes des intensités de transition (et font surtout intervenir au moins deux intensités de transition), il est compliqué de calculer l'effet des variables explicatives sur ces indicateurs.

Pour répondre aux besoins d'une « vraie » valeur pour le paramètre Γ, un jeu de données de 50 000 sujets a été créé. Les trois indicateurs épidémiologiques ont été calculés à partir des variables explicatives de ce grand jeu de données en utilisant les distributions de Weibull utilisées pour générer les données. Les indicateurs épidémiologiques calculés théoriquement ont été utilisés comme variable réponse dans un modèle de régression linéaire : les estimations ponctuelles du modèle ont alors servi de valeurs « théoriques » pour la comparaison des résultats des simulations, en supposant que les variables explicatives ont un effet linéaire sur les quantités d'intérêt.

Tableau IV.2 – Résultats des simulations : comparaison de l'approche par linéarisation ($m = 100$ et $J = 1000$) pour le risque vie-entière de démence d'un sujet vivant non-dément à 75 ans ($LTR(75, Z_1, Z_2)$) suivant deux méthodes d'estimation et trois scénarios. Schéma C : 500 échantillons de 3500 sujets.

Γ	Scénario et méthode	$\overline{\hat{\Gamma}}$	Biais	ASE	ESE	RMSE × 1000	Taux de couv.
$\Gamma_0 = 0,381$	1- Splines	0,380	-0,001	0,013	0,027	13	94,8
	1- Weibull	0,381	<0,001	0,013	0,013	13	96,0
	2- Splines	0,380	-0,001	0,016	0,016	16	94,8
	2- Weibull	0,381	<0,001	0,015	0,015	15	94,6
	3- Splines	0,380	-0,001	0,016	0,015	16	95,2
	3- Weibull	0,381	<0,001	0,015	0,015	15	94,8
$\Gamma_1 = 0,201$	1- Splines	0,199	-0,002	0,016	0,021	17	93,4
	1- Weibull	0,200	-0,001	0,016	0,017	16	94,2
	2- Splines	0,200	-0,001	0,020	0,020	20	93,6
	2- Weibull	0,200	-0,001	0,020	0,020	20	93,0
	3- Splines	0,200	-0,001	0,020	0,020	20	92,4
	3- Weibull	0,200	-0,001	0,020	0,020	20	93,2
$\Gamma_2 = 0,018$	1- Splines	0,018	<0,001	0,002	0,003	2	95,4
	1- Weibull	0,018	<0,001	0,002	0,002	2	94,4
	2- Splines	0,018	<0,001	0,003	0,003	3	94,8
	2- Weibull	0,018	<0,001	0,003	0,003	3	94,0
	3- Splines	0,018	<0,001	0,003	0,003	3	94,4
	3- Weibull	0,018	<0,001	0,003	0,003	3	95,0

scénario 1 sans censure par intervalle
scénario 2 avec censure par intervalle et 2,5 ans entre deux visites en moyennes
scénario 3 avec censure par intervalle et 4,5 ans entre deux visites en moyennes
$\overline{\hat{\Gamma}}$ = moyenne des estimations
ASE = écart-type asymptotique / ESE = écart-type empirique
RMSE = racine carrée de l'erreur quadratique moyenne
Taux de couv. = pourcentage des 500 échantillons ayant $\Gamma \in \mathrm{BC}_{95\%}(\hat{\Gamma})$

IV.2.2.4 Résultats

Pour le risque de démence vie-entière à 75 ans, les résultats sont sensiblement comparable entre la méthode semi-paramétrique et la méthode paramétrique. Le tableau IV.2 montre qu'un individu vivant non-dément à 75 ans ayant $Z_1 = 0$ et $Z_2^* = 24$ a 38% de chance de développer une démence jusqu'à son décès. Le fait d'avoir $Z_1 = 1$ augmente la probabilité de développer une démence de 20% par rapport à $Z_1 = 0$, à niveau de Z_2^* équivalent. La probabilité de développer une démence avant de décéder pour un sujet non-dément de 75 ans augmente de 1,8% pour une augmentation d'une unité de Z_2^*. Les résultats des simulations montrent des biais très faibles pour les trois paramètres, les trois scénarios de simulation et les deux méthodes d'estimation. Les RMSE sont beaucoup plus faible pour le paramètre de régression de Z_2^*. D'une manière générale, les RMSE augmentent au fur et à mesure que la taille de l'intervalle de censure de l'âge de démence augmente. Les taux de couvertures sont bons sauf pour Γ_1 pour le scénario 2 estimé à partir d'une méthode paramétrique et le scénario 3 pour la méthode semi-paramétrique.

Tableau IV.3 – Résultats des simulations : comparaison de l'approche par linéarisation ($m = 100$ et $J = 1000$) pour l'espérance de vie sans démence d'un sujet vivant non-dément à 75 ans ($LE_{00}(75, Z_1, Z_2)$) suivant deux méthodes d'estimation et trois scénarios. Schéma C : 500 échantillons de 3500 sujets.

Γ	Scénario et méthode	$\overline{\hat{\Gamma}}$	Biais	ASE	ESE	RMSE × 1000	Taux de couv.
$\Gamma_3 = 10,039$	1- Splines	10,128	0,089	0,170	0,958	192	94,6
	1- Weibull	10,058	0,019	0,157	0,157	158	93,0
	2- Splines	10,067	0,028	0,174	0,166	176	95,6
	2- Weibull	10,063	0,024	0,157	0,161	159	92,0
	3- Splines	10,068	0,029	0,174	0,166	176	95,4
	3- Weibull	10,062	0,022	0,157	0,159	159	92,8
$\Gamma_4 = 1,444$	1- Splines	1,422	-0,022	0,200	0,210	201	95,2
	1- Weibull	1,446	0,003	0,197	0,186	197	96,0
	2- Splines	1,431	-0,013	0,204	0,191	204	95,8
	2- Weibull	1,442	-0,002	0,200	0,196	200	95,0
	3- Splines	1,430	-0,014	0,204	0,189	204	96,0
	3- Weibull	1,443	-0,001	0,200	0,196	200	95,6
$\Gamma_5 = -0,243$	1- Splines	-0,242	0,001	0,030	0,035	30	93,8
	1- Weibull	-0,245	-0,002	0,029	0,032	30	93,0
	2- Splines	-0,244	-0,001	0,031	0,032	31	93,2
	2- Weibull	-0,245	-0,001	0,030	0,032	30	92,8
	3- Splines	-0,244	-0,001	0,031	0,032	31	92,4
	3- Weibull	-0,244	-0,001	0,030	0,032	30	92,4

scénario 1 sans censure par intervalle

scénario 2 avec censure par intervalle et 2,5 ans entre deux visites en moyennes

scénario 3 avec censure par intervalle et 4,5 ans entre deux visites en moyennes

$\overline{\hat{\Gamma}}$ = moyenne des estimations

ASE = écart-type asymptotique / ESE = écart-type empirique

RMSE = racine carrée de l'erreur quadratique moyenne

Taux de couv. = pourcentage des 500 échantillons ayant $\Gamma \in \mathrm{BC}_{95\%}(\hat{\Gamma})$

Tableau IV.4 – Résultats des simulations : comparaison de l'approche par linéarisation ($m = 100$ et $J = 1000$) pour l'espérance de vie totale d'un sujet vivant non-dément à 75 ans ($LE_{0.}(75, Z_1, Z_2)$) suivant deux méthodes d'estimation et trois scénarios. Schéma C : 500 échantillons de 3500 sujets.

Γ	Scénario et méthode	$\overline{\hat{\Gamma}}$	Biais	ASE	ESE	RMSE $\times$ 1000	Taux de couv.
$\Gamma_6 = 11,902$	1- Splines	11,994	0,093	0,177	0,843	200	94,2
	1- Weibull	11,931	0,029	0,171	0,165	173	95,0
	2- Splines	11,937	0,035	0,182	0,172	185	94,4
	2- Weibull	11,932	0,031	0,173	0,167	176	93,8
	3- Splines	11,937	0,035	0,182	0,172	185	94,6
	3- Weibull	11,932	0,030	0,173	0,166	176	94,2
$\Gamma_7 = 3,059$	1- Splines	3,023	-0,036	0,213	0,279	216	96,4
	1- Weibull	3,062	0,003	0,213	0,202	213	97,0
	2- Splines	3,040	-0,019	0,216	0,203	217	96,8
	2- Weibull	3,060	0,001	0,216	0,206	216	96,6
	3- Splines	3,039	-0,020	0,216	0,203	217	97,2
	3- Weibull	3,061	0,002	0,216	0,206	216	95,6
$\Gamma_8 = -0,140$	1- Splines	-0,140	<0,001	0,033	0,035	33	93,8
	1- Weibull	-0,143	-0,002	0,032	0,033	32	93,0
	2- Splines	-0,141	-0,001	0,033	0,033	33	94,2
	2- Weibull	-0,142	-0,002	0,033	0,034	33	94,0
	3- Splines	-0,141	-0,001	0,033	0,033	33	94,0
	3- Weibull	-0,142	-0,001	0,033	0,034	33	93,6

scénario 1 sans censure par intervalle
scénario 2 avec censure par intervalle et 2,5 ans entre deux visites en moyennes
scénario 3 avec censure par intervalle et 4,5 ans entre deux visites en moyennes
$\overline{\hat{\Gamma}}$ = moyenne des estimations
ASE = écart-type asymptotique / ESE = écart-type empirique
RMSE = racine carrée de l'erreur quadratique moyenne
Taux de couv. = pourcentage des 500 échantillons ayant $\Gamma \in \mathrm{BC}_{95\%}(\hat{\Gamma})$

L'espérance de vie sans démence à 75 ans est estimée à environ 10 ans sur ce schéma de simulation (voir Γ_3 du tableau IV.3) pour un sujet avec $Z_1 = 0$ et $Z_2 = 24$. Avoir la variable Z_1 égale à 0 augmente l'espérance de vie sans démence à 75 ans d'un peu moins d'un an et demi. La variable Z_2^* a un effet délétère sur cette espérance de vie, avec une diminution de 2 mois de l'espérance de vie pour une augmentation d'une unité de Z_2^*. Les biais semblent plus importants pour Γ_3 et Γ_4 mais cela est dû à la valeur du paramètre : le biais relatif pour Γ_3 pour le scénario 1 estimé par splines est inférieur à 1%. Cette remarque est également valable pour les RMSE. Les biais sont toujours plus importants avec la méthode semi-paramétrique que la méthode paramétrique. Les écart-types asymptotiques et empiriques sont proches, sauf pour la première ligne du tableau IV.3 car quelques échantillons ont estimés $\Gamma_3 = 25$. Les taux de couvertures sont bons pour la méthode semi-paramétrique. Pour la méthode paramétrique, les taux de couverture sont raisonnables pour l'intercept et le paramètre de régression associé à Z_2^* et sont bons pour le paramètre de régression de Z_1. Les résultats ne se sont pas dégradés en présence de censure par intervalle.

Les résultats pour l'espérance de vie totale d'un sujet non-dément de 75 ans sont semblables

à ceux de l'espérance de vie sans démence. L'espérance de vie à 75 ans vaut presque 12 ans pour un sujet ayant comme caractéristiques $Z_1 = 0$ et $Z_2 = 24$, elle augmente de 3 ans pour ceux qui ont $Z_1 = 1$ par rapport aux autres et Z_2 a un effet négatif sur cette espérance de vie en diminuant légèrement l'espérance de vie lorsque Z_2 augmente d'une unité. Le biais montre en général une sur-estimation des paramètres de régression. Il faut noter qu'ici certains taux de couvertures sont trop importants (comme par exemple pour Γ_7 estimé paramétriquement dans les scénarios 1 et 3, avec des valeurs à 97% et 97,2%).

IV.3 Application

L'application de l'approche par linéarisation a été faite à partir des données de la cohorte PAQUID. Les sujets de la cohorte PAQUID n'ont pas été inclus aux mêmes âges donc plusieurs illustrations sont proposées pour tenir compte ou non de la troncature à gauche.

IV.3.1 Effet du niveau d'éducation et du sexe

La première illustration s'est concentrée sur l'effet du certificat d'étude primaire et du sexe sur le risque vie-entière de démence à 75 ans, les espérances de vie totale et sans démence d'un sujet non-dément à 75 ans.

IV.3.1.1 Modélisation

Ces trois indicateurs ont un sens épidémiologiques s'ils sont calculés pour un âge donné. L'âge a donc été utilisé comme temps de base dans cette illustration, en tenant compte de l'entrée retardée dans la cohorte.

Le modèle *illness-death* a été estimé par des splines avec stratification sur les variables *sexe* et *cep*, pour ne pas supposer d'effet proportionnel de ces variables :

$$\alpha_{01}(t, sexe, cep) = \begin{cases} \alpha_{01,0,\bullet 1}(t) \text{ si } sexe = 0 \text{ et } cep = 0 \\ \alpha_{01,0,\bullet 2}(t) \text{ si } sexe = 0 \text{ et } cep = 1 \\ \alpha_{01,0,\bullet 3}(t) \text{ si } sexe = 1 \text{ et } cep = 0 \\ \alpha_{01,0,\bullet 4}(t) \text{ si } sexe = 1 \text{ et } cep = 1 \end{cases}$$

$$\alpha_{02}(t, sexe, cep) = \begin{cases} \alpha_{02,0,\bullet 1}(t) \text{ si } sexe = 0 \text{ et } cep = 0 \\ \alpha_{02,0,\bullet 2}(t) \text{ si } sexe = 0 \text{ et } cep = 1 \\ \alpha_{02,0,\bullet 3}(t) \text{ si } sexe = 1 \text{ et } cep = 0 \\ \alpha_{02,0,\bullet 4}(t) \text{ si } sexe = 1 \text{ et } cep = 1 \end{cases}$$

$$\alpha_{12}(t, sexe, cep) = \begin{cases} \alpha_{12,0,\bullet 1}(t) \text{ si } sexe = 0 \text{ et } cep = 0 \\ \alpha_{12,0,\bullet 2}(t) \text{ si } sexe = 0 \text{ et } cep = 1 \\ \alpha_{12,0,\bullet 3}(t) \text{ si } sexe = 1 \text{ et } cep = 0 \\ \alpha_{12,0,\bullet 4}(t) \text{ si } sexe = 1 \text{ et } cep = 1 \end{cases}$$

Le modèle *illness-death* $\Phi(s, t, Z)$ est donc composé de 4 sous-modèles

$$\Phi(s, t, Z) = \left\{\Phi^{\bullet 1}(s, t), \Phi^{\bullet 2}(s, t), \Phi^{\bullet 3}(s, t), \Phi^{\bullet 4}(s, t)\right\}$$

.

Trois modèles linéaires pondérés ont été estimé avec $J = 1000$ retirages d'une loi normale multivarié pour chaque modèle stratifié (c.à.d. $\varphi^{\bullet 1} \sim \mathcal{NM}(\hat{\varphi}^{\bullet 1}; \hat{V}_{\varphi^{\bullet 1}})$, de même pour $\varphi^{\bullet 2}$, etc...)

$$\begin{aligned} LTR(75, sexe, cep) &\leftrightsquigarrow \Gamma_0 + \Gamma_1 \times sexe + \Gamma_2 \times cep \\ LE_{00}(75, sexe, cep) &\leftrightsquigarrow \Gamma_3 + \Gamma_4 \times sexe + \Gamma_5 \times cep \\ LE_{0.}(75, sexe, cep) &\leftrightsquigarrow \Gamma_6 + \Gamma_7 \times sexe + \Gamma_8 \times cep \end{aligned}$$

Dans cette illustration, 4 valeurs différentes (pour m) ont été calculés pour chaque indicateur épidémiologique pour représenter les 4 modalités du couple de variable *sexe* et *cep*.

IV.3.1.2 Descriptif de l'échantillon

Les sujets de cet échantillon de PAQUID avaient entre 65 ans et 101 ans (âge médian à 74 ans et 9 mois), la méthode de linéarisation a utilisé l'âge comme temps de base avec entrée retardée (c.à.d. avec troncature à gauche). Cette illustration a été conduite sur un échantillon de 3673 sujets non déments à l'inclusion. L'échantillon était composé de 13% d'hommes sans CEP, de 29% d'hommes avec CEP, de 22% de femmes sans CEP et de 36% de femmes avec CEP. Le suivi à 27

ans a été utilisé : 25% des sujets ont eu un diagnostic de démence au cours du suivi et 7% des sujets étaient toujours vivant 27 ans après l'inclusion.

IV.3.1.3 Résultats

Les résultats des estimations de cette illustration sont présentés dans le tableau IV.5 pour les trois indicateurs épidémiologiques. D'après ce dernier, un homme vivant non-dément à 75 ans, sans certificat d'étude primaire a un risque de démence vie-entière égal à 40%, c'est-à-dire qu'il a 40% de risque de développer une démence avant son décès. Son espérance de vie sans démence est d'un peu plus de 8 ans et son espérance de vie totale est d'un peu moins de 10 ans.

Le sexe a une influence significative sur ces trois indicateurs épidémiologiques : à niveau d'éducation similaire, les femmes ont 20% de chance en plus de développer démence avant de décéder par rapport aux hommes, si elles ont atteint l'âge de 75 ans sans démence. Elles ont leur espérance de vie sans démence qui augmente en moyenne d'environ 2 ans par rapport aux hommes, pour deux sujets de 75 ans sans démence. Les femmes ont aussi une espérance de vie plus importante que les hommes : elles peuvent espérer vivre 3 ans et 4 mois de plus que les hommes, pour celles vivantes non-démentes à 75 ans.

Le niveau d'éducation joue significativement un rôle protecteur dans cette illustration. Pour deux personnes vivantes et non-démentes à 75 ans, de même sexe, celle qui a au moins le CEP comme diplôme voit son risque de démence vie-entière diminué de 0,06 par rapport à celle qui n'a pas le CEP. L'espérance de vie sans démence et l'espérance de vie totale à 75 ans sont augmentées de moins de 2 ans pour une personne qui a le CEP par rapport à une personne qui ne l'a pas, ajusté sur le sexe, d'après le tableau IV.5.

IV.3.2 Effet du score au MMS

La seconde illustration a été faite pour comparer les résultats de l'approche par linéarisation aux résultats de l'approche par pseudo-valeurs. L'échantillon est donc le même que celui utilisé dans l'application du chapitre précédant dont le descriptif a été présenté en section III.3.2.

Pour rappel, l'objectif de cette application est de mesurer l'effet du MMS sur le risque absolu de démence, la probabilité d'être vivant non-dément et la moyenne restreinte des temps de survie sans démence à 10 ans après l'inclusion, ajusté sur l'âge à l'inclusion, le sexe et le niveau d'étude (modélisé par avoir ou ne pas avoir le CEP). Soient mms^* et age^* les variables telles que $mms^* = mms - 26$ et $age^* = age - 75$

Le modèle *illness-death* estimé dans la première étape de l'approche par linéarisation a été

Tableau IV.5 – Modélisation du risque vie entière de démence ($LTR(75, sexe, CEP)$), de l'espérance de vie sans démence ($LE_{00}(75, sexe, cep)$) et de l'espérance de vie totale ($LE_{0.}(75, sexe, cep)$) pour un sujet vivant non-dément de 75 ans en fonction du sexe et du niveau d'éducation. Estimation par linéarisation ($m = 4$ et $J = 1000$) d'après 3673 sujets de la cohorte PAQUID.

	$\hat{\Gamma}$	$\text{BC}_{95\%}(\hat{\Gamma})$
Risque vie-entière de démence		
Homme sans CEP	0.416	[0.363 ; 0.477]
Femmes versus hommes	0.213	[0.148 ; 0.261]
Niveau d'éducation (avec vs. sans CEP)	-0.065	[-0.117 ;-0.004]
Espérance de vie sans démence		
Homme sans CEP	8.141	[7.534 ; 8.516]
Femmes versus hommes	1.959	[1.490 ; 2.458]
Niveau d'éducation (avec vs. sans CEP)	1.861	[1.329 ; 2.389]
Espérance de vie totale		
Homme sans CEP	9.921	[9.314 ;10.284]
Femmes versus hommes	3.350	[2.827 ; 3.803]
Niveau d'éducation (avec vs. sans CEP)	1.137	[0.627 ; 1.687]

BC : bande de confiance issue des 2,5^{e} et 97,5^{e}percentiles

stratifié sur le sexe avec :

$$\alpha_{01}(t \mid sexe^{\bullet}, Z) = \mathbb{1}_{\{sexe=0\}}\alpha_{01,0,\bullet=0}(t)e^{\beta_{01,\bullet=0}{}^{\top}Z} + \mathbb{1}_{\{sexe=1\}}\alpha_{01,0,\bullet=1}(t)e^{\beta_{01,\bullet=1}{}^{\top}Z} \quad \text{(IV.10)}$$

$$\alpha_{02}(t \mid sexe^{\bullet}, Z) = \mathbb{1}_{\{sexe=0\}}\alpha_{02,0,\bullet=0}(t)e^{\beta_{02,\bullet=0}{}^{\top}Z} + \mathbb{1}_{\{sexe=1\}}\alpha_{02,0,\bullet=1}(t)e^{\beta_{02,\bullet=1}{}^{\top}Z} \quad \text{(IV.11)}$$

$$\alpha_{12}(t \mid sexe^{\bullet}, Z) = \mathbb{1}_{\{sexe=0\}}\alpha_{12,0,\bullet=0}(t)e^{\beta_{12,\bullet=0}{}^{\top}Z} + \mathbb{1}_{\{sexe=1\}}\alpha_{12,0,\bullet=1}(t)e^{\beta_{12,\bullet=1}{}^{\top}Z} \quad \text{(IV.12)}$$

avec Z correspondant aux variables $sexe$, MMS^*, age^* et cep. On peut voir d'après équations (IV.10) à (IV.12) que le score au MMS et l'âge ont été modélisés suivant une hypothèse de proportionnalité et de log-linéarité.

Les modèles suivant ont été estimés respectivement pour le risque absolu de démence, pour la probabilité d'être vivant non-dément et pour la moyenne restreinte des temps de survie sans démence, en supposant un effet linéaire des quatre variables explicatives :

$$F_{01}(10 \mid MMS^*, age^*, sexe, cep) \leftrightsquigarrow \Gamma_0 + \Gamma_1 MMS^* + \Gamma_2 age^* + \Gamma_3 sexe + \Gamma_4 cep \quad \text{(IV.13)}$$

$$P_{00}(10 \mid MMS^*, age^*, sexe, cep) \leftrightsquigarrow \Gamma_5 + \Gamma_6 MMS^* + \Gamma_7 age^* + \Gamma_8 sexe + \Gamma_9 cep \quad \text{(IV.14)}$$

$$RM(10 \mid MMS^*, age^*, sexe, cep) \leftrightsquigarrow \Gamma_{10} + \Gamma_{11} MMS^* + \Gamma_{12} age^* + \Gamma_{13} sexe + \Gamma_{14} cep \quad \text{(IV.15)}$$

avec un retirage de $J = 1000$ valeurs pour $\varphi^{\bullet=0}$ et $J = 1000$ valeurs pour $\varphi^{\bullet=1}$ d'après des lois normale multivariées. Le nombre de valeurs différentes de θ était de $m = 200$ sujets dont $m_0 = 82$ et $m_1 = 118$ pour respecter la proportion des sujets utilisés dans les modèles *illness-death* stratifiés sur le sexe.

Cet échantillon de la cohorte PAQUID a utilisé les mêmes données que dans l'application du chapitre des pseudo-valeurs (voir section III.3.2 pour un descriptif détaillé).

IV.3.2.1 Résultats

Tableau IV.6 – Modélisation du risque absolu de démence ($F_{01}(10, sexe, CEP)$), de la probabilité d'être vivant non-dément ($P_{00}(10, sexe, cep)$) et de la moyenne restreinte des temps de survie sans démence ($RM(10, sexe, cep)$) à 10 ans après l'inclusion, en fonction du score au MMS, de l'âge à l'inclusion, du sexe et du niveau d'éducation. Estimation par linéarisation ($m = 200$ et $J = 1000$) d'après 2641 sujets de la cohorte PAQUID.

	$\hat{\Gamma}$	$BC_{95\%}(\hat{\Gamma})$
Risque absolu de démence		
Homme sans CEP de 75 ans et 26 points de MMS	0,188	[0,153 ; 0,246]
MMS à l'inclusion (points)	-0,017	[-0,024 ;-0,009]
Age à l'inclusion (année)	0,012	[0,009 ; 0,016]
Femmes versus hommes	0,064	[0,008 ; 0,104]
Niveau d'éducation (avec / sans CEP)	-0,019	[-0,067 ; 0,027]
Probabilité d'être vivant non-dément		
Homme sans CEP de 75 ans et 26 points de MMS	0,422	[0,384 ; 0,448]
MMS à l'inclusion (points)	0,019	[0,014 ; 0,023]
Age à l'inclusion (année)	-0,029	[-0,031 ;-0,028]
Femmes versus hommes	0,101	[0,070 ; 0,133]
Niveau d'éducation (avec / sans CEP)	0,015	[-0,016 ; 0,047]
Moyenne restreinte des temps de survie sans démence		
Homme sans CEP de 75 ans et 26 points de MMS	7,084	[6,797 ; 7,274]
MMS-26 à l'inclusion (en points)	0,138	[0,105 ; 0,169]
Age-75 à l'inclusion (en années)	-0,182	[-0,197 ;-0,172]
Femmes versus hommes	0,812	[0,604 ; 1,059]
Niveau d'éducation (avec / sans CEP)	-0,012	[-0,214 ; 0,193]

BC : bande de confiance issue des 2,5^{e} et 97,5^{e} percentiles

Pour cette illustration, les résultats sont présentés dans le tableau IV.6. Un homme de 75 ans avec un score de 26 points au moment de l'inclusion a une probabilité de 19% de développer une démence dans les 10 ans suivant l'inclusion, il a une probabilité d'être vivant non-dément de 42% à 10 ans de suivi et il peut espérer rester vivant sans démence environ 7 ans dans les 10 ans suivant l'inclusion.

Le score au MMS à l'inclusion a montré une diminution de 0,01 de la probabilité d'avoir développé une démence dans les 10 ans après l'inclusion pour une personne ayant un point de MMS en plus qu'une autre personne, ajusté sur l'âge à l'inclusion, le sexe et le niveau d'éducation. Une différence d'un an d'âge entre deux sujets à l'inclusion entraine une augmentation de 1% de la probabilité de développer une démence dans les 10 ans de suivi pour le sujet le plus âgés des deux. Les femmes ont une probabilité de 6% de plus de développer une démence que les hommes, toutes choses égales par ailleurs. Le certificat d'étude primaire ne change pas le risque absolu de démence à 10 ans après l'inclusion, ajusté sur les autres variables du modèle.

Pour la probabilité d'être vivant non-dément à 10 ans de suivi, l'effet du MMS va dans le

sens inverse de l'effet pour le risque absolu de démence : pour une augmentation d'un point de MMS à l'inclusion, la probabilité d'être vivant non-dément augmente de 0,02. La probabilité d'être vivant non-dément à $t = 10$ change en fonction de l'âge à l'inclusion. Deux sujets ayant un an de différence à l'inclusion ont une probabilité différente de 0,029 : le sujet le plus vieux a une probabilité diminuée par rapport au plus jeune. Les femmes ont 10% de probabilité en plus que les hommes d'être vivantes et non-démentes. Ces trois augmentations ou diminutions sont significatives et sont ajustées sur les autres variables du modèle.

Pour la moyenne restreinte des temps de survie en bonne santé, les résultats vont dans le mêmes sens que pour la probabilité d'être vivant non-dément, bien que les gains ou pertes soient tous inférieur à 1 an. Ici, le score au MMS a un impact significatif sur le temps passé dans l'état 0 dans un horizon de 10 ans suivant l'inclusion : lorsque le MMS augmente de 1 point, l'espérance du temps passé dans l'état 0 augmente d'un peu plus d'un mois et demi. Lorsque l'on compare deux sujets qui ont une différence d'un an à l'inclusion, le plus jeune peut espérer rester vivant non-dément deux mois de plus que le plus âgés des deux. Les femmes peuvent espérer vivre sans démence 9 mois de plus que les hommes dans un horizon de 10 ans après l'inclusion, ajusté sur le score au MMS et l'âge à l'inclusion et le niveau d'éducation.

IV.3.3 Effet de l'âge à la ménopause

Cette illustration a été faite pour pouvoir utiliser des variables quantitatives lorsque l'âge est utilisé comme temps de base. Comme les sujets de l'étude PAQUID ont été inclus à des âges différents, les variables mesurées à l'inclusion n'avaient pas nécessairement les mêmes valeurs 10 ans plus tôt. En particulier, les variables quantitatives comme le score au MMS sont dépendantes de l'âge. Par exemple, si un sujet est inclus à 80 ans et que l'on utilise son score au MMS mesuré à l'inclusion comme variable explicative, on fait l'hypothèse que son score au MMS à 80 ans ait été le même qu'à 65 ans. Cette hypothèse étant trop forte, il n'a pas été fait de modèle avec entrée retardée avec des variables quantitatives dépendantes de l'âge (comme des scores psychométriques). Dans cette illustration, nous proposons de travailler sur les femmes incluses dans l'étude PAQUID et de regarder l'effet de l'âge à la ménopause sur l'espérance de vie sans démence et l'espérance de vie totale, ajusté sur le fait d'avoir eu ou non des enfants. L'intérêt de la variable « âge à la ménopause » est d'avoir une valeur constante au cours du temps pour des femmes entrées dans l'étude après 65 ans.

IV.3.3.1 Modélisation

Le modèle *illness-death* dans la première étape de la linéarisation a été estimé en utilisant une hypothèse de proportionnalité et de log-linéarité de l'âge à la ménopause ($agem$) ajusté sur le fait

d'avoir eu des enfants (*mere*) en utilisant l'âge comme temps de base :

$$\alpha_{01}(t \mid agem_i, mere_i) = \alpha_{01,0}(t) \exp[\beta_{01,1} \times agem_i + \beta_{01,2} \times mere_i]$$
$$\alpha_{02}(t \mid agem_i, mere_i) = \alpha_{02,0}(t) \exp[\beta_{02,1} \times agem_i + \beta_{02,2} \times mere_i]$$
$$\alpha_{12}(t \mid agem_i, mere_i) = \alpha_{12,0}(t) \exp[\beta_{12,1} \times agem_i + \beta_{12,2} \times mere_i]$$

avec *agem* l'âge à la ménopause - 50 et *mere* qui vaut 1 si les femmes ont eu au moins un enfant, 0 sinon.

Les estimations du maximum de vraisemblance pénalisée du modèle *illness-death* ont été approchées par des M-splines cubiques avec 5 nœuds intérieur. Le choix du paramètre de lissage κ a été choisi en fonction des représentations graphiques des intensités de transition de base $\alpha_{kl,0}(.)$.

L'effet de l'âge à la ménopause sur l'espérance de vie sans démence et l'espérance de vie totale à 70 ans, ajusté sur le fait d'avoir eu des enfants a été estimé par :

$$LE_{00}(70 \mid agem, mere) \leftrightsquigarrow \Gamma_0 + \Gamma_1 agem + \Gamma_2 mere$$
$$LE_{0.}(70 \mid agem, mere) \leftrightsquigarrow \Gamma_3 + \Gamma_4 agem + \Gamma_5 mere$$

Cent femmes ont été tirées au sort de manière aléatoire et un retirage de 1000 nouvelles valeurs du vecteur φ a été fait.

IV.3.3.2 Descriptif de l'échantillon

Cette illustration s'est basée sur l'échantillon des femmes de la cohorte PAQUID n'ayant pas eu une ménopause précoce (c.à.d. un âge à la ménopause supérieur à 40 ans). Le suivi à 27 ans a été utilisé pour cette illustration : les femmes étaient inclus entre 65 et 97 ans, 31% d'entre elles ont reçu un diagnostic de démence entre l'inclusion et le suivi à 27 ans. Au moment de la censure administrative, 9% étaient toujours vivantes. L'âge moyen à l'inclusion était de 75 ans (écart-type de 7,0 ans, au minimum 65 ans et au maximum 97 ans). Un histogramme de la distribution pour l'âge à l'inclusion et l'âge à la ménopause sont présentés en figure C.2 en annexe.
L'échantillon était composé de 1909 femmes, avec un âge moyen à la ménopause de 50 ans (au minimum 40 ans et au maximum 65 ans et un écart type de 4,32 ans). Parmi les femmes incluses, 84% d'entre elles avaient au moins un enfant.

IV.3.3.3 Résultats

Les résultats présentés dans le tableau IV.7 montrent que l'espérance de vie sans démence et l'espérance de vie totale à 70 ans sont respectivement de 15 ans et 18 ans pour des femmes ayant

Tableau IV.7 – Modélisation de l'espérance de vie sans démence ($LE_{00}(70, agem, mere)$) et de l'espérance de vie totale ($LE_{0.}(70, agem, mere)$) pour des femmes vivantes et non-démentes de 70 ans, en fonction de l'âge à la ménopause ajusté sur le fait d'avoir eu un enfant. Estimation par linéarisation ($m = 100$ et $J = 1000$) d'après 1909 sujets de la cohorte PAQUID.

	$\hat{\Gamma}$	$\text{BC}_{95\%}(\hat{\Gamma})$
Espérance de vie sans démence		
Femme de 50 ans à la ménopause et sans enfant	15,440	[14,569 ;16,052]
Age à la ménopause (année)	0,100	[0,030 ; 0,169]
Avoir des enfants (oui versus non)	-0,402	[-1,097 ; 0,363]
Espérance de vie totale		
Femme de 50 ans à la ménopause et sans enfant	18,0245	[17,248 ;18,600]
Age à la ménopause (année)	0,095	[0,028 ; 0,171]
Avoir des enfants (oui versus non)	-0,300	[-0,930 ; 0,397]

BC : bande de confiance issue des 2,5^{e} et 97,5^{e}percentiles

eu leur ménopause à 50 ans et n'ayant pas eu d'enfant. Plus une femme a eu une ménopause tard, plus cette espérance de vie augmente : elles gagnent environ 4 mois d'espérance de vie (sans démence ou totale) pour une différence de 4 ans d'âge à la ménopause, ajusté sur le fait d'avoir eu ou non des enfants. Cette illustration montre un effet de l'âge sur ces espérances de vie, bien que ne soit pas important (environ 36 jours d'espérance de vie en plus pour une augmentation d'un an d'âge à la ménopause). Ces résultats sont conditionnés sur des hypothèses de log-linéarité du modèle *illness-death* et de linéarité de l'effet sur chaque espérance. De plus, seulement 100 femmes ont été tirées au sort dans cette illustration : les résultats sont équivalent en augmentant la taille de m, jusqu'à utiliser toutes les femmes de l'échantillon. Les résultats pour le risque de démence vie-entière sont présentés dans le tableau D.2 en annexe car l'âge à la ménopause ou le fait d'avoir eu des enfants n'influence pas cet indicateur épidémiologique .

IV.4 Conclusion et discussion

L'approche par linéarisation proposée dans ce chapitre permet d'estimer l'effet de variables explicatives dans le cadre d'un modèle *illness-death* avec temps de maladie censurés par intervalle et troncature à gauche du temps d'entrée dans l'état sain. Cette approche originale a été proposée par Daniel Commenges et permet d'estimer l'effet de variables explicatives sur des quantités complexes (c'est-à-dire des fonctions de plusieurs intensités de transition) par linéarisation de l'estimateur du maximum de vraisemblance. Elle a les mêmes objectifs que l'approche par pseudo-valeurs mais semble plus intuitive que l'approche par pseudo-valeurs car les pseudo-valeurs n'ont pas un sens « palpable », contrairement à l'approche par linéarisation. L'approche par linéarisation utilise des indicateurs estimés en fonction des variables explicatives, ce qui a l'avantage d'avoir une interprétation directe. Dans ce chapitre, deux types d'estimations ont été utilisés pour calculer le maximum de vraisemblance : un estimateur semi-paramétrique de la vraisemblance pénalisée avec une approximation des intensités de transition par des splines et un estimateur paramétrique

qui suppose que les intensités de transition suivent des distributions de Weibull. Ces deux types d'estimateurs permettent de prendre en compte la censure par intervalle du temps de maladie d'un modèle *illness-death*, la troncature à gauche du temps d'entrée dans l'état sain et la non-observation de toutes les transitions (en particulier les sujets décédés après être passé par l'état 1 sans avoir été diagnostiqué). Cette approche permet de quantifier l'effet de variables explicatives sur des espérances de vie par exemple. Ces espérances de vie sont des indicateurs épidémiologiques qui dépendent de l'âge, la prise en compte de l'entrée retardée dans les études de cohorte est un des enjeux des modèles de régression pour ce type d'indicateur.

Estimer les intensités de transition en fonction des variables explicatives est l'inconvénient majeur de cette approche. C'est aussi la plus grande différence entre cette approche et l'approche par pseudo-valeurs au niveau des intensités de transition car l'approche par pseudo-valeurs estime le modèle *illness-death* marginalement.
La majorité des modèles utilisés dans ce chapitre ont supposé un effet proportionnel pour au moins une variable explicative. Des modèles *illness-death* stratifiés sur une ou deux variables explicatives binaires ont été estimés dans les illustrations pour relâcher cette hypothèse. Lorsque des variables quantitatives ont été utilisées, une hypothèse de log-linéarité s'ajoute à l'hypothèse de proportionnalité des risques. Cette hypothèse de log-linéarité peut être réduite en utilisant par exemple des bases de spline sur les variables explicatives.

Dans ce chapitre, des modèles linéaires ont été estimés dans l'étape 3 et l'étape 4 de l'approche par linéarisation pour résumer l'effet des variables explicatives sur les indicateurs épidémiologiques de la démence. Des modèles plus complexes (comme des modèles additifs généralisés par exemple) pourraient être estimés (tout en faisant de nouvelles investigations pour regarder le comportement des GAM par des simulations).

Nous avons limité volontairement les modélisations proposées dans ce chapitre (comme les hypothèses de (log-)linéarité des variables explicatives sur les intensités de transition et les indicateurs épidémiologiques) car nous trouvons qu'il vaut mieux proposer une approche simple dans un premier temps. Dans un second temps, des raffinements pourront être proposés (comme l'utilisation de GAM ou l'utilisation de splines pour estimer les effets des variables dans le modèle *illness-death*).

Le calcul des bandes de confiance supposent qu'asymptotiquement, les paramètres des intensités de transition suivent une loi normale multivariée. Cette hypothèse permet de calculer des bandes de confiances des paramètres de régression. Nous proposons ici de choisir l'estimation de l'étape 3 comme valeur ponctuelle des paramètres de régression. Un autre choix aurait aussi pu être fait par la moyenne des estimations des $J+1$ modèles estimés par les étapes 3 et 4. Pareillement, nous proposons d'utiliser les percentiles de la distribution des paramètres de régression $\hat{\Gamma}$ sur les $(J+1)$ estimations du modèle linéaire. Une autre idée pour calculer les bandes de confiance aurait été d'utiliser la formule d'un intervalle de confiance classique. Ces deux alternatives supposent que pour un paramètre Γ_q sa distribution sur les $J+1$ modèles linéaires suit une loi normale.

Cette approche par linéarisation fait aussi face aux choix du nombre de sujets à retirer (choix de m) et du nombre de paramètres φ à retirer (choix de J). Plus m et J augmentent, plus les

temps de calculs augmentent. Par exemple, pour $m = 100$ et $J = 1000$, il y a un total de 10^5 intégrales à calculer pour un indicateur de santé (voir 10^5 doubles intégrales pour l'espérance de vie totale) et J modèles de régression sont aussi à estimer. De l'autre côté, plus m augmente, plus les estimations sont précises, il y a donc un choix à faire entre précision et temps de calcul. Bien sûr, il est possible de choisir m égale à n. Choisir $m = n$ a été fait en analyse de sensibilité pour l'illustration de l'effet de l'âge à la ménopause sur les deux espérances de vie ; les résultats sont sensiblement les mêmes (que ce soit en terme d'estimation ou en terme de significativité).

Les illustrations ont dû tenir compte de l'entrée retardée dans la cohorte PAQUID. En particulier, les variables biologiques (comme l'IMC ou la tension artérielle) et les scores psychométriques sont des variables dépendantes de l'âge. À cause de l'entrée retardée, ces variables quantitatives n'ont pas été utilisées comme variables d'ajustement dans le modèle *illness-death* pour ne pas faire d'hypothèses épidémiologiques trop fortes. Cette approche est donc adéquate pour d'autres pathologies où les variables explicatives quantitatives ne seraient pas dépendantes de l'âge lorsque le temps de base est celui-ci et en présence de troncature à gauche.

L'approche par linéarisation a montré des effets du score au MMS sur les indicateurs épidémiologiques calculés pour un horizon fini. Plus le score au MMS est élevé à l'inclusion, plus la probabilité de développer une démence diminue et plus la probabilité d'être vivant non-dément et la moyenne restreinte des temps de survie à 10 ans de suivi augmentent. Ces résultats vont dans le mêmes sens que les résultats énoncés par l'approche par pseudo-valeurs. De même, les estimations sont proches entre les deux approches.

En conclusion, à partir du moment où un indicateur de santé peut être estimé en fonction des variables explicatives et qu'une matrice de variance-covariance est disponible, il est possible de résumer l'effet de ces variables explicatives par des modèles plus ou moins complexes. L'approche a été proposée pour des données censurées par intervalle mais rien n'indique *a priori* qu'il n'est pas possible d'utiliser des estimateurs non-paramétriques lorsque la censure par intervalle n'est pas présente.

Chapitre V

Discussion générale

V.1 Résumé des approches

Dans cette thèse, deux approches différentes ont été proposées pour quantifier, mesurer ou bien résumer l'effet de déterminants de la démence sur six indicateurs épidémiologiques de cette pathologie. Ce travail a été motivé par les données de la cohorte PAQUID. Cette cohorte initiée à la fin des années 1980 a plusieurs avantages comme le suivi régulier des sujets pendant plus de 27 ans et la recherche active des cas incidents de démence des participants. À l'opposé, les études comme la cohorte PAQUID ont l'inconvénient de suivre les sujets par intermittence, ce qui entraine une censure par intervalle de l'âge d'apparition de la démence. En outre, les sujets inclus dans la cohorte PAQUID sont âgés d'au moins 65 ans : ils ont donc un risque important de décès. Ne pas observer toutes les trajectoires de chaque individu est une conséquence de la censure par intervalle en présence du risque compétitif de décès. En particulier, lorsqu'un sujet décède alors qu'il n'était pas dément à sa dernière visite de suivi, il a pu développer une démence puis décéder ou décéder directement sans démence.

Les indicateurs épidémiologiques auxquels ce travail s'est intéressé dépendent du risque de démence et du risque de décès : la relation entre les facteurs de risque de la démence et ces indicateurs n'est pas directe. Les deux approches proposées répondent donc à la volonté de travailler sur des quantités complexes définies à partir des intensités de transition d'un modèle *illness-death* tout en tenant compte de la censure par intervalle (et du fait que certaines transitions ne soient pas observées). De plus, une des deux approches a été développée pour tenir compte de l'entrée retardée des sujets dans une cohorte et permet donc de regarder l'effet de facteurs de risque sur le risque de démence vie-entière ou des espérances de vie pour un sujet vivant non-dément d'un âge donné.

Un résumé rapide des deux approches est maintenant présenté. La première approche est l'extension de l'approche par pseudo-valeurs (Andersen *et al.*, 2003) pour un modèle *illness-death* avec temps de maladie censurés par intervalle. Ce travail a fait l'objet d'une publication (Sabathé *et al.*, 2019). L'approche peut être résumée comme ceci :

Soient $\phi(t)$ et $\phi^{-i}(t)$ les intensités de transition d'un modèle *illness-death* pour temps de maladie censurés par intervalle respectivement calculés depuis l'échantillon avec tous les sujets et depuis l'échantillon sans le sujet i :

$$\phi(t) = \{\alpha_{01}(t), \alpha_{02}(t), \alpha_{12}(t)\}$$

$$\phi^{-i}(t) = \left\{\alpha_{01}^{-i}(t), \alpha_{02}^{-i}(t), \alpha_{12}^{-i}(t)\right\} \text{ pour } i = 1, ..., n$$

Soit θ une quantité d'intérêt définie à partir des intensités de transition du modèle *illness-death* :

$$\theta(t) = h\left(\phi(t)\right)$$

$$\theta^{-i}(t) = h\left(\phi^{-i}(t)\right)$$

pour $i = 1, ..., n$ et n le nombre total de sujets de l'échantillon, $\theta^{-i}(t)$ la quantité calculée depuis l'échantillon sans le sujet i et $h(.)$ une fonction définissant un indicateur épidémiologique.
Pour chaque sujet i, une pseudo-valeur Y_i est alors calculée par

$$\hat{Y}_i(t) = n \times \hat{\theta}(t) - (n-1) \times \hat{\theta}^{-i}(t)$$

avec $i = 1, ..., n$.
Les pseudo-valeurs sont utilisées comme variable réponse dans un modèle linéaire généralisé :

$$g\left(\mathbb{E}\left[\hat{Y}_i(t)\right]\right) = \gamma^\top Z$$

avec $g(.)$ une fonction de lien. Le vecteur $\hat{\gamma}$ estime l'effet de Z sur θ.

La seconde approche de ce travail est basée sur une « linéarisation » d'estimateurs du maximum de vraisemblance d'un modèle *illness-death*. La méthode peut être résumée ainsi :

Soit $\Phi(s,t,Z)$ les intensités de transition d'un modèle *illness-death* estimées conditionnellement aux variables explicatives

$$\Phi(s,t,Z) = \{\alpha_{01}(t,Z), \alpha_{02}(t,Z), \alpha_{12}(t,Z)\}$$

Soient φ le vecteur des estimateurs des paramètres associés à $\Phi(s,t,Z)$ et V_φ la matrice de variance-covariance de φ.
Soit $\varphi^{[j]}$ le j^{e} élément retiré parmi $\varphi^{[J]}$ avec $\varphi^{[J]} \sim \mathcal{NM}(\varphi; V_\varphi)$
Pour $i = 1, ..., m \leq n$, et $j = 1, ..., J$, la méthode estime les quantités d'intérêt :

$$\hat{\theta}_i(s,t,Z_i) = h(\Phi(s,t,Z))$$

$$\hat{\theta}_i^{[j]}(s,t,Z_i) = h(\Phi^{[j]}(s,t,Z))$$

avec $h(.)$ une fonction définissant un indicateur épidémiologique et $\Phi^{[j]}$ les intensités de transition du modèle *illness-death* dont les paramètres sont issus de $\varphi^{[j]}$.
L'estimation ponctuelle de l'effet de Z est faite par :

$$\mathbb{E}\left[\hat{\theta}_i(s,t,Z_i)\right] = \Gamma^\top Z$$

et l'estimation de la bande de confiance de l'effet de Z par les 2,5$^{\text{e}}$ et 97,5$^{\text{e}}$ percentiles de la distribution des J modèles :

$$\mathbb{E}\left[\hat{\theta}_i^{[j]}(s,t,Z_i)\right] = \Gamma^{[j]\top} Z$$

avec $j = 1, ..., J$.

V.2 Limites

Certaines limites de ce travail ont déjà été évoquées dans les deux chapitres présentant les approches par pseudo-valeurs et par linéarisation. Les deux approches reposent sur des estimations d'un modèle *illness-death* pour temps de maladie censurés par intervalle. La limite principale de l'approche par pseudo-valeurs est l'hypothèse d'indépendance de la censure. Les hypothèses des effets des variables sur les intensités de transition sont les limites propres à l'approche par linéarisation. Les autres limites énoncées sont communes aux deux approches (comme par exemple des hypothèses de linéarité de la variable réponse dans un modèle linéaire).

Une autre limite provient de l'entrée retardée des sujets dans les études de cohorte.
L'approche par pseudo-valeurs en présence de troncature à gauche a été étudiée par Grand *et al.* (2018) dans le cadre d'un modèle de survie. Le calcul des pseudo-valeurs nécessite de pondérer par n l'estimateur et $n-1$ l'estimateur de type *leave-one-out*. Ils proposent de modifier la pondération du calcul de la pseudo-valeur en présence de troncature à gauche par le nombre de sujets à risque

à l'âge considéré (au lieu d'utiliser le nombre total de sujets de l'échantillon). Le sujet est encore ouvert, nous ne conseillons pas d'utiliser l'approche par pseudo-valeurs en présence de troncature à gauche et de temps de maladie d'un modèle *illness-death* censuré par intervalle tel quel.
L'approche par linéarisation estime des modèles *illness-death* ajustés sur les variables explicatives. Les mesures biologiques (comme l'IMC ou la pression artérielle) ou les tests psychométriques sont des variables qui varient beaucoup en fonction de l'âge. Nous déconseillons d'ajuster des modèles *illness-death* par ce type de variables pour ne pas faire d'hypothèses épidémiologiques trop importantes en présence de troncature à gauche.

Le choix de la méthode d'estimation est un point à discuter. Pour laisser le plus de liberté possible, les approches par linéarisation et par pseudo-valeurs ont été proposées à partir d'estimateur du maximum de vraisemblance pénalisée et à partir d'estimateur paramétrique (avec des distributions de Weibull). Outre la prise en compte de la censure par intervalle du temps de maladie et la possibilité de décéder en ayant été malade sans diagnostic, les avantages et les limites de ces deux méthodes d'estimation sont présentés dans les deux paragraphes ci-dessous.

V.2.1 Discussion des méthodes d'estimation

V.2.1.1 Estimateur du maximum de vraisemblance pénalisée

La méthode par vraisemblance pénalisée a l'avantage de ne pas faire d'hypothèse sur la distribution des temps d'évènement en approchant les intensités de transition par des M-splines. L'inconvénient est le nombre important de coefficients associés aux bases de splines à estimer. Ce nombre total de paramètre est égal à $(m_{01} + 2) + (m_{02} + 2) + (m_{12} + 2)$ avec m_{kl} le nombre de nœuds intérieurs. Un autre inconvénient à relever pour cette méthode est aussi le choix du paramètre de lissage. Dans le cadre des simulations, le choix n'a été fait que pour un seul échantillon de chaque schéma et chaque scénario en deux temps. La première étape a constitué à estimer les paramètres de lissage par validation croisée. La seconde étape s'est appuyée sur des graphiques des intensités de transition pour rechercher les paramètres de lissage en partant des valeurs initiales estimées par validation croisée (étape 1). Ce choix est assez déterminant car des paramètres de lissage trop petits ou trop grands peuvent affecter toutes les estimations qui en découlent. De plus, cette méthode estime beaucoup de paramètres, les temps de calculs sont importants, en particulier pour l'approche par pseudo-valeurs, même pour les applications (voir annexe B.2 pour les détails).

V.2.1.2 Estimateur paramétrique

Une autre possibilité pour un modèle *illness-death* avec temps de maladie censurés par intervalle est d'estimer les intensités de transition de manière paramétrique, en supposant que les intensités de transition du modèle suivent des distributions de Weibull. Le nombre de paramètres à estimer est alors de 6 pour les intensités de transition, ce qui permet une estimation plus rapide du modèle *illness-death* que par la méthode d'estimation du maximum de la vraisemblance pénalisée.

Pour conclure sur les avantages et limites propres aux méthodes d'estimation du modèle *illness-death*, il y a un choix à opérer entre un modèle très flexible qui demande des temps de calcul et des ressources-machines importants ou un modèle qui suppose une certaine forme des distributions mais qui est plus rapide à estimer.

V.2.2 Adéquation des modèles

L'adéquation de l'estimation d'un modèle *illness-death* n'est pas évidente. Les techniques classiques de l'analyse de données de survie ne sont pas appropriées car la présence de plusieurs transitions rend impossible des méthodes de résidus, l'utilisation de variables dépendantes du temps est compliquée et toutes les transitions ne sont pas observées. L'approche par linéarisation se base sur l'estimation d'un modèle *illness-death* en fonction des variables explicatives. Pour certaines applications, une hypothèse de proportionnalité des intensités a été faite. Elle peut être vérifiée graphiquement pour des variables qualitatives (en estimant des modèles stratifiés et en regardant si les différentes intensités de transition de bases respectent l'hypothèse), mais ce choix n'est pas possible pour des variables quantitatives. De plus, l'hypothèse de log-linéarité, non nécessaire, des variables quantitatives a aussi été supposée pour ne pas compliquer la modélisation.

En plus de l'adéquation du modèle *illness-death*, les adéquations propres à l'approche par pseudo-valeurs et à l'approche par linéarisation n'ont pas été proposées. L'approche par pseudo-valeurs utilise des modèles de régressions avec les pseudo-valeurs comme variable d'intérêt. La littérature pour l'approche par pseudo-valeurs montre quelques exemples d'études de l'adéquation des pseudo-valeurs. À notre connaissance, l'approche par linéarisation n'a pas été proposée ailleurs. Les débuts des travaux n'ont pas permis à l'heure actuelle de proposer une méthode d'adéquation pour cette approche.

V.2.3 Limites des études de simulations

De nombreuses simulations ont été conduites dans ce travail de thèse. En effet, mieux comprendre l'impact des déterminants de la démence sur des indicateurs épidémiologiques de la démence n'était pas possible avant. Nous avons proposé des nouveaux outils statistiques et étendu des techniques existantes pour répondre à ce besoin. Pour pouvoir évaluer le comportement d'une (nouvelle) méthode, nous avons choisi de réaliser des simulations. Un autre moyen aurait été de passer par des développements statistiques plus théoriques, cela n'a pas été le choix fait ici.

L'intérêt des simulations est de savoir *a priori* les résultats qu'une méthode doit donner (si elle fonctionne). Une étude de simulation se découpe en trois « phases » :

— la génération de données

— l'estimation de la nouvelle méthode sur les données simulées

— l'évaluation, grâce à différents critères, de la qualité des estimations.

V.2.3.1 Génération des données

Dans cette thèse, toutes les données ont été simulées en partant du principe que le jeu de données créé se calque sur les données observées issues des études de cohorte. Les données sont composées d'une partie « survie » (avec T_0 le temps de sortie de l'état 0 et T le temps de décès, accompagnée de deux indicatrices d'évènement d_1 et d_2) et d'une partie « covariables », qui représentent les variables mesurées dans les études de cohorte. De plus, la censure par intervalle n'est qu'une notion que l'on peut facilement ajouter à des données en temps continu. La génération des données s'est donc faite au niveau d'un modèle *illness-death*.

Ce point est important car il amène le point suivant : comme les données sont simulées suivant un modèle *illness-death*, on ne connait pas les valeurs théoriques que l'on doit estimer.

Ce point est assez peu discuté dans la littérature et beaucoup d'articles sur l'approche par pseudo-valeurs font des simulations uniquement dans le cadre d'un modèle de survie (où le lien entre la survie et le risque par exemple est direct).

V.2.3.2 Valeurs théoriques en simulation

L'étude du comportement de nouveaux modèles statistiques peut se faire grâce à des simulations.

Les critères mesurant les qualités d'un modèle sont multiples mais ils supposent de comparer une valeur théorique à une valeur estimée. Comme les données simulées ne se trouvent pas sur la même échelle que les indicateurs épidémiologiques, il a fallu calculer des valeurs de référence. En particulier, les méthodes choisies pour calculer les valeurs de référence dans l'approche par pseudo-valeurs et dans l'approche par linéarisation ne sont pas les mêmes.

V.2.3.3 Critères de jugement

Les choix faits dans cette thèse se sont limités aux moyennes des estimations et des biais pour mesurer les biais des méthodes. La mesure de la précision des modèles s'est faite par comparaison entre les écarts-types empiriques et les écart-types asymptotiques. La justesse [1] a été mesurée par les RMSE et les taux de couverture. Mais toutes ces mesures supposent une valeur théorique (ou attendue). Dans cette thèse, nous avons comparé des méthodes dans leur globalité : certaines mesures sont plus favorables à une méthode qu'une autre. Il faut garder en tête que chaque mesure présentée seule n'évalue pas complètement une méthode. Par exemple, un taux de couverture satisfaisant ne montre pas que la méthode peut être biaisée si les intervalles de confiance associés sont larges.

1. la *justesse* a été librement traduit du terme anglais *accuracy* pour ne pas être confondu avec le terme de *précision*.

V.3 Perspectives

Les deux approches proposées dans ce travail ont été appliquées à la cohorte PAQUID. Cependant, les résultats épidémiologiques nécessitent d'être plus approfondis en prenant en compte plus de facteurs par exemple. Pour cela, un package R pour l'approche par pseudo-valeurs est en cours de développement pour mettre en œuvre facilement cette approche. Le package `SmoothHazard` a été développé pour étudier l'effet de déterminants sur les intensités de transition et propose de calculer une multitude de quantités à partir de ces intensités de transition. Une réflexion sur la manière de calculer ces quantités pour l'approche par linéarisation doit être menée car actuellement, le package n'est pas optimisé pour les estimer. Les applications épidémiologiques seront facilitées et les approches pourront aussi être appliquées à d'autres pathologies que la démence et d'autres indicateurs épidémiologiques.

Des analyses complémentaires doivent être menées pour regarder le comportement des modèles dans des cas mal-spécifiés (par exemple en cas de censure dépendantes des variables explicatives, là où l'approche par pseudo-valeurs nécessite une correction). En effet, la génération des données pour les études de simulation s'est faite sous les hypothèses des modèles *illness-death*. Pour l'approche par linéarisation par exemple, des données à intensités proportionnelles ont été générées, sachant que les modèles *illness-death* sont estimés avec un effet proportionnel. L'approche par linéarisation suppose de choisir m et J, le nombre de sujets à tirer au sort et le nombre de vecteurs des paramètres du modèle *illness-death*. Une étude approfondie de l'impact de m et j doit aussi être réalisée en complément.

Une des perspectives découle directement des limites énoncées précédemment. Le comportement des pseudo-valeurs pour des données tronquées à gauche n'a pas été étudié en présence de censure par intervalle du temps de maladie d'un modèle *illness-death* et troncature à gauche. Une des perspectives est donc d'étendre l'approche par pseudo-valeurs grâce aux propositions de Grand *et al.* (2018) pour ce type de schéma d'observation.

Plus largement, toute extension du modèle *illness-death* pour temps de maladie censurés par intervalle permettrait d'étendre les approches par linéarisation et par pseudo-valeurs. Par exemple, l'utilisation d'un modèle avec plus de trois états peut être une piste d'extension. Des extensions semblent aussi imaginables grâce aux modèles conjoints pour prendre en compte par exemple des variables dépendantes du temps. Les applications des approches par pseudo-valeurs ou par linéarisation peuvent être encore plus vastes à partir du moment où un estimateur marginal est disponible pour l'approche par pseudo-valeurs et à partir du moment où un estimateur conditionnel aux variables explicatives est disponible pour l'approche par linéarisation.

Bibliographie

3C Study Group . Vascular Factors and Risk of Dementia : Design of the Three-City Study and Baseline Characteristics of the Study Population. *Neuroepidemiology*, 22(6) : 316–325, 2003.

Aalen O. O. A linear regression model for the analysis of life times. *Statistics in Medicine*, 8(8) : 907–925, 1989.

Alioum A. et Commenges D. A proportional hazards model for arbitrarily censored and truncated data. *Biometrics*, 52(2) : 512–524, 1996.

Alzheimer's Disease International . World Alzheimer Report 2015, The Global Impact of Dementia : An analysis of prevalence, incidence, cost and trends. Rapport, Alzheimer's Disease International, Londres, 2015.

Alzheimer's Disease International . World Alzheimer Report 2018 - The state of the art of dementia research : New frontiers. Rapport, Alzheimer's Disease International, Londres, 2018.

American Psychiatric Association . *Diagnostic and Statistical Manual of Mental Disorders, 3rd Ed. Rev. : DSM-III-R*. American Psychiatric Association, Washington DC, 1987.

American Psychiatric Association . *Diagnostic and Statistical Manual of Mental Disorders, 4th Ed. : DSM-IV*. American Psychiatric Association, Washington DC, 1994.

Andersen P. K. et Pohar Perme M. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, 19(1) : 71–99, 2010.

Andersen P. K. et Klein J. P. Regression Analysis for Multistate Models Based on a Pseudo-value Approach, with Applications to Bone Marrow Transplantation Studies. *Scandinavian Journal of Statistics*, 34(1) : 3–16, 2007.

Andersen P. K., Borgan Ø., Gill R. D. et Keiding N. *Statistical Models Based on Counting Processes*. Springer US, New York, NY, 1993.

Andersen P. K. et Pohar Perme M. Inference for outcome probabilities in multi-state models. *Lifetime Data Analysis*, 14(4) : 405–431, 2008.

Andersen P. K., Klein J. P. et Rosthøj S. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1) : 15–27, 2003.

Andersen P. K., Hansen M. G. et Klein J. P. Regression Analysis of Restricted Mean Survival Time Based on Pseudo-Observations. *Lifetime Data Analysis*, 10(4) : 335–50, 2004.

Bennett D. A., Schneider J. A., Buchman A. S., Mendes de Leon C., Bienias J. L. et Wilson R. S. The Rush Memory and Aging Project : Study Design and Baseline Characteristics of the Study Cohort. *Neuroepidemiology*, 25(4) : 163–175, 2005.

Beyersmann J. et Scheike T. H. Classical Regression Models for Competing Risks. In Klein J. P., van Houwelingen H. C., Ibrahim J. et Scheike T. H., éditeurs, *Handbook of Survival Analysis*. CRC Press, Boca Raton, Florida, 2014.

Binder N. et Schumacher M. Missing information caused by death leads to bias in relative risk estimates. *Journal of Clinical Epidemiology*, 67(10) : 1111–1120, 2014.

Binder N., Gerds T. A. et Andersen P. K. Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime Data Analysis*, 20(2) : 303–315, 2014.

Brayne C., Huppert F., Paykel E. et Gill C. The Cambridge Project for Later Life : Design and Preliminary Results. *Neuroepidemiology*, 11(1) : 71–75, 1992.

Carcaillon L., Quintin C., Moutengou E., Boussac-Zarebska M., Moisan F., Ha C. et Elbaz A. Peut-on estimer la prévalence de la maladie d'Alzheimer et autres démences à partir des bases de données médico-administratives ? Comparaison aux données de cohortes populationnelles. *Bulletin épidémiologique hebdomadaire*, 28-29 : 459–467, 2016.

Cox D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society : Series B (Methodological)*, 34(2) : 187–202, 1972.

Dartigues J.-F., Gagnon M., Barberger-Gateau P., Letenneur L., Commenges D., Sauvel C., Michel P. et Salamon R. The Paquid Epidemiological Program on Brain Ageing. *Neuroepidemiology*, 11(1) : 14–18, 1992.

Dartigues J.-F., Gagnon M., Michel P., Letenneur L., Commenges D., Barberger-Gateau P., Auriacombe S., Rigal B., Bedry R., Alpérovitch A., Orgogozo J.-M., Henry P., Loiseau P. et Salamon R. Le programme de recherche Paquid sur l'épidémiologie de la démence. Méthodes et résultats initiaux. *Revue Neurologique (Paris)*, 147 : 225–230, 1991.

Do G. et Kim Y.-J. Analysis of interval censored competing risk data with missing causes of failure using pseudo values approach. *Journal of Statistical Computation and Simulation*, 87 (4) : 631–639, 2017.

Farmer M. E., White L. R., Kittner S. J., Kaplan E., Moes E., Mcnamara P., Wolz M. M., Wolf P. A. et Feinleib M. NEUROPSYCHOLOGICAL TEST PERFORMANCE IN FRAMINGHAM : A DESCRIPTIVE STUDY. *Psychological Reports*, 60(3c) : 1023–1040, 1987.

Fine J. P. et Gray R. J. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, 94(446) : 496–509, 1999.

Finkelstein D. M. A Proportional Hazards Model for Interval-Censored Failure Time Data. *Biometrics*, 42(4) : 845–854, 1986.

Folstein M. F., Folstein S. E. et McHugh P. R. "Mini-mental state". *Journal of Psychiatric Research*, 12(3) : 189–198, 1975.

France Alzheimer & Maladie Apparentées . La maladie d'Alzheimer en chiffres. https ://www.francealzheimer.org/maladie-dalzheimer-vos-questions-nos-reponses/maladie-dalzheimer-chiffres/, 2019.

Fratiglioni L., Launer L. J., Andersen K., Breteler M. M., Copeland J. R., Dartigues J. F., Lobo A., Martinez-Lage J., Soininen H. et Hofman A. Incidence of dementia and major subtypes in Europe : A collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group. *Neurology*, 54(11 Suppl 5) : S10–15, 2000.

Grand M. K., Putter H., Allignol A. et Andersen P. K. A note on pseudo-observations and left-truncation. *Biometrical Journal*, 2018.

Grand M. K. et Putter H. Regression models for expected length of stay. *Statistics in Medicine*, 35(7) : 1178–1192, 2016.

Graw F., Gerds T. A. et Schumacher M. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis*, 15(2) : 241–255, 2009.

Grøn R. et Gerds T. A. Binomial Regression Models. In Klein J. P., van Houwelingen H. C., Ibrahim J. et Scheike T. H., éditeurs, *Handbook of Survival Analysis*. CRC Press, Boca Raton, Florida, 2014. ISBN 978-1-4665-5566-2 978-1-4665-5567-9. OCLC : 884099914.

Helmer C., Joly P., Letenneur L., Commenges D. et Dartigues J. F. Mortality with dementia : Results from a French prospective community-based cohort. *American Journal of Epidemiology*, 154(7) : 642–648, 2001.

Helmer C., Pasquier F. et Dartigues J.-F. Épidémiologie de la maladie d'Alzheimer et des syndromes apparentés. *médecine/sciences*, 22(3) : 288–296, 2006.

Hofman A., Grobbee D. E., de Jong P. T. et van den Ouweland F. A. Determinants of disease and disability in the elderly : The Rotterdam Elderly Study. *European Journal of Epidemiology*, 7 (4) : 403–422, 1991.

Hofman A., Ott A., Breteler M. M., Bots M. L., Slooter A. J., van Harskamp F., van Duijn C. N., Van Broeckhoven C. et Grobbee D. E. Atherosclerosis, apolipoprotein E, and prevalence of dementia and Alzheimer's disease in the Rotterdam Study. *The Lancet*, 349(9046) : 151–154, 1997.

Ikram M. A., Brusselle G. G. O., Murad S. D., van Duijn C. M., Franco O. H., Goedegebure A., Klaver C. C. W., Nijsten T. E. C., Peeters R. P., Stricker B. H., Tiemeier H., Uitterlinden A. G., Vernooij M. W. et Hofman A. The Rotterdam Study : 2018 update on objectives, design and main results. *European Journal of Epidemiology*, 32(9) : 807–850, 2017.

Institut national de la statistique et des études économiques (France) , de Plazaola J. et Rignols E. *Tableaux de l'économie française*. 2019.

Jacobsen M. et Martinussen T. A Note on the Large Sample Properties of Estimators Based on Generalized Linear Models for Correlated Pseudo-observations : Asymptotics of pseudo value regression. *Scandinavian Journal of Statistics*, 43(3) : 845–862, 2016.

Jellinger K. A. Clinicopathological analysis of dementia disorders in the elderly – An update. *Journal of Alzheimer's Disease*, 9(s3) : 61–70, 2006.

Joly P., Commenges D. et Letenneur L. A penalized likelihood approach for arbitrarily censored and truncated data : Application to age-specific incidence of dementia. *Biometrics*, 54(1) : 185–194, 1998.

Joly P., Commenges D., Helmer C. et Letenneur L. 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, 2002.

Jorm A. F. Is Depression a Risk Factor for Dementia or Cognitive Decline ? *Gerontology*, 46 : 9, 2000.

Kalmijn S., Launer L. J., Ott A., Witteman J. C. M., Hofman A. et Breteler M. M. B. Dietary fat intake and the risk of incident dementia in the Rotterdam study. *Annals of Neurology*, 42(5) : 776–782, 1997.

Kaplan E. L. et Meier P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282) : 457, 1958.

Katz S. Assessing Self-maintenance : Activities of Daily Living, Mobility, and Instrumental Activities of Daily Living. *Journal of the American Geriatrics Society*, 31(12) : 721–727, 1983.

Kim S. et Kim Y.-J. Regression analysis of interval censored competing risk data using a pseudovalue approach. *Communications for Statistical Applications and Methods*, 23(6) : 555–562, 2016.

Klein J. P. Modelling competing risks in cancer studies. *Statistics in Medicine*, 25(6) : 1015–1034, 2006.

Klein J. P. et Andersen P. K. Regression Modeling of Competing Risks Data Based on Pseudovalues of the Cumulative Incidence Function. *Biometrics*, 61(1) : 223–29, 2005.

Klein J. P., Logan B., Harhoff M. et Andersen P. K. Analyzing survival curves at a fixed point in time. *Statistics in Medicine*, 26(24) : 4505–4519, 2007.

Klein J. P., Gerster M., Andersen P. K., Tarima S. et Pohar Perme M. SAS and R functions to compute pseudo-values for censored data regression. *Computer Methods and Programs in Biomedicine*, 89(3) : 289–300, 2008.

Law C. G. et Brookmeyer R. Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in Medicine*, 11(12) : 1569–1578, 1992.

Lawton M. P. et Brody E. M. Assessment of Older People : Self-Maintaining and Instrumental Activities of Daily Living. *Gerontologist*, 9 : 179–186, 1969.

Leffondré K., Touraine C., Helmer C. et Joly P. 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(4) : 1177–1186, 2013.

Lemeshow S., Letenneur L., Dartigues J.-F., Lafont S., Orgogozo J.-M. et Commenges D. Illustration of Analysis Taking into Account Complex Survey Considerations : The Association between Wine Consumption and Dementia in the PAQUID Study. *American Journal of Epidemiology*, 148(3) : 298–306, 1998.

Letenneur L., Commenges D., Dartigues J. F. et Barberger-Gateau P. Incidence of Dementia and Alzheimer's Disease in Elderly Community Residents of South-Western France. *International Journal of Epidemiology*, 23(6) : 1256–1261, 1994.

Letenneur L., Gilleron V., Commenges D., Helmer C., Orgogozo J. M. et Dartigues J. F. Are sex and educational level independent predictors of dementia and Alzheimer's disease ? Incidence data from the PAQUID project. *Journal of Neurology, Neurosurgery & Psychiatry*, 66(2) : 177–183, 1999.

Levenberg K. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2(2) : 164–168, 1944.

Liang K.-Y. et Zeger S. L. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73(1) : 13, 1986.

Lim H. J. et Zhang X. Semi-parametric additive risk models : Application to injury duration study. *Accident Analysis & Prevention*, 41(2) : 211–216, 2009.

Lobo A., Launer L. J., Fratiglioni L., Andersen K., Di Carlo A., Breteler M. M., Copeland J. R., Dartigues J. F., Jagger C., Martinez-Lage J., Soininen H. et Hofman A. Prevalence of dementia and major subtypes in Europe : A collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group. *Neurology*, 54(11 Suppl 5) : S4–9, 2000.

Logan B. R., Zhang M.-J. et Klein J. P. Marginal Models for Clustered Time-to-Event Data with Competing Risks Using Pseudovalues. *Biometrics*, 67(1) : 1–7, 2011.

Mandel M. Simulation-Based Confidence Intervals for Functions With Complicated Derivatives. *The American Statistician*, 67(2) : 76–81, 2013.

Marquardt D. W. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2) : 431–441, 1963.

Merchant C., Tang M., Albert S., Manly J., Stern Y. et Mayeux R. The influence of smoking on the risk of Alzheimer's disease. *Neurology*, 52(7) : 6, 1999.

Mogensen U. B. et Gerds T. A. A random forest approach for competing risks based on pseudo-values. *Statistics in Medicine*, 32(18) : 3102–3114, 2013.

Moreno-Betancur M. et Latouche A. Regression modeling of the cumulative incidence function with missing causes of failure using pseudo-values. *Statistics in Medicine*, 32(18) : 3206–3223, 2013.

Mura T., Dartigues J.-F. et Berr C. How many dementia cases in France and Europe ? Alternative projections and scenarios 2010-2050 : Future numbers of dementia cases. *European Journal of Neurology*, 17(2) : 252–259, 2010.

Nelson W. Hazard Plotting for Incomplete Failure Data. *Journal of Quality Technology*, 1(1) : 27–52, 1969.

Nicolaie M. A., van Houwelingen J. C., de Witte T. M. et Putter H. Dynamic Pseudo-Observations : A Robust Approach to Dynamic Prediction in Competing Risks : Dynamic Pseudo-Observations : A Robust Approach to Dynamic Prediction in Competing Risks. *Biometrics*, 69(4) : 1043–1052, 2013.

Ott A., Stolk R. P., Hofman A., van Harskamp F., Grobbee D. E. et Breteler M. M. Association of diabetes mellitus and dementia : The Rotterdam Study. *Diabetologia*, 39(11) : 1392–1397, 1996.

Overgaard M., Parner E. T. et Pedersen J. Estimating the variance in a pseudo-observation scheme with competing risks : Variance in a pseudo-observation scheme. *Scandinavian Journal of Statistics*, 45(4) : 923–940, 2018.

Overgaard M., Parner E. T. et Pedersen J. Pseudo-observations under covariate-dependent censoring. *Journal of Statistical Planning and Inference*, 202 : 112–122, 2019.

Pan W. et Chappell R. Estimation in the Cox Proportional Hazards Model with Left-Truncated and Interval-Censored Data. *Biometrics*, 58(1) : 64–70, 2002.

Pavlič K. et Pohar Perme M. Using pseudo-observations for estimation in relative survival. *Biostatistics*, 20(3) : 384–399, 2019.

Pavlič K., Martinussen T. et Andersen P. K. Goodness of fit tests for estimating equations based on pseudo-observations. *Lifetime Data Analysis*, 2018.

Pérès K., Matharan F., Allard M., Amieva H., Baldi I., Barberger-Gateau P., Bergua V., Bourdel-Marchasson I., Delcourt C., Foubert-Samier A., Fourrier-Réglat A., Gaimard M., Laberon S., Maubaret C., Postal V., Chantal C., Rainfray M., Rascle N. et Dartigues J.-F. Health and aging in elderly farmers : The AMI cohort. *BMC Public Health*, 12(1), 2012.

Pohar Perme M. et Gerster M. {p}seudo : Pseudo - observations. R package version 1.1. http ://CRAN.R-project.org/package=pseudo, 2012.

Prasher V. P., Sajith S. G., Rees S. D., Patel A., Tewari S., Schupf N. et Zigman W. B. Significant effect of APOE epsilon 4 genotype on the risk of dementia in Alzheimer's disease and mortality in persons with Down Syndrome. *International Journal of Geriatric Psychiatry*, 23(11) : 1134–1140, 2008.

Ramsay J. O. Monotone Regression Splines in Action. *Statistical Science*, 3(4) : 425–441, 1988.

Ruitenberg A., van Swieten J. C., Witteman J. C., Mehta K. M., van Duijn C. M., Hofman A. et Breteler M. M. Alcohol consumption and risk of dementia : The Rotterdam Study. *The Lancet*, 359(9303) : 281–286, 2002.

Sabathé C., Andersen P. K., Helmer C., Gerds T. A., Jacqmin-Gadda H. et Joly P. Regression analysis in an illness-death model with interval-censored data : A pseudo-value approach. *Statistical Methods in Medical Research*, 2019.

Touraine C., Helmer C. et Joly P. Predictions in an illness-death model. *Statistical Methods in Medical Research*, 25(4) : 1452–1470, 2016.

Touraine C., Gerds T. A. et Joly P. **SmoothHazard** : An *R* Package for Fitting Regression Models to Interval-Censored Observations of Illness-Death Models. *Journal of Statistical Software*, 79 (7), 2017.

Tunes-da-Silva G. et Klein J. P. Regression analysis of mean quality-adjusted survival time based on pseudo-observations. *Statistics in Medicine*, 28(7) : 2009, 2009.

von Cube M., Schumacher M., Putter H., Timsit J.-F., van de Velde C. et Wolkewitz M. The population-attributable fraction for time-dependent exposures using dynamic prediction and landmarking. *Biometrical Journal*, 2019.

World Health Organization . Mental and behavioural disorders. In *The International Clasification of Diseases, 10th*. World Heatlh Organization, Genève, 1992.

World Health Organization . *Global Action Plan on the Public Health Response to Dementia 2017-2025*. World Heatlh Organization, Geneva, 2017.

World Health Organization . WHO — Life expectancy. http ://www.who.int/gho/mortality_burden_disease/life_tables/situation_trends_text/en/, 2019.

Wu Y.-T., Clare L., Hindle J. V., Nelis S. M., Martyr A., Matthews F. E. et on behalf of the Improving the experience of Dementia and Enhancing Active Life study . Dementia subtype and living well : Results from the Improving the experience of Dementia and Enhancing Active Life (IDEAL) study. *BMC Medicine*, 16(1), 2018.

Zöller D., Schmidtmann I., Weinmann A., Gerds T. A. et Binder H. Stagewise pseudo-value regression for time-varying effects on the cumulative incidence. *Statistics in Medicine*, 35(7) : 1144–1158, 2016.

Annexe A

Indicateurs épidémiologiques en fonction des variables explicatives

Les indicateurs épidémiologiques de la démence définis en section II.2 peuvent être définis en fonction des variables explicatives (voir annexe A.2 pour le détail). Pour rappel, soient :

— $\alpha_{kl}(t, Z_{kl})$ l'intensité de transition conditionnellement aux variables explicatives, de l'état k à l'état l au temps t

— $\alpha_{kl}(t \mid Z_{kl})$ l'intensité de transition conditionnellement aux variables explicatives suivant un modèle à intensités proportionnelles, de l'état k à l'état l au temps t avec $\alpha_{kl}(t \mid Z_{kl}) = \alpha_{kl,0} \exp(\beta_{kl}^{\top} Z_{kl})$

— $A_{kl}(s, t, Z_{kl})$ l'intensité de transition cumulée conditionnellement aux variables explicatives, de l'état k à l'état l entre les temps s et t

— $A_{kl}(s, t \mid Z_{kl})$ l'intensité de transition cumulée conditionnellement aux variables explicatives selon un modèle à intensités proportionnelles, de l'état k à l'état l entre les temps s et t et $A_{kl}(s, t \mid Z_{kl}) = A_{kl,0}(s, t) \exp(\beta_{kl}^{\top} Z_{kl})$

A.1 Indicateurs épidémiologiques calculés pour un horizon fini

Les trois indicateurs épidémiologiques de ce travail calculés pour un horizon fini en fonction des variables explicatives sont :

$$
\begin{aligned}
F_{01}(s,t,Z) &= \int_s^t \exp\left[-A_{01}(s,u,Z_{01}) - A_{02}(s,u,Z_{02})\right] \alpha_{01}(u,Z_{01})\mathrm{d}u \\
F_{01}(s,t \mid Z) &= \Bigg(\int_s^t \exp\left[-A_{01,0}(s,u)\exp\left(\beta_{01}^\top Z_{01}\right) - A_{02,0}(s,u)\exp\left(\beta_{02}^\top Z_{02}\right)\right] \\
&\qquad \alpha_{02,0}(u)\exp\left(\beta_{01}^\top Z_{01}\right)\mathrm{d}u\Bigg) \\
P_{00}(s,t,Z) &= \exp\left[-A_{01}(s,t,Z_{01}) - A_{02}(s,t,Z_{02})\right] \\
P_{00}(s,t \mid Z) &= \exp\left[-A_{01,0}(s,t)\exp(\beta_{01}^\top Z_{01}) - A_{02,0}(s,t)\exp\left(\beta_{02}^\top Z_{02}\right)\right] \\
RM(s,t,Z) &= \int_s^t \exp\left[-A_{01}(s,u,Z_{01}) - A_{02}(s,u,Z_{02})\right]\mathrm{d}u \\
RM(s,t \mid Z) &= \int_s^t \exp\left[-A_{01,0}(s,u)\exp\left(\beta_{01}^\top Z_{01}\right) - A_{02,0}(s,u)\exp\left(\beta_{02}^\top Z_{02}\right)\right]\mathrm{d}u
\end{aligned}
$$

A.2 Indicateurs épidémiologiques calculés pour un horizon infini

Le risque vie-entière de démence, l'espérance de vie sans démence et l'espérance de vie totale suivant les variables explicatives se définissent par :

$$
\begin{aligned}
LTR(s,Z) &= \int_s^{+\infty} \exp\left[-A_{01}(s,u,Z_{01}) - A_{02}(s,u,Z_{02})\right] \alpha_{01}(u,Z_{01})\mathrm{d}u \\
LTR(s \mid Z) &= \Bigg(\int_s^{+\infty} \exp\left[-A_{01,0}(s,u)\exp\left(\beta_{01}^\top Z_{01}\right) - A_{02,0}(s,u)\exp\left(\beta_{02}^\top Z_{02}\right)\right] \\
&\qquad \alpha_{02,0}(u)\exp\left(\beta_{01}^\top Z_{01}\right)\mathrm{d}u\Bigg) \\
LE_{00}(s,Z) &= \int_s^{+\infty} \exp\left[-A_{01}(s,u,Z_{01}) - A_{02}(s,u,Z_{02})\right]\mathrm{d}u \\
LE_{00}(s \mid Z) &= \int_s^{+\infty} \exp\left[-A_{01,0}(s,u)\exp\left(\beta_{01}^\top Z_{01}\right) - A_{02,0}(s,u)\exp\left(\beta_{02}^\top Z_{02}\right)\right]\mathrm{d}u \\
LE_{0.}(s \mid Z) &= \int_s^{+\infty} \Bigg(\exp\left[-A_{01}(s,u \mid Z_{01}) - A_{02}(s,u \mid Z_{02})\right] \\
&+ \int_s^v \exp\left[-A_{01}(s,v \mid Z_{01}) - A_{02}(s,v \mid Z_{02})\right] \alpha_{01}(v \mid Z_{01}) \exp\left[-A_{12}(v,u \mid Z_{12})\right]\mathrm{d}v\Bigg)\mathrm{d}u
\end{aligned}
$$

Annexe B

Implémentation

B.1 Amélioration du package `SmoothHazard`

Toutes les estimations effectuées dans le cadre de cette thèse ont été réalisées grâce au package `SmoothHazard`. Des implémentions supplémentaires ont été réalisé au cours de cette thèse. En particulier, la moyenne restreinte des temps de survie en bonne santé ($RM(.)$) a été implémentée pour être calculée en même temps que le risque absolu de la maladie ($F_{01}(.)$) et la probabilité d'être dans l'état 0 ($P_{00}(t.)$). De plus, le risque vie-entière de la maladie ($LTR(.)$) a été ajouté au package et est calculé en même temps que les espérances de vie disponibles. La fonction de risque absolu de la maladie a été recodée pour être calculée en concordance avec les équations présentées dans ce chapitre (c.à.d. sans faire appel à l'intensité de transition $\alpha_{12}(.)$). Ces différentes implémentations sont disponibles sur la plateforme GitHub dans le dépôt disponible à l'adresse suivante : `https://github.com/tagteam/SmoothHazard`.

B.2 Temps de calcul

Les temps de calcul des deux approches proposées sont à prendre en considération lors de l'application des approches sur des données réelles. Notez que les programmes ont été parallélisés pour les deux approches :

— pour l'approche par pseudo-valeurs, les modèles *illness-death* définis sur des échantillons par méthode *leave-one-out* sont indépendants les uns des autres et peuvent donc être estimés parallèlement (tout comme les indicateurs épidémiologiques)

— pour l'approche par linéarisation, les différents indicateurs estimés à partir du retirage d'une loi normale multivariée (c.à.d. le calcul de $\hat{\theta}_i^{[j]}(s,t,Z_i)$) sont indépendants les uns des autres et peuvent être estimés en parallèle.

Pour l'application des pseudo-valeurs sur un échantillon de la cohorte PAQUID, les temps de calcul sont les suivants :

Tableau B.1 – Paramètres des lois de Weibull utilisés pour générer les données des différents schémas de simulations.

Schéma	Transition	Forme	Échelle	β_1	β_2
A	0-1	1,74	0,05	0,33	NA
	0-2	1,22	0,06	-0,58	NA
	1-2	1,11	0,25	-0,31	NA
B	0-1	2,14	0,04	0,51	NA
	0-2	1,36	0,30	0,50	NA
	1-2	1,63	0,08	0,30	NA
C	entrée dans la cohorte	11,13	0,01	NA	NA
	0-1	10,50	0,01	0,19	0,0800
	0-2	8,90	0,01	-0,66	0,0005
	1-2	7,57	0,01	-0,40	0,0200

NA : non applicable

— il a fallu 45 secondes pour estimer le modèle *illness-death* avec tous les sujets, en ayant choisi au préalable les paramètres de lissage κ de la vraisemblance pénalisée

— environ 2 heures 30 minutes pour calculer les modèles *illness-death* par une technique de *leave-one-out* en parallélisant sur 20 cœurs sur un serveur dédiés exclusivement aux calculs scientifiques

— moins d'une minute pour calculer la pseudo-valeur de chaque sujet pour un indicateur épidémiologique en parallélisant sur 20 cœurs sur un serveur dédiés exclusivement aux calculs scientifiques

Pour estimer l'effet de l'âge à la ménopause, ajusté sur le fait d'avoir eu des enfants par l'approche par linéarisation, les temps de calculs ont été les suivants :

— 4 minutes au total pour estimer pour un tirage aléatoire de 100 femmes parmi les 1909 femmes composant l'échantillon

— 2 heures 30 minutes pour estimer l'effet en utilisant les 1909 femmes de l'échantillon ($m = n$)

avec un tirage aléatoire de 1000 valeurs du vecteur des paramètres du modèle *illness-death* pour ces deux points en parallélisant sur 50 cœurs d'un serveur de calculs.

B.3 Paramètres des distributions de Weibull

Pour ne pas alourdir les notations des modèles *illness-death* générés dans les différents scénarios de simulation, les paramètres de forme et d'échelle de chaque transition n'ont pas été présentés dans le corps de la thèse. Ils sont disponibles dans le tableau B.1.

Annexe C

Descriptif complémentaire des échantillons

Les histogrammes C.1 montrent la répartition de l'âge à l'entrée (en haut) et du score au MMS à l'inclusion (en bas) des sujets de la cohorte PAQUID utilisés dans l'application du chapitre III et de la seconde illustration du chapitre IV.

Les histogrammes C.2 montrent la répartition de l'âge à l'entrée (en haut) et de l'âge à la ménopause (en bas) des femmes de la cohorte PAQUID utilisées dans la troisième illustration du chapitre IV.

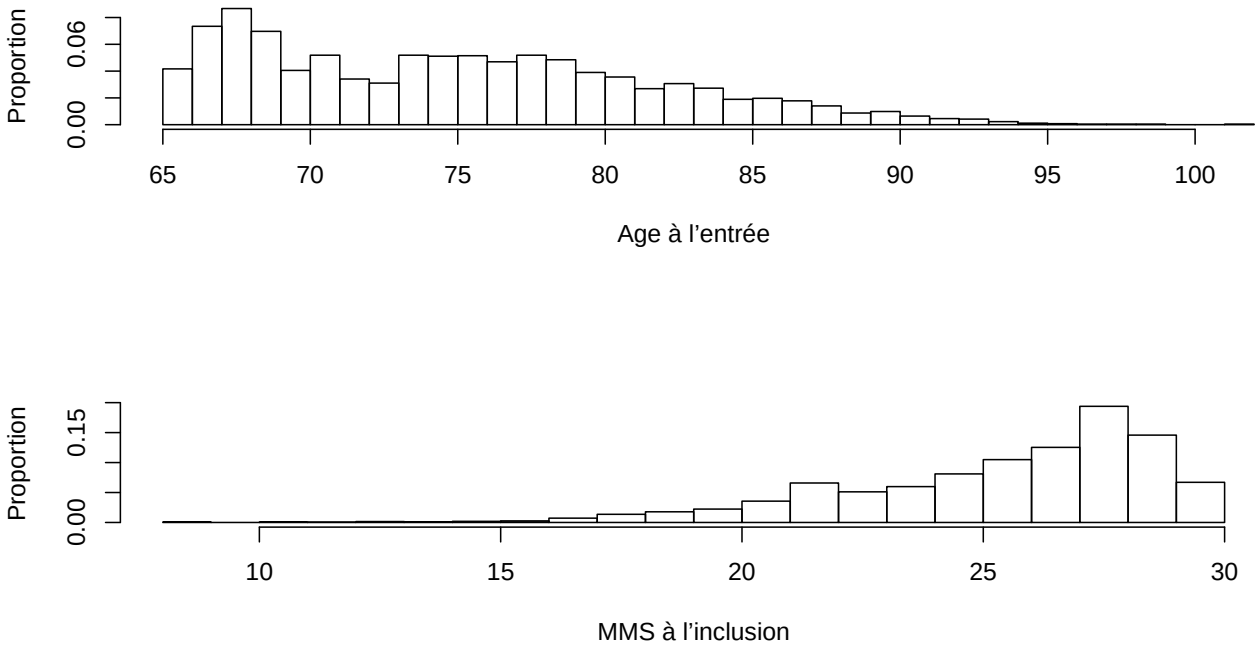

FIGURE C.1 – Répartition de l'âge à l'inclusion et du MMS à l'inclusion pour l'application du chapitre III et la seconde illustration du chapitre IV. $n = 3673$.

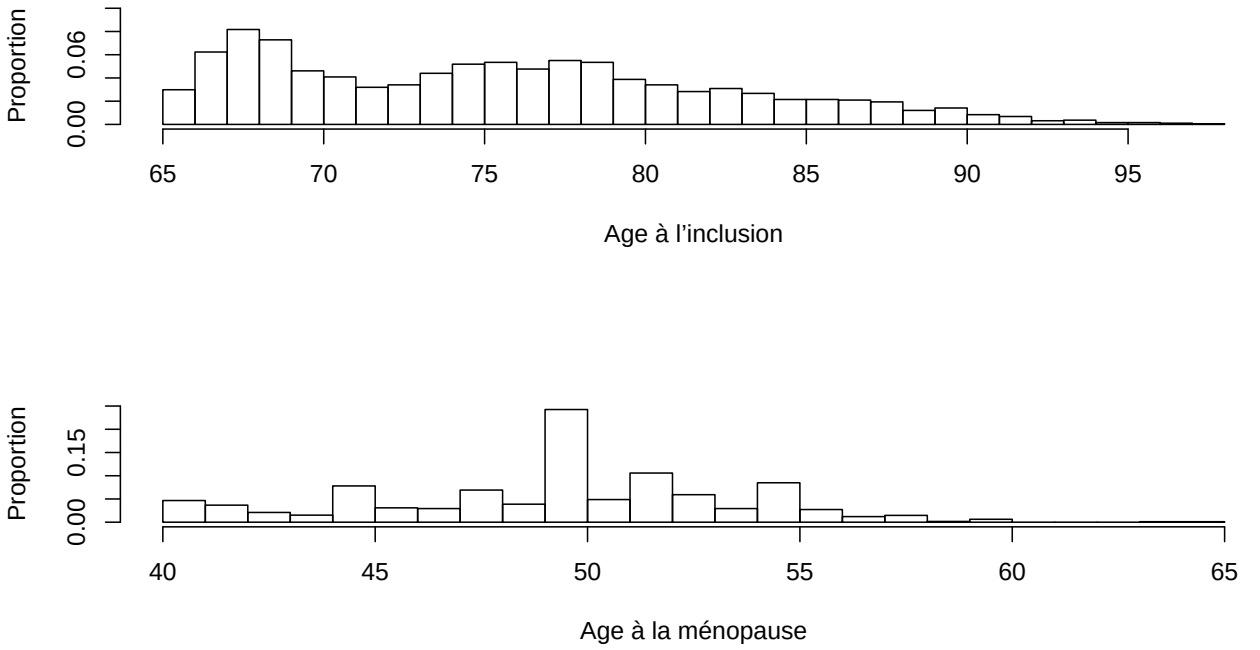

FIGURE C.2 – Répartition de l'âge à l'inclusion (en haut) et de l'âge à la ménopause (en bas) des femmes de la troisième illustration du chapitre IV. $n = 1906$.

Annexe D

Résultats complémentaires

Le tableau D.1 retranscrit les résultats du Schéma A présenté dans le chapitre III pour un lien logarithmique. Les résultats présentés sont ceux des paramètres de régression et s'interprète grâce à la fonction exponentielle. Par exemple, pour le paramètre $\gamma_{1,10,F}$ l'interprétation est la suivante : un sujet qui a pour caractéristique $Z = 1$ a un risque absolu de démence multiplié par 1,49 à 10 ans après l'inclusion, comparé à un sujet ayant $Z = 0$ comme caractéristique ($\exp(0,403) \approx 1,496$).

Les figures D.1 et D.2 retranscrivent les résultats de l'approche par pseudo-valeurs pour le schéma B. La figure D.1 illustre les estimations pour le risque absolu de démence ($F_{01}(5)$) en haut, pour la probabilité d'être vivant non-dément ($P_{00}(5)$) au milieu et pour la moyenne restreinte des temps de survie sans démence ($RM(5)$) en bas pour un horizon de 5 ans après l'inclusion. Les colonnes de gauche à droite représente la présence de censure par intervalle : à gauche, le temps de maladie est observé en temps continu (c.à.d. sans censure par intervalle), au milieu le temps de maladie est censuré par intervalle avec un espace entre deux visites consécutives d'au moins 2 ans et d'au plus 3 ans et la colonne de droite correspond aux données censurées par intervalle avec un temps compris entre 3 et 6 ans entre deux visites consécutives. La figure D.2 retranscrit les mêmes informations que la figure D.2 pour un horizon de 15 ans après l'inclusion. Pour ces deux figures, les valeurs théoriques de l'effet de Z sur l'indicateur épidémiologique sont présentées par la courbe noire et les estimations sur les 50 échantillons sont présentés par les courbes grises.

Tableau D.1 – Résultats des simulations : comparaison de l'approche par pseudo-valeurs et un lien logarithmique pour le risque absolu de démence ($F_{01}(10)$), la probabilité d'être vivant non-dément ($P_{00}(10)$) et la moyenne restreinte des temps de survie sans démence ($RM(10)$) 10 ans après l'inclusion suivant deux méthodes d'estimation et trois scénarios. Schéma A : 500 échantillons de 500 sujets.

Indicateur	γ	Scénario et méthode	$\overline{\hat{\gamma}}$	Biais	ASE	ESE	RMSE × 1000	Taux de couv.
$F_{01}(10)$	-1,645	1- np	-1,665	-0,021	0,144	0,151	152	93,8
		1- splines	-1,663	-0,019	0,136	0,147	148	94,6
		1- Weibull	-1,658	-0,013	0,126	0,136	137	94,2
		2- splines	-1,658	-0,013	0,178	0,175	175	95,4
		2- Weibull	-1,669	-0,024	0,171	0,171	173	95,8
		3- splines	-1,652	-0,007	0,225	0,244	244	93,4
		3- Weibull	-1,674	-0,029	0,228	0,263	265	93,4
	0,403	1- np	0,426	0,023	0,171	0,188	190	92,0
		1- splines	0,425	0,022	0,161	0,179	180	92,2
		1- Weibull	0,415	0,012	0,148	0,168	169	92,2
		2- splines	0,427	0,024	0,210	0,212	213	94,6
		2- Weibull	0,421	0,018	0,202	0,206	207	94,8
		3- splines	0,404	0,001	0,264	0,286	286	93,4
		3- Weibull	0,412	0,009	0,266	0,277	277	94,0
$P_{00}(10)$	-0,836	1- np	-0,840	-0,004	0,079	0,079	79	95,2
		1- splines	-0,839	-0,004	0,074	0,075	75	96,0
		1- Weibull	-0,840	-0,005	0,064	0,065	65	96,4
		2- splines	-0,841	-0,005	0,074	0,074	74	95,8
		2- Weibull	-0,840	-0,004	0,065	0,066	67	95,8
		3- splines	-0,839	-0,003	0,073	0,074	74	94,8
		3- Weibull	-0,840	-0,004	0,067	0,069	69	94,8
	0,119	1- np	0,116	-0,004	0,100	0,110	110	93,6
		1- splines	0,115	-0,004	0,093	0,101	102	93,8
		1- Weibull	0,118	-0,001	0,081	0,087	87	94,2
		2- splines	0,116	-0,003	0,093	0,101	101	94,0
		2- Weibull	0,118	-0,001	0,083	0,090	90	93,6
		3- splines	0,116	-0,004	0,092	0,099	99	94,8
		3- Weibull	0,119	-0,000	0,085	0,092	92	94,4
$RM(10)$	1,982	1- np	1,980	-0,001	0,030	0,031	31	95,0
		1- splines	1,980	-0,001	0,036	0,040	40	92,2
		1- Weibull	1,980	-0,001	0,029	0,029	29	95,6
		2- splines	1,981	-0,001	0,030	0,031	31	94,6
		2- Weibull	1,981	-0,001	0,029	0,029	29	95,4
		3- splines	1,980	-0,001	0,031	0,032	32	94,4
		3- Weibull	1,980	-0,001	0,030	0,034	34	94,8
	0,056	1- np	0,056	-0,000	0,038	0,040	40	93,6
		1- splines	0,052	-0,004	0,046	0,051	51	94,4
		1- Weibull	0,056	-0,000	0,036	0,038	38	94,2
		2- splines	0,055	-0,001	0,038	0,040	40	93,8
		2- Weibull	0,055	-0,001	0,037	0,039	39	95,0
		3- splines	0,055	-0,001	0,039	0,042	42	95,0
		3- Weibull	0,055	-0,001	0,038	0,040	40	94,8

scénario 1 sans censure par intervalle / scénarios 2 et 3 avec censure par intervalle

ASE = écart-type asymptotique / ESE = écart-type empirique

Tableau D.2 – Modélisation du risque de démence vie-entière ($LTR(70, agem, mere)$) pour des femmes vivantes et non-démentes de 70 ans, en fonction de l'âge à la ménopause ajusté sur le fait d'avoir eu un enfant. Estimation par linéarisation ($m = 100$ et $J = 1000$) d'après 1909 sujets de la cohorte PAQUID.

	$\hat{\Gamma}$	$\text{BC}_{95\%}(\hat{\Gamma})$
Risque de démence vie-entière à 70 ans		
Femme de 50 ans à la ménopause et sans enfant	0,537	[0,474 ;0,595]
Age à la ménopause (année)	0,027	[-0,032 ; 0,091]
Avoir des enfants (oui versus non)	-0,001	[-0,008 ; 0,008]

BC : bande de confiance

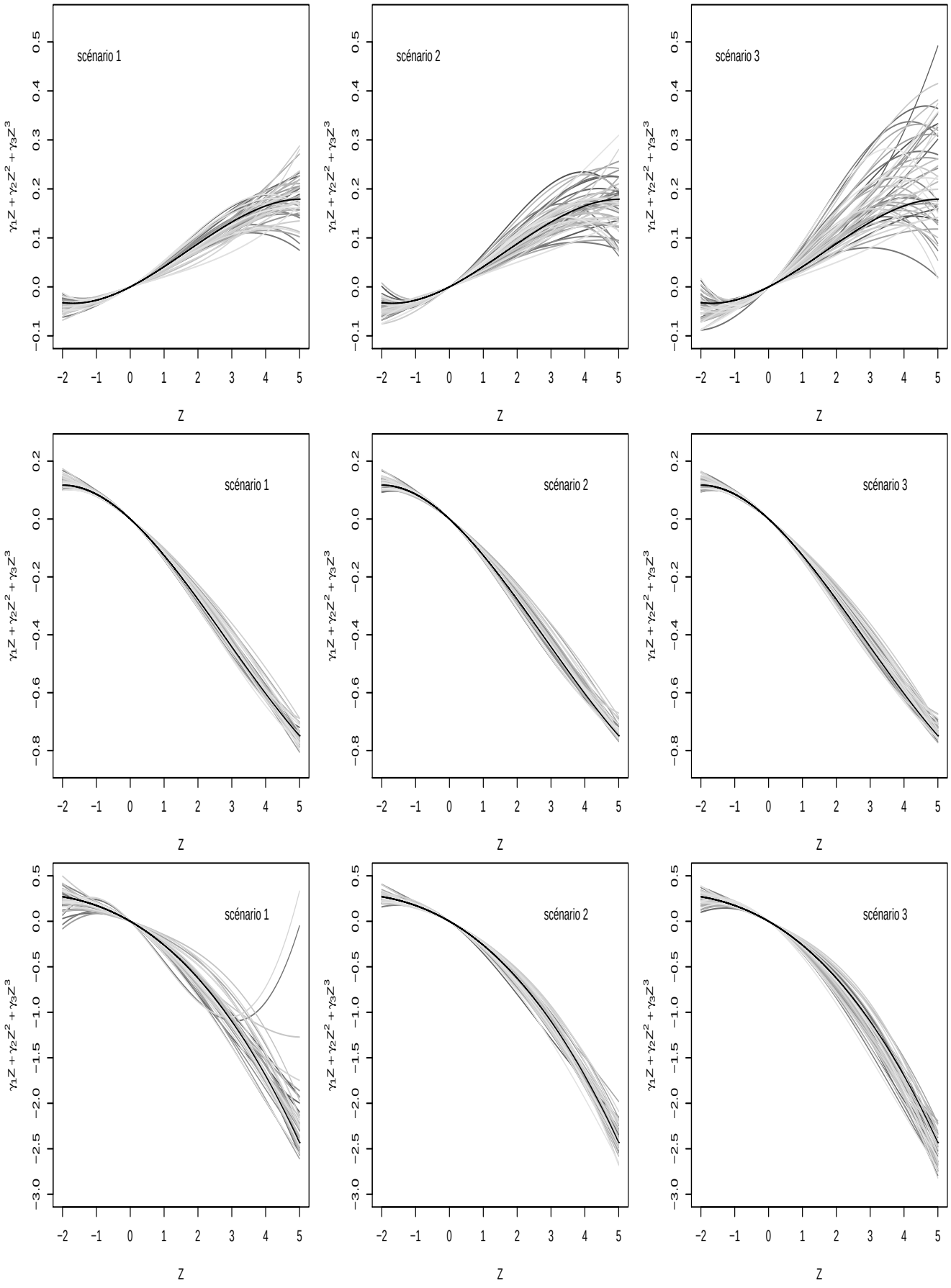

FIGURE D.1 – Résultats des simulations : comparaison de l'approche par pseudo-valeurs trois indicateurs épidémiologiques de la démence (de haut en bas) calculés à 5 ans de suivi ($F_{01}(5), P_{00}(5)$ et $RM(5)$) suivant trois scénarios (de gauche à droite). Schéma B : 50 répliques de 3200 sujets.

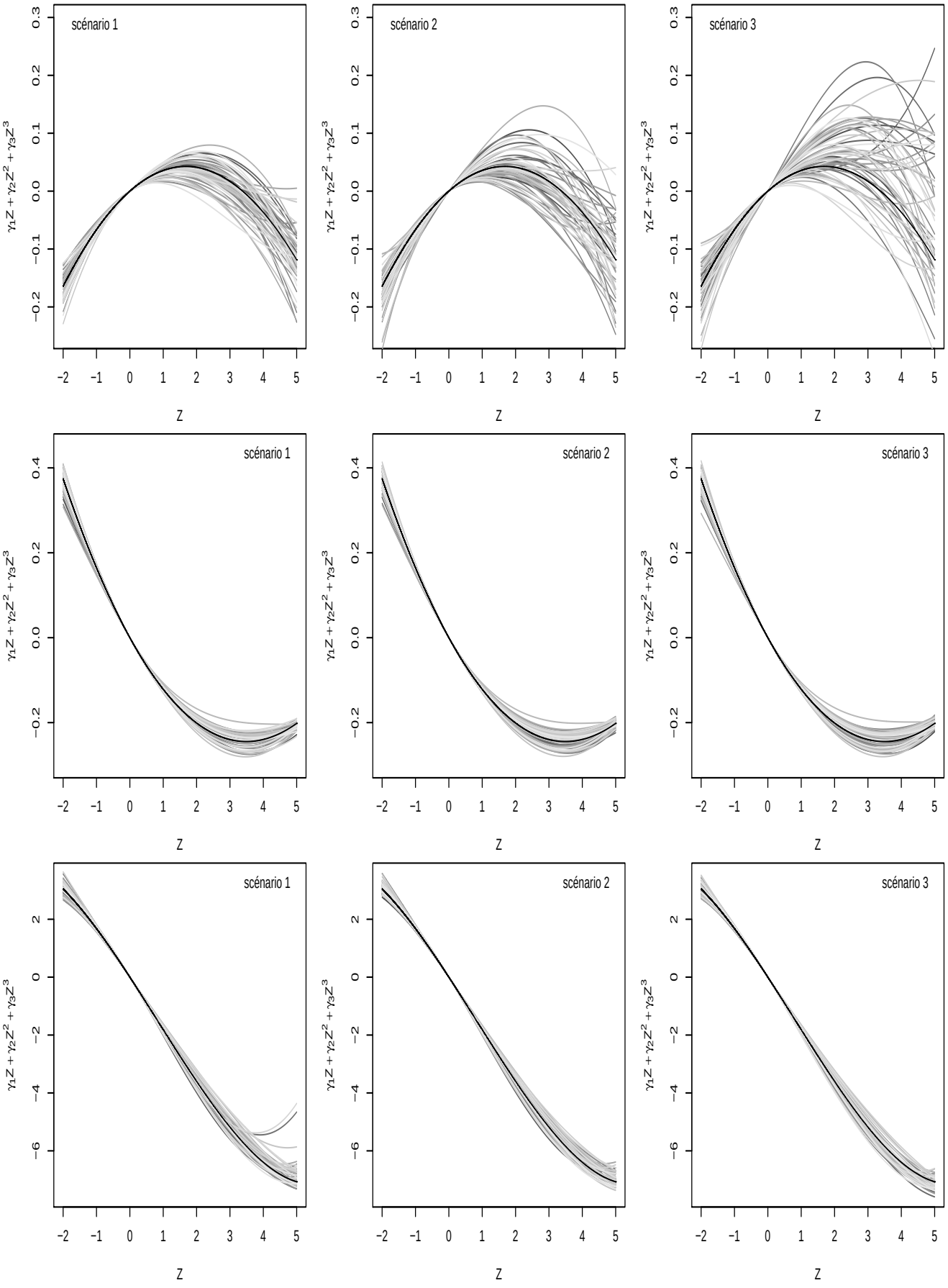

FIGURE D.2 – Résultats des simulations : comparaison de l'approche par pseudo-valeurs trois indicateurs épidémiologiques de la démence (de haut en bas) calculés à 5 ans de suivi ($F_{01}(15), P_{00}(15)$ et $RM(15)$) suivant trois scénarios (de gauche à droite). Schéma B : 50 répliques de 3200 sujets.