UNIVERSITÉ DE BORDEAUX

UNIVERSITÉ BORDEAUX SEGALEN

Année 2013 Thèse n°2099

THÈSE

pour le

DOCTORAT DE L'UNIVERSITÉ BORDEAUX 2

École doctorale Sociétés, Politique, Santé Publique

Mention : Sociétés, Politique, Santé Publique

Spécialité : Santé Publique

Option : Biostatistique

Présentée et soutenue publiquement

Le 10 décembre 2013 par

Célia Touraine

Née le 31 mars 1985 à Oloron Sainte-Marie

Modèles illness-death pour données censurées par intervalle : Application à l'étude de la démence

JURY

Mme. Hélène JACQMIN-GADDA	Directrice de Recherche, Bordeaux	Présidente
M. Laurent BORDES	Professeur, Pau	Rapporteur
M. Aurélien LATOUCHE	Professeur, Paris	Rapporteur
M. Jean-François DARTIGUES	Professeur, Bordeaux	Examinateur
M. Yohann FOUCHER	Maître de conférences, Nantes	Examinateur
M. Philippe SAINT PIERRE	Maître de conférences, Paris	Examinateur
M. Pierre JOLY	Maître de conférences, Bordeaux	Directeur

Cette thèse a été préparée au sein de l'équipe Biostatistique du centre de recherche INSERM U897 à l'ISPED. Elle a été rendue possible grâce au financement de la Région Aquitaine.

À ma mère

Remerciements

Je remercie en premier lieu mon directeur de thèse, Pierre Joly, qui m'a accordé sa confiance en m'offrant la possibilité d'effectuer cette thèse. Ces trois années avec lui ont été très enrichissantes d'un point de vue scientifique. J'ai eu la chance de travailler avec quelqu'un de non seulement très compétent mais aussi de profondément humain. Je le remercie particulièrement pour s'être rendu disponible en toutes circonstances et pour s'être montré compréhensif dans les moments difficiles.

Merci à Hélène Jacqmin-Gadda de me faire l'honneur de présider mon jury de thèse. Ses compétences, son expérience et son dynamisme m'ont énormément apporté durant ces trois années. Ses nombreuses qualités sont précieuses pour cette équipe au sein de laquelle j'ai eu tant de plaisir à évoluer.

Je remercie vivement M. Laurent Bordes et M. Aurélien Latouche de me faire le plaisir d'être les rapporteurs de ma thèse. Je suis très honorée que vous ayez accepté de juger mon travail. Je remercie M. Yohann Foucher d'avoir accepté d'être membre de mon jury de thèse. Je remercie également M. Philippe Saint-Pierre de participer à l'évaluation de cette thèse. Son manuscrit de thèse m'a lui-même été d'une grande aide lors de mon initiation aux modèles multi-états. Je remercie enfin M. Jean-François Dartigues de bien vouloir examiner mon travail de thèse. De ses connaissances en épidémiologie et de sa compréhension des biostatistiques émergent toujours des commentaires pertinents et un éclairage nouveau sur les travaux de l'équipe.

Mes remerciements vont aussi à toute l'équipe Biostat et aux autres ispédiens. À Karen avec qui j'ai eu beaucoup de plaisir à travailler ; merci pour ton accessibilité, tes conseils. À Amadou pour son investissement, sa gentillesse, la bonne humeur qu'il insufflait dans le bureau. À Émilie, la super détectrice des bugs de SmoothHazard et ma regrettée partenaire de badminton. À Robin, merci de nous conter tes péripéties rigolotes et de penser à nous, les doctorants. Merci aussi aux membres extérieurs à l'équipe : Fleur Delva et Fanny Artaud qui ont utilisé SmoothHazard pour leur travaux d'épidémiologie ; les échanges que j'ai eu avec vous ont été très constructifs et enrichissants pour moi. Merci aussi aux Paquidettes Fanny et Mélanie pour leur aide. Merci à Mathilde qui à sa manière m'a bien aidé pour la fin de thèse ; garde le même caractère et tu vas faire du super boulot ! Merci bien sûr à tous les autres qui m'ont fait passer d'agréables moments à et en dehors de l'ISPED : Julie, Boris, Mélanie, Hind, Yassin, Ling-ling, Riccardo, Lucie, Viviane, Henri, Jérémie, Alexandre. Un grand merci à Audrey, continue à mettre autant de vie autour de toi ; à Jérémie, merci pour les discussions partagées, ta gentillesse ; à Mbéry, merci pour ta générosité, pour les discussions avec toi qui permettent de relativiser ou de s'évader un peu ; à Paul, tu m'as écouté, soutenu, conseillé et aidé pour beaucoup de choses, que ce soit dans le travail ou pour finir mes bières, merci !

Un immense merci aussi à ceux qui m'ont soutenu dans les moments difficiles ou permis de me sortir un peu le nez de l'ISPED pendant ces trois années. Aux bordelais : Fabien, Gabriel ; aux toulousains : Fabien, Clément ; aux oloronaises : Émilie, je n'ai pas été très présente mais tu as toujours été compréhensive, tu as pensé à moi, merci pour tous tes messages de soutien ; Ludivine, heureusement que je t'ai, tu as toujours été là pour moi quoi qu'il arrive et je sais que je pourrai toujours compter sur toi. À Arnaud ; tu es la personne qui a été la plus directement impliquée dans cette thèse, tu as supporté mes humeurs, m'a conseillé, aidé et rassuré à tous points de vue, merci pour tout.

Mes derniers remerciements, et non les moindres, vont à ma famille. À Mitou, à mon frère. À mes parents ; si j'en suis arrivée là, c'est grâce à vous. Maman, il n'y avait pas besoin de mots pour que tu saches quand ça n'allait pas, ta présence cette année m'a énormément manqué mais j'ai trouvé les ressources pour y arriver et je continuerai car même absente tu seras toujours près de moi. Papa, tu es mon pilier, ne l'oublie pas.

Liste des abréviations et notations

Abréviations

resp.	respectivement
vs	*versus*
cep	certificat d'études primaires
HR	rapport des risques instantanés (*hazard ratio*)
IC	intervalle de confiance
3C	étude des 3 Cités
LR	statistique de test de rapport de vraisemblance
VIH	virus de l'immunodéficience humaine
SIDA	syndrome d'immunodéficience acquise

Notations

Analyse de survie

n	effectif
$\wedge$	minimum
$\lambda(.)$	fonction de risque
$\Lambda(.)$	fonction de risque cumulée

L	vraisemblance
L_{Cox}	vraisemblance partielle de Cox
T_i	durée de survie du sujet i
C_i	temps de censure à droite du sujet i
$\tilde{T}_i$	durée observée du sujet i ($C_i \wedge T_i$)
$R(t)$	Effectif à risque à l'instant t^-
Z_i	vecteur des variables explicatives du sujet i

Modèles multi-états

MV	maximum de vraisemblance
MVP	maximum de vraisemblance pénalisée
X	processus (de Markov si pas de précision supplémentaire)
$\alpha_{kl}(.)$	intensités de transition d'un modèle multi-état ou *illness-death*
$\alpha_l(.)$	intensité de transition d'un modèle à risques concurrents associée à la cause l de décès
$A_{kl}(.)$	intensités de transition cumulées d'un modèle multi-état ou *illness-death*
$A_l(.)$	intensité de transition cumulée d'un modèle à risque concurrents associée à la cause l de décès
$p_{kl}(.,.)$	probabilités de transition de l'état k à l'état l
$p^0_{02}(.,.)$, $p^1_{02}(.,.)$	probabilités de transiter de l'état 0 à l'état 2 directement/en passant par l'état 1
$F_j(t)$	dans un modèle à risques concurrents, incidence cumulée associée au décès par la cause j (*i.e.* probabilité cumulée de décéder de la cause j)
$F_{01}(s,t)$, $F_{02}(s,t)$	dans un modèle *illness-death* et pour un sujet en l'état 0 au temps s, probabilités d'atteindre respectivement l'état 1 et l'état 2 avant t
$F_{0\bullet}(s,t)$	dans un modèle *illness-death* et pour un sujet en l'état 0 au temps s, probabilité de sortie de l'état 0 avant t
i.i.d	indépendants et identiquement distribués
I	matrice identité

Table des matières

Valorisation scientifique

Publications

— Touraine C, Helmer C, Joly P (2013), Predictions in an illness-death model. *Statistical Methods in Medical Research.*

— Touraine C, Gerds T A, Joly, P. The SmoothHazard package for R : Fitting regression models to interval-censored observations of illness-death models. *Research report 13/12. Department of Biostatistics, University of Copenhagen*, 2013.

— Leffondré K, Touraine C, Helmer C, Joly P (2013), Interval-censored time-to-event and competing risk with death : is the illness-death model more accurate than the Cox model ? *International Journal of Epidemiology* **42**(4).

— Joly P, Touraine C, Georget A, Dartigues J-F, Commenges D and Jacqmin-Gadda H (2013), Prevalence Projections of Chronic Diseases and Impact of Public Health Intervention. *Biometrics* **69**(1), 109-117.

— Dumurgier J, Artaud F, Touraine C, Rouaud O, Tavernier B, Singh-Manoux A, Tzourio C, Elbaz A. Slow walking speed associated with incident dementia : 10-year follow-up of the 3C Dijon study. Soumis, 2013.

— Jacqmin-Gadda H, Blanche P, Chary E, Touraine C, Dartigues J-F. ROC curve estimation for time-to-event with semi-competing risks and interval censoring. *Statistical Methods in Medical Research.* En révision, 2013.

Communications orales

— Touraine C, Joly P. A comparison between the estimates of regression parameters using an illness-death model and a two-state survival model. 3^{rd} *International Biometric Society Channel Network Conference*, avril 2011, Bordeaux (France).

— Touraine C, Joly P. Prédictions dans un modèle illness-death. $44^{èmes}$ *Journées de Statistique*, mai 2012, Bruxelles (Belgique).

— Touraine C, Joly P. Predictions from a Markov illness-death model : application to dementia disease. 27^{th} *International Workshop on Statistical Modelling*, juillet 2012, Prague (République Tchèque).

— Joly P, Touraine C, Diakité A, Gerds T A. Analyse de données de survie en présence de censure par intervalle : le package SmoothHazard. $1^{ères}$ *rencontres R*, juillet 2012, Bordeaux (France).

— Touraine C, Joly P. Modèle illness-death pour données censurées par intervalle avec effets aléatoires : Application à la cohorte Paquid. $45^{èmes}$ *Journées de Statistique*, mai 2013, Toulouse (France).

— Joly P, Blanche P, Touraine C. Modèle de régression pour des probabilités cumulées en présence de risques concurrents et de censure par intervalles. $45^{èmes}$ *Journées de Statistique*, mai 2013, Toulouse (France).

— Jacqmin-Gadda H, Touraine C, Dufouil C, Elbaz A, Dartigues J-F, Joly P. Prévalence de la démence en France de 2010 à 2030 et impact des politiques de prévention. *ADELF*, octobre 2013, Bordeaux (France).

Communications affichées

— Joly P, Touraine C. Predictions and life expectancies in illness-death models. 33^{rd} *Annual Conference of the International Society of Clinical Biostatistics*, août 2012, Bergen (Norvège).

— Touraine C, Joly P. Illness-death model for interval-censored and left-truncated data with random effects : Application to dementia. 27^{th} *International Workshop on Statistical Modelling*, juillet 2013, Palerme (Italie).

— Leffondré K, Touraine C, Helmer C, Joly P. Comment étudier les déterminants de la démence avec des méthodes d'analyse de la survie ? *Congrès ADELF*, octobre 2013, Bordeaux (France).

Introduction générale

Le nombre de personnes atteintes de démence ne cesse d'augmenter avec le vieillissement de la population faisant de l'étude de la démence un enjeu majeur de santé publique. Des études de cohorte ont été mises en place et des méthodes statistiques plus ou moins sophistiquées sont nécessaires pour en exploiter les données. Le présent travail a pour objet de proposer des méthodes visant à une meilleure connaissance de l'épidémiologie de la démence du sujet âgé.

Contexte

La démence

Le vieillissement cognitif est caractérisé par une détérioration avec l'âge des capacités mentales comme la mémoire, le langage, l'attention, la capacité d'abstraction ou d'orientation dans le temps et l'espace. Il peut être normal avec des troubles restant légers ou pathologique avec des troubles de plus en plus marqués au cours du temps. Selon la quatrième révision du manuel diagnostique et statistique des troubles mentaux (DSM-IV), la démence est caractérisée par l'apparition de déficits cognitifs multiples : l'altération de la mémoire et au moins une des perturbations cognitives suivantes :

— aphasie (perturbation du langage) ;
— apraxie (diminution de la capacité à réaliser une activité motrice malgré des fonctions motrices intactes) ;
— agnosie (incapacité de reconnaître ou d'identifier des objets malgré des fonctions sensorielles intactes) ;

— perturbation des fonctions exécutives (planifier ou faire des projets, organiser, séquencer ou ordonner dans le temps, avoir une pensée abstraite).

Ces déficits doivent de plus être à l'origine d'une altération significative du fonctionnement social ou professionnel et représenter un déclin significatif par rapport au niveau de fonctionnement antérieur.

Plusieurs types de démence sont définis en fonction du processus causal. On distingue les démences non-dégénératives (démences vasculaires par exemple) et les démences dégénératives. Parmi celles-ci, on trouve la démence à corps de Lewy, la démence associée à la maladie de Parkinson, la démence fronto-temporale et la maladie d'Alzheimer qui représente plus de deux tiers des cas de démences. Lorsqu'une démence de type Alzheimer est associée à une démence vasculaire, on parle de démence mixte.

La démence d'un sujet a un retentissement sur son autonomie et impose une prise en charge lourde et de surcroît, coûteuse. Il a été montré que les démences étaient de loin la cause principale de dépendance du sujet âgé (Dartigues et al., 2012). L'épidémiologie joue un rôle clé dans la mise en place des politiques de santé publique visant à améliorer la prise en charge de ces sujets dépendants.

Épidémiologie de la démence

Prévalence

Sur le plan de la santé publique, la prévalence de la démence (nombre de déments dans la population) est la quantité la plus importante à estimer. Une méta-analyse réunissant les données de 11 études européennes réalisées avant l'année 2000 indique une prévalence de 6.4 % pour les démences toutes causes confondues et de 4.4 % pour la maladie d'Alzheimer (Lobo et al., 2000). Les taux de prévalence des études européennes les plus récentes varient entre 5.9 % et 9.4 % (Berr et al., 2009). La prévalence de la démence dépend à la fois de son incidence (nombre de nouveaux cas dans une période de temps donnée) et de la durée de la survie des patients. Joly et al. (2013) ont proposé une méthode pour estimer la prévalence future de la démence en utilisant des prévisions du taux d'incidence de la démence et des taux de mortalité des déments et des non déments. Ils ont en particulier estimé l'impact que pourrait avoir une campagne de prévention visant à réduire la fréquence dans la population d'un facteur de risque de démence sur l'estimation de la prévalence future. Les données de Paquid, l'une des deux principales cohortes pour l'étude de la démence en France (voir le chapitre IV pour une description) ont été utilisées pour faire des prévisions de prévalence entre 2010 et 2030.

Facteurs de risque

De multiples facteurs de risque de démence ont été identifiés grâce à la cohorte Paquid. Ceux faisant l'objet d'un consensus dans la littérature scientifique sont :

- l'âge ;
- le sexe (les femmes sont plus à risque que les hommes) ;
- l'apoE4 (l'allèle E4 du gène de l'apolipoprotéine E) ;
- le niveau d'études (les personnes ayant un niveau d'études bas sont plus à risque) ;
- la consommation de tabac (augmente le risque) et la consommation modérée de vin (diminue le risque).

La symptomatologie dépressive et les activités sociales et de loisir sont également associées à respectivement une augmentation et une diminution du risque de démence, mais un rapport de causalité entre ces facteurs et l'apparition de la maladie n'a pas été établi.

D'autres facteurs plus controversés ont été mis en évidence par l'étude Paquid :

- le statut marital (le fait d'être marié diminuerait le risque de démence) ;
- la prise d'anti-inflammatoires ou de benzodiazépines.

Les facteurs suivants ont été dégagés mais des analyses supplémentaires sont nécessaires pour confirmer leur effet sur le risque de démence :

- la présence d'aluminium dans l'eau (augmenterait le risque) ;
- la présence de silice dans l'eau (diminuerait le risque) ;
- la réactivation de l'infection herpétique.

Enfin, des facteurs protecteurs liés notamment à une alimentation riche en poissons et en flavonoïdes sont suspectés et sont plus amplement explorés à l'aide de 3 Cités, la seconde des principales cohortes en France dédiées à l'étude de la démence (voir le chapitre IV pour une présentation de cette étude de cohorte).

Les données de cohorte

Les cohortes constituent l'une des principales sources de données pour l'épidémiologie de la démence en particulier, et la recherche épidémiologique en général. C'est en observant la survenue d'une maladie dans le temps qu'on peut par exemple mettre en évidence des facteurs de risque associés à cette survenue, ou estimer l'incidence et la prévalence de la maladie et décrire leur évolution.

Lorsque l'évènement observé est la survenue d'une démence, les données de cohorte

comportent certaines caractéristiques. Tout d'abord, la date de survenue de la démence ne peut être connue exactement. Les sujets de la cohorte subissent des visites au cours du suivi et leur statut « dément » ou « non dément » est connu seulement à ces dates de visites. Lorsqu'un sujet est diagnostiqué dément, on sait donc seulement que la démence est survenue entre la visite de diagnostic et la visite précédente. Cela donne lieu à des données dites *censurées par intervalle*. En outre, les sujets participant à ces cohortes sont relativement âgés avec une risque important de décès au cours du suivi. Ils peuvent décéder après la survenue d'une démence ou avant. À cause de la concomitance de la censure par intervalle et du risque de décès, le traitement de certains sujets dans un analyse statistique devient plus complexe. La difficulté vient en particulier des sujets décédés qui ont été vus non déments à leur dernière visite, car ceux-ci ont pu devenir déments avant de décéder.

Modélisation

Modèle de survie

Les modèles de survie permettent de modéliser la durée jusqu'à l'apparition d'un évènement. La spécificité des méthodes d'estimation en analyse de survie est de tenir compte des données censurées à droite, c'est-à-dire des sujets n'ayant pas subi l'évènement durant leur temps de suivi. Les fonctions associées à la distribution de la durée de survie peuvent être estimées non paramétriquement (estimateur de Kaplan-Meier de la fonction de survie, estimateur de Nelson-Aalen de la fonction de risque cumulé). Les modèles de survie permettent aussi d'associer des facteurs d'exposition au risque de survenue de l'évènement grâce à un modèle de régression. Le modèle à risques proportionnels de Cox est le plus utilisé. L'estimation des effets des facteurs est alors semi-paramétrique et consiste à maximiser une vraisemblance dite *partielle*. Cependant, dans ces techniques d'analyse de survie, les temps d'évènements doivent être connus de façon exacte. Des modèles de survie pour données censurées par intervalle, notamment des modèles à risques proportionnels ont alors été développés (Betensky et al., 2002; Finkelstein and Wolfe, 1985; Pan and Chappell, 2002). Une façon de traiter les données censurées par intervalle, plus triviale mais relativement courante, consiste à imputer les temps d'évènements (par exemple au milieu de l'intervalle de censure) pour se ramener à un modèle de survie standard. Cependant, dans un contexte où l'évènement d'intérêt est la démence, la difficulté ne provient pas de la censure par

intervalle à elle toute seule mais plutôt de sa concomitance avec les risques de décès des non déments et des déments. Comment traiter les sujets décédés qui ont été vus non déments à leur dernière visite ? Puisqu'on ne sait pas s'ils sont devenus déments ou non entre leur dernière visite et leur décès, nous n'avons d'autre choix que de les censurer à droite à leur date de dernière visite. Nous verrons, notamment dans le chapitre II, que cette censure à droite artificielle n'est pas sans conséquences et qu'un modèle *illness-death* qui modélise conjointement les deux évènements « démence » et « décès » serait plus approprié.

Modèle *illness-death*

Les modèles multi-états sont de plus en plus utilisés en analyse de survie. Ils permettent de modéliser les différents états occupés par les sujets au cours du temps. Les transitions d'un état à l'autre peuvent être observées en temps continu ou l'état d'un sujet ne peut être connu qu'en des temps discrets. Un modèle *illness-death* est un modèle multi-état particulier. Il est composé de trois états dits « sain », « malade » et « décédé ». Dans un modèle *illness-death* irréversible (*i.e.* sans rétablissement possible), un sujet peut transiter de l'état sain à l'état décédé directement ou *via* l'état malade. Dans le contexte présent, un modèle *illness-death* composé des états « non dément », « dément » et « décédé » paraît le plus approprié. Malgré la censure par intervalle, ce modèle multi-état reste simple, ce qui permet d'utiliser des méthodes d'estimation relativement flexibles (Frydman, 1995; Joly et al., 2002). En effet, dans certains modèles multi-états plus complexes, le nombre de trajectoires qui ont pu être prises par un sujet entre deux temps d'observation peut être grand, voire infini. Ici, seulement deux trajectoires sont possibles pour un sujet qui décède sans avoir été vu dément : le passage direct de l'état non dément à l'état décédé ou le passage par l'état intermédiaire dément. Les intensités de transition sont les pendants de la fonction de risque d'un modèle de survie. Et de manière similaire à un modèle de survie, on peut, grâce à trois modèles de régression, associer des facteurs d'exposition à chacune des trois intensités de transition.

Objectifs de la thèse

Les objectifs de cette thèse sont multiples :

— justifier une approche *illness-death* tenant compte des données censurées par

intervalle plutôt qu'une approche d'analyse de survie standard pour l'étude du risque de démence et des facteurs associés (chapitre II) ;
— lorsqu'on considère un modèle *illness-death*, passer en revue les quantités pertinentes d'un point de vue épidémiologique et proposer une méthode d'estimation (chapitre III) ;
— rendre accessible l'utilisation d'un modèle *illness-death* pour données censurées par intervalle (chapitre V) ;
— prendre en compte des effets aléatoires dans le modèle afin de tenir compte de certaines particularités des données de cohorte dont nous disposons (chapitre IV).

Bien que nous nous placions dans le cadre de l'étude de la démence, le présent travail ne se réduit pas à cette application. Il concerne de façon plus générale n'importe quelle application s'inscrivant dans une modélisation de type *illness-death* et pour laquelle les dates de transition vers l'état intermédiaire sont censurées par intervalle.

Structure du mémoire

Le premier chapitre de ce mémoire est une présentation des principaux modèles pour l'analyse de survie et multi-état. Il se termine par le cas particulier du modèle *illness-death* pour données censurées par intervalle. Les remarques n'y sont pas anodines ; elles mettent l'accent sur les insuffisances de certains modèles dans le contexte qui nous intéresse ou soulèvent des points importants. Le chapitre II est consacré à l'estimation des effets des facteurs influant sur le risque de démence. Des simulations mettent en évidence les biais qui peuvent survenir dans l'estimation de ces effets lorsqu'on utilise un modèle de survie qui ne prend pas en compte la censure par intervalle plutôt qu'un modèle *illness-death* pour données censurées par intervalle. Les mécanismes sous-jacents à l'apparition de ces biais sont expliqués et illustrés. Le chapitre III est consacré aux prévisions qui peuvent être faites dans un modèle *illness-death*. Des quantités pertinentes d'un point de vue épidémiologique sont développées et expliquées. Des illustrations donnent un aperçu de l'éventail des prévisions possibles. Le chapitre IV est dédié à l'ajout d'effets aléatoires dans le modèle *illness-death* pour données censurées par intervalle dans l'objectif de prendre en compte des facteurs d'exposition partagés par les sujets d'un même groupe. Le chapitre V décrit la structure d'un paquet R en général et du paquet **SmoothHazard** en particulier, dont le développement

a constitué une partie importante de cette thèse. Son utilisation est brièvement décrite dans ce chapitre et, de façon plus détaillée dans un article fourni dans l'annexe A. Les différentes annexes fournissent également divers compléments qui seront mentionnés dans les différents chapitres.

Chapitre I

État de l'art des principaux modèles de survie et multi-états

1 Analyse de survie

L'analyse de survie est l'étude du délai de survenue d'un évènement d'intérêt. Cet évènement est souvent associé à un changement d'état, communément le passage de l'état « vivant » à l'état « décédé ». Cependant, on s'intéresse souvent à d'autres types de délais que la durée de vie proprement dite : la durée jusqu'à l'apparition d'une maladie, le délai entre la prise d'un traitement et la guérison d'une maladie, la durée de séropositivité sans symptômes de patients infectés par le VIH, ou encore en fiabilité, la durée de fonctionnement d'une machine.

1.1 Notions générales

Soit T la variable aléatoire positive et continue qui représente la durée de survie ou délai, c'est-à-dire la durée écoulée jusqu'à la survenue de l'évènement d'intérêt. Pour définir cette durée, il faut définir une date d'origine qui est généralement propre aux sujets et dont le choix va dépendre de l'évènement d'intérêt. Dans un contexte d'essais cliniques par exemple, si l'on souhaite comparer deux traitements, on choisira la date de mise sous traitement comme date d'origine. Lorsque l'évènement étudié est très dépendant de l'âge, on choisit souvent la date de naissance comme date d'origine et la variable T est alors un âge.

1.1.1 Censure et troncature

La difficulté en analyse de survie réside dans le fait que les données recueillies sont en partie incomplètes.

Censure à droite

Le phénomène le plus souvent à l'origine de ces données incomplètes est la censure à droite. La durée de survie du sujet i, T_i, est dite *censurée à droite* si le sujet i n'a pas subi l'évènement à sa date de dernières nouvelles C_i, c'est-à-dire que la seule information dont on dispose est que $T_i > C_i$.

Généralement, dans des données de cohorte, une durée T_i est censurée à droite si le sujet i est :

— perdu de vue : sa surveillance est interrompue alors qu'il n'a pas encore subi l'évènement (pour cause de déménagement par exemple) ;

— exclu vivant : à la date de fin d'étude le sujet n'a pas encore subi l'évènement ;

De façon formelle, on associe à chaque sujet i la variable aléatoire $\tilde{T}_i$:

$$\tilde{T}_i = T_i \wedge C_i$$

qui est la durée réellement observée et un indicateur $\delta_i = \mathbb{1}_{\{T_i \leq C_i\}}$ tel que :

$$\delta_i = \begin{cases} 1 & \text{si la « vraie » durée est observée (dans ce cas } \tilde{T}_i = T_i) \\ 0 & \text{si la durée est censurée à droite (dans ce cas } \tilde{T}_i = C_i) \end{cases}$$

Dans les modèles classiques d'analyse de survie, on fait l'hypothèse que les variables T_i et C_i sont indépendantes, c'est-à-dire que la censure est indépendante de l'évènement. Lorsque cette hypothèse n'est pas vérifiée, par exemple quand la censure est due à un arrêt du traitement ou lorsque les sujets les plus malades et donc les plus à risque de décéder sont perdus de vue, on parle de *censure informative*.

Censure par intervalle

La durée de survie T_i est dite *censurée par intervalle* si au lieu de l'observer de façon exacte, la seule information dont on dispose est qu'elle est comprise entre deux dates connues. La censure par intervalle se rencontre généralement dans les études de cohorte lorsque les sujets ne sont pas observés en temps continu mais par intermittence lors de

visites. Par exemple, si l'on s'intéresse à l'âge de survenue d'une maladie et que le sujet i est diagnostiqué malade au cours d'une visite, on sait seulement que $T_i \in [L_i, R_i]$ où R_i est l'âge à la visite de diagnostic et L_i est l'âge à la visite précédente.

La durée T_i est dite *doublement censurée par intervalle* lorsqu'elle représente un délai entre deux variables aléatoires censurées par intervalle. On trouve souvent dans la littérature l'exemple où T_i représente le délai entre l'infection par le VIH et le début du SIDA.

Troncature à gauche

La durée de survie T est dite *tronquée* si son observation est conditionnelle à un autre évènement. La durée de survie T est *tronquée à gauche* si elle n'est observable qu'à la condition $T > A$, où A est une variable que l'on suppose indépendante de T. S'il y a troncature à gauche, on n'étudie que le sous-échantillon des sujets dont la durée de survie est supérieure à une certaine valeur. Par exemple, lorsqu'on étudie l'âge de décès et que les sujets ne sont pas suivis depuis leur date de naissance, les données sont tronquées à gauche puisque seuls les sujets vivants à la date d'inclusion sont observables et A représente leur âge à l'inclusion.

1.1.2 Distribution de la durée de survie

On suppose que la durée de survie T est une variable positive ou nulle, et absolument continue. La distribution de T est caractérisée par l'une des cinq fonctions suivantes définies pour $t \geq 0$, chacune pouvant être obtenue à partir de l'une des autres.

La *fonction de survie* S au temps t est la probabilité de survie jusqu'au temps t :

$$S(t) = \mathbb{P}(T > t)$$

La *fonction de répartition* F au temps t est la probabilité de subir l'évènement avant le temps t :

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

La *densité de probabilité* f au temps t représente la probabilité instantanée de subir l'évènement dans un petit intervalle de temps après t :

$$f(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leq t + \Delta t)}{\Delta t} = F'(t) = -S'(t)$$

Elle est telle que

$$F(t) = \int_0^t f(u)\mathrm{d}u$$

La *fonction de risque* λ au temps t représente la probabilité de subir l'évènement dans un petit intervalle de temps après t, conditionnellement au fait d'avoir survécu jusqu'à t :

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leq t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$$

Enfin, la *fonction de risque cumulé* est :

$$\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u = -\ln(S(t))$$

La fonction de survie peut aussi s'exprimer en fonction du risque cumulé :

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

1.1.3 Estimation non paramétrique

Dans le cas où l'on ne fait pas d'hypothèse *a priori* sur la distribution de la durée de survie T, les principaux estimateurs non paramétriques sont l'estimateur de Kaplan-Meier de la fonction de survie et l'estimateur de Nelson-Aalen du risque cumulé.

Estimateur de Kaplan-Meier de la fonction de survie

L'estimateur de Kaplan-Meier découle de l'idée suivante : ne pas avoir subi l'évènement à l'instant t, c'est ne pas l'avoir subi juste avant t et ne pas le subir en t. Notons t' le temps « juste avant t » et t'' le temps « juste avant t' ». On a :

$$\begin{aligned} S(t) = \mathbb{P}(T > t) &= \mathbb{P}(T > t', T > t) \\ &= \mathbb{P}(T > t | T > t')\mathbb{P}(T > t') \\ &= \mathbb{P}(T > t | T > t')\mathbb{P}(T > t' | T > t'')\mathbb{P}(T > t'') \\ &= \dots \end{aligned}$$

En considérant les durées observées (temps d'évènement ou temps de censure) rangées par ordre croissant, $\tilde{T}_i, i = 1, \dots, n$, et en les supposant distinctes, et avec $\tilde{T}_0 = 0$, on

a :

$$\mathbb{P}(T > \tilde{T}_i) = \prod_{j=1}^{i} p_j, \qquad j = 1, \ldots, n$$

où $p_j = \mathbb{P}(T > \tilde{T}_j | T > \tilde{T}_{j-1})$ est la probabilité de ne pas subir l'évènement dans l'intervalle $[\tilde{T}_{j-1}, \tilde{T}_j]$ sachant qu'il ne s'est toujours pas produit en $\tilde{T}_{j-1}$.

Considérons $R(\tilde{T}_j)$ l'*effectif à risque* à l'instant $\tilde{T}_j^-$, c'est-à-dire le nombre de sujets n'ayant pas encore subi ni l'évènement ni la censure à droite juste avant $\tilde{T}_j$. Rappelons que $\delta_j = 1$ si le sujet j a subi l'évènement $(\tilde{T}_j = T_j)$; $\delta_j = 0$ si sa durée de survie est censurée à droite $(\tilde{T}_j = C_j)$. Un estimateur naturel pour p_j est :

$$\hat{p}_j = \frac{R(\tilde{T}_j) - \delta_j}{R(\tilde{T}_j)} = 1 - \frac{\delta_j}{R(\tilde{T}_j)}$$

et on obtient l'estimateur de Kaplan-Meier :

$$\hat{S}_{KM}(t) = \prod_{j:\tilde{T}_j \leq t} \left(1 - \frac{\delta_j}{R(\tilde{T}_j)}\right)$$

Dans le cas où il y a des *ex-aequo*, c'est-à-dire que tout les $\tilde{T}_j$ ne sont pas distincts, on note $D(\tilde{T}_j)$ le nombre de sujets subissant l'évènement au temps $\tilde{T}_j$ et l'estimateur de Kaplan-Meier devient :

$$\hat{S}_{KM}(t) = \prod_{j:\tilde{T}_j \leq t} \left(1 - \frac{D(\tilde{T}_j)}{R(\tilde{T}_j)}\right)$$

Estimateur de Nelson-Aalen du risque cumulé

Nous avons abordé jusqu'ici analyse de survie en considérant la variable aléatoire de durée T mais elle peut aussi être abordée en terme de processus en considérant le processus ponctuel qui vaut 0 tant que l'évènement n'a pas lieu, 1 après. L'estimateur de Nelson-Aalen a été introduit par Aalen (1978) pour généraliser celui de Nelson (1972) aux processus de comptage. Il est donné par :

$$\hat{\Lambda}(t) = \sum_{j:T_j \leq t} \frac{D(T_j)}{R(T_j)}$$

L'estimateur du risque cumulé a une interprétation moins immédiate que celui de la fonction de survie. Son intérêt réside surtout dans la pente de la courbe correspondante qui estime la fonction de risque λ.

Remarque I.1 *Les estimateurs de Kaplan-Meier et de Nelson-Aalen se généralisent aux données tronquées à gauche mais pas aux données censurées par intervalle car dans ce dernier cas, les temps exacts d'évènements ne sont pas connus.*

1.1.4 Modèles paramétriques

Supposons maintenant que la distribution des durées de survie appartient à une famille de loi paramétrique donnée. Bien que chacune des cinq fonctions S, F, f, λ, Λ caractérise la loi de T, on spécifie souvent la forme de la fonction de risque λ qui donne la description la plus intéressante, à savoir celle du futur immédiat du sujet qui n'a pas encore subi l'évènement.

Loi exponentielle

La loi exponentielle, qui ne dépend que d'un paramètre θ, est la seule distribution continue qui admet un risque instantané constant. Pour $t \geq 0$ et avec $\theta > 0$:

$$\begin{aligned} \lambda(t) &= \theta \\ S(t) &= \mathrm{e}^{-\theta t} \\ f(t) &= \theta \mathrm{e}^{-\theta t} \end{aligned}$$

Cette loi est dite « sans mémoire » et est peu adaptée dans le domaine du vivant. Cependant, dans certaines applications, on peut découper le temps en plusieurs intervalles et considérer un risque constant sur chacun des intervalles (et différent d'un intervalle à l'autre) de façon à obtenir une fonction de risque constante par morceaux.

Loi de Weibull

La loi de Weibull, qui dépend d'un paramètre de forme a et d'un paramètre d'échelle b, admet un risque instantané monotone. C'est une généralisation de la loi exponentielle que l'on retrouve en prenant $a = 1$. Pour $t \geq 0$ et avec $a > 0$ et $b > 0$:

$$\begin{aligned} \lambda(t) &= a \left(\frac{1}{b}\right)^a t^{a-1} \\ S(t) &= \exp\left\{-\left(\frac{t}{b}\right)^a\right\} \\ f(t) &= a \left(\frac{1}{b}\right)^a t^{a-1} \exp\left\{-\left(\frac{t}{b}\right)^a\right\} \end{aligned}$$

Il existe d'autres lois avec des risques instantanés monotones. Citons notamment la loi Gamma et la loi de Gompertz.

D'autres lois permettent aussi de modéliser des risques instantanés en forme de cloche ($\cup$ ou $\cap$), en particulier la loi de Weibull généralisée.

Vraisemblance

Soit θ le vecteur des paramètres du modèle. Les estimateurs des paramètres sont obtenus en maximisant la vraisemblance détaillée ci-après. En pratique, on utilise des méthodes itératives de type algorithme de Newton-Raphson.

La vraisemblance représente la probabilité d'observer l'échantillon d'après le modèle et est le produit des n contributions individuelles :

$$L = \prod_{i=1}^{n} L_i$$

Soit t_i le temps de participation du sujet i. Dans le cas le plus fréquent de données censurées à droite, la contribution du sujet i à la vraisemblance est :

— $f(t_i;\theta)$ si $\delta_i = 1$ (le sujet i a subi l'évènement)
— $S(t_i;\theta)$ si $\delta_i = 0$ (l'observation du sujet i est censurée à droite)

et la vraisemblance s'écrit :

$$\begin{aligned} L &= \prod_{i=1}^{n} f(t_i;\theta)^{\delta_i} S(t_i;\theta)^{1-\delta_i} \\ &= \prod_{i=1}^{n} S(t_i;\theta)\lambda(t_i;\theta)^{\delta_i} \end{aligned}$$

Dans le cas de données tronquées à gauche, la contribution individuelle d'un sujet i dont l'observation est tronquée à gauche en a_i est :

$$L_i = \begin{cases} \dfrac{f(t_i;\theta)}{S(a_i;\theta)} & \text{si } \delta_i = 1 \\ \dfrac{S(t_i;\theta)}{S(a_i;\theta)} & \text{si } \delta_i = 0 \end{cases}$$

Dans le cas de données censurées par intervalle, la contribution individuelle d'un sujet

i dont l'observation est censurée dans intervalle $[l_i, r_i]$ est :

$$L_i = \begin{cases} S(r_i;\theta) - S(l_i;\theta) & \text{sans troncature à gauche} \\ \dfrac{S(r_i;\theta) - S(l_i;\theta)}{S(a_i;\theta)} & \text{avec troncature à gauche} \end{cases}$$

Remarque I.2 ***Avantages et inconvénients des approches paramétrique et non paramétrique*** *L'approche paramétrique induit des hypothèses sur la distribution des données mais elle a l'avantage de fournir des estimations en temps continu de n'importe quelle fonction caractérisant la distribution. L'approche non paramétrique fournit des estimations en temps discret et donc les fonctions estimées sont continues par morceaux. La fonction de risque étant la plus intéressante en terme d'interprétation, un lissage a posteriori de l'estimateur de Nelson-Aalen est envisageable en utilisant une méthode à noyau (Ramlau-Hansen, 1983). Une autre approche pour estimer des fonctions lisses sans faire d'hypothèse paramétrique est d'utiliser des fonctions splines (Rosenberg, 1995). Ce type d'approche et l'approche paramétrique, étant basés sur la vraisemblance, ont l'avantage de prendre en compte aisément des données censurées par intervalle.*

1.2 Prise en compte des facteurs de risque

Nous nous sommes concentrés jusqu'ici sur le délai jusqu'à la survenue de l'évènement d'intérêt mais l'un des principaux objectifs de l'analyse de survie et de l'épidémiologie est d'évaluer l'impact sur ce délai de facteurs auxquels sont exposés les sujets.

1.2.1 Modèles de régression

Les deux principaux types de modèles qui permettent d'exprimer le risque d'évènement en fonction de variables explicatives sont : les modèles à temps de vie accélérée et les modèles à risques proportionnels. Dans un modèle à durée de vie accélérée, une variable a pour effet de multiplier la durée de survie par une constante tandis que dans un modèle à risques proportionnels, une variable a pour effet de multiplier le risque instantané par une constante. Le modèle à risques proportionnels s'inscrit plus généralement dans la famille des modèles multiplicatifs. Les modèles additifs constituent une alternative aux modèles multiplicatifs : au lieu de faire le produit entre la fonction

de risque de base et une fonction des variables explicatives, on en fait la somme. Dans ce cadre, la prise en compte de variables explicatives dont l'effet varie au cours du temps a été considéré par Aalen (1980, 1989). D'autres modèles de survie dynamiques permettent de prendre en compte des effets variables dans le temps dans un cadre multiplicatif (Gamerman, 1991).

Modèle de Cox

Le modèle de régression le plus largement utilisé est le modèle à risques proportionnels de Cox (1972) :

$$\lambda(t|Z_i) = \lambda_0(t)\mathrm{e}^{\beta^T Z_i} \tag{I.2}$$

où i est l'indice du sujet,

Z_i est le vecteur des variables explicatives,

β est le vecteur des coefficients de régression,

λ_0 est la fonction de risque de base, c'est-à-dire le risque instantané des sujets pour lesquels toutes les variables explicatives sont nulles.

Le risque de survenue d'évènement à l'instant t pour un sujet qui a pour caractéristiques Z_i par rapport à un sujet qui a pour caractéristiques Z_j est :

$$\frac{\lambda(t|Z_i)}{\lambda(t|Z_j)} = \frac{\lambda_0(t)\mathrm{e}^{\beta^T Z_i}}{\lambda_0(t)\mathrm{e}^{\beta^T Z_j}} = \frac{\mathrm{e}^{\beta^T Z_i}}{\mathrm{e}^{\beta^T Z_j}} = \mathrm{e}^{\beta^T (Z_i - Z_j)}$$

Dans le modèle de Cox, le rapport des risques instantanés (*hazard ratio*) est constant au cours du temps. C'est dans ce sens que le modèle de Cox est dit à risques proportionnels. La proportionnalité des risques est une conséquence du modèle mais aussi une hypothèse à vérifier *a posteriori*.

Si la variable Z_i est une variable quantitative ou qualitative ordonnée à plus de deux modalités, il conviendra aussi de vérifier une hypothèse de log-linéarité. En effet, quand on ajoute une unité à la valeur de Z_i, on multiplie le risque instantané par e^{β} (c'est-à-dire qu'on ajoute β à son logarithme) quelle que soit la valeur de Z_i.

Vraisemblance partielle

Afin d'estimer les paramètres de régression β, l'idée de Cox a été de considérer le risque de base λ_0 comme un paramètre de nuisance en maximisant une vraisemblance dite *partielle*.

Soit D le nombre de sujets ayant subi l'évènement. Soient $T_i, i = 1, \ldots, D$ les différents temps d'évènement supposés distincts et rangés par ordre croissant avec i les indices des sujets correspondant. Notons $R(T_i)$ l'ensemble des sujets encore à risque à l'instant T_i^-.

La probabilité que le sujet i subisse l'évènement en T_i sachant qu'il y a eu un évènement en T_i s'écrit :

$$p_i = \frac{\lambda_0(T_i)\mathrm{e}^{\beta^T Z_i}}{\sum_{j \in R(T_i)} \lambda_0(T_i)\mathrm{e}^{\beta^T Z_j}} = \frac{\mathrm{e}^{\beta^T Z_i}}{\sum_{j \in R(T_i)} \mathrm{e}^{\beta^T Z_j}}$$

Le produit sur les temps d'évènement des termes p_i qui ne dépendent que du paramètre β définit alors la vraisemblance partielle de Cox :

$$L_{Cox} = \prod_{i=1}^{D} \frac{\mathrm{e}^{\beta^T Z_i}}{\sum_{j \in R(T_i)} \mathrm{e}^{\beta^T Z_j}}$$

L'estimateur de β est le vecteur des paramètres qui maximisent cette vraisemblance.

Remarque I.3
1. *Naturellement, d'autres méthodes n'utilisant pas la vraisemblance partielle de Cox peuvent être utilisées pour estimer les coefficients β dans l'équation I.2, par exemple en spécifiant un modèle paramétrique pour la fonction de risque de base λ_0.*
2. *Par extension, lorsqu'on parle du modèle de Cox, on désigne souvent non seulement l'équation I.2 mais aussi la méthode d'estimation des effets des facteurs de risque par vraisemblance partielle.*

Remarque I.4 *De la même façon que les estimateurs non paramétriques, la vraisemblance partielle de Cox se généralise aux données tronquées à gauche mais pas aux données censurées par intervalle.*

1.2.2 Modèles à fragilité

Modèles simples à fragilité

Toutes les variables pertinentes ne sont pas toujours incluses dans un modèle de régression, soit parce qu'elles n'ont pas été mesurées, soit parce qu'elles ne sont pas suspectées être liées à l'évènement d'intérêt. Ces variables omises peuvent créer une

sélection de la population au cours du suivi : les sujets les plus fragiles subissent l'évènement plus tôt que les autres, entraînant au fil du temps une modification de la structure de la population observée (Aalen, 1994; Vaupel and Yashin, 1985). Ignorer ces variables peut engendrer un biais dans l'estimation de la fonction de risque (sous-estimation) (Bretagnolle and Huber-Carol, 1988). La notion de fragilité introduite par Vaupel et al. (1979) traduit le fait que certains individus sont plus susceptibles de subir l'évènement, et donc sont plus « fragiles » que d'autres. La fragilité est représentée par un effet aléatoire qui, comme les variables explicatives, agit de façon multiplicative sur le risque instantané. Le modèle à fragilité est donc une extension du modèle à risques proportionnels de Cox :

$$\lambda_j(t|\omega_j, Z_j) = \lambda_0(t)\omega_j \mathrm{e}^{\beta^T Z_j} \tag{I.3}$$

où j est l'indice du sujet et ω_j est un effet aléatoire spécifique à chaque sujet, appelé variable à fragilité.

Modèles à fragilité partagée

Les modèles à fragilité sont aussi utilisés pour modéliser des données de survie corrélées telles que les données répétées et les données groupées (Clayton, 1978; Hougaard, 1995; Petersen et al., 1996). Dans le cas de données répétées, les individus correspondent aux différentes observations d'un même sujet ; dans le cas de données groupées qui est celui qui nous intéressera par la suite, ils correspondent aux différentes observations des sujets appartenant à un même groupe. Le modèle à fragilité partagée s'écrit :

$$\lambda_{ij}(t|\omega_i, Z_{ij}) = \lambda_0(t)\omega_i \mathrm{e}^{\beta^T Z_{ij}} \tag{I.4}$$

où i est l'indice du groupe, j l'indice du sujet, ω_i un effet aléatoire spécifique au groupe i.

Plusieurs distributions peuvent être utilisées pour ω. La plus courante est la loi Gamma car elle possède des propriétés mathématiques entraînant la simplification du calcul de la vraisemblance, ce qui évite de recourir à des méthodes d'intégration gourmandes en temps de calcul. Cette loi a cependant quelques propriétés indésirables (Hougaard, 2000, p. 256). En particulier, la non proportionnalité des risques pourrait avoir une influence plus grande sur les estimations que celle de la corrélation des données. La distribution log-normale est aussi largement utilisée. Contrairement à la loi

Gamma, elle ne permet pas de simplification de la vraisemblance mais se révèle très pratique dans un contexte multivarié avec plusieurs effets aléatoires non indépendants.

Dans les cas d'une distribution Gamma ou log-normale, la fragilité ω doit être d'espérance égale à 1 et de variance finie pour des questions d'identifiabilité. La variance est un paramètre à estimer et exprime l'hétérogénéité des données. Remarquons que l'équation I.4 peut se réécrire :

$$\lambda_{ij}(t|U_i, Z_{ij}) = \lambda_0(t)\mathrm{e}^{\beta^T Z_{ij}+U_i}$$

avec $\omega_i = \mathrm{e}^{U_i}$. Dans le cas log-normal, la variable U suit une loi normale $\mathcal{N}(0, \sigma^2)$ où σ^2 est le paramètre de variance à estimer.

Tester la significativité de l'effet aléatoire revient à tester si sa variance est significativement non nulle : $H_0 : \sigma^2 = 0$. La valeur 0 étant à la frontière de l'espace des paramètres, le test de rapport de vraisemblance basé sur une distribution asymptotique du χ^2 n'est pas applicable. La distribution asymptotique de la statistique de test de rapport des vraisemblances suit un mélange de χ^2 à 0 et 1 degré de liberté avec des poids égaux : $0.5\chi^2_0 + 0.5\chi^2_1$ (Verbeke and Molenberghs, 2009).

2 Modèles multi-états

L'analyse de survie étudie le délai de survenue d'un évènement d'intérêt, ou autrement dit, le délai entre deux états successifs (communément état « vivant » et état « décédé »). Les modèles multi-états, aujourd'hui très populaires, permettent d'étudier des dynamiques plus complexes en utilisant la notion de processus pour représenter l'évolution d'un sujet à travers différents états successifs (Andersen et al., 1993; Hougaard, 1999). En épidémiologie, ils permettent par exemple de modéliser son évolution à travers les différents stades d'une maladie. Un état peut être *transitoire* ou *absorbant* lorsqu'on y reste avec une probabilité égale à 1. Les figures I.1, I.2 et I.3 représentent trois exemples simples de modèles multi-états à trois états :

- un modèle progressif : les sujets transitent de l'état initial (0) à l'état absorbant (2) en passant obligatoirement par l'état transitoire (1) (figure I.1) ;
- un modèle *illness-death* : les sujets transitent de l'état initial (0) à l'état absorbant (2) directement ou en passant par l'état transitoire (1) (figure I.3) ;
- un modèle à risques concurrents : les sujets peuvent transiter vers deux états absorbants (qui correspondent typiquement à deux causes de décès) (figure I.2).

Figure I.1 – Modèle progressif à 3 états

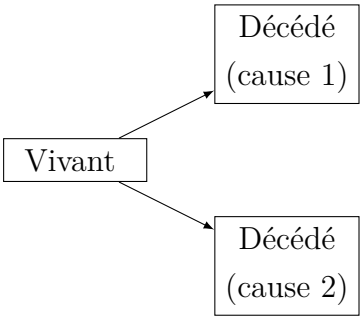


Figure I.2 – Modèle à 2 risques concurrents

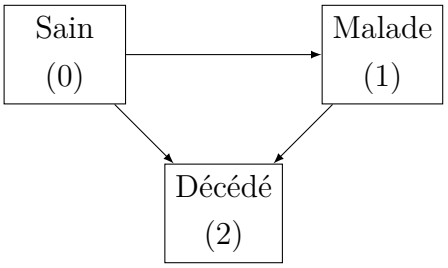


Figure I.3 – Modèle *illness-death*

2.1 Définitions

Soit $X = \{X(t), t \geq 0\}$ un processus à temps continu et à espace d'états fini *i.e.* $X(t) \in \mathcal{S} = \{0, 1, \ldots, K\}$ où $K+1$ est le nombre d'états possibles. X peut être caractérisé par les *probabilités de transition* entre les différents états :

$$p_{kl}(s,t) = \mathbb{P}\Big(X(t) = l \,\Big|X(s) = k, \mathcal{H}_{s-}\Big) \;, \quad k,l \in \mathcal{S}$$

où $\mathcal{H}_{s-}$ représente l'histoire du processus générée par $\{X(u), u < s\}$, ou, par ses *intensités de transition* :

$$\alpha_{kl}(t) = \lim_{\Delta t \to 0} \frac{p_{kl}(t, t+\Delta t)}{\Delta t} \;, \quad k,l \in \mathcal{S}$$

qui sont le pendant de la fonction de risque dans un modèle de survie.

Les *intensités de transition cumulées* sont définies par :

$$A_{kl}(t) = \int_0^t \alpha_{kl}(u)\mathrm{d}u$$

X est un ***processus de Markov*** si l'évolution future du processus dépend uniquement de la connaissance du temps présent s et de l'état en ce temps $X(s)$ (le passé peut être oublié) :

$$p_{kl}(s,t) = \mathbb{P}\Big(X(t) = l \,\Big|X(s) = k\Big) \;, \quad k,l \in \mathcal{S}$$

On définit un *processus de Markov homogène* en supposant en plus que les probabilités de transition dépendent uniquement du délai entre chaque observation et non du temps auquel se produisent ces observations *i.e.* du délai depuis le temps d'origine :

$$p_{kl}(s,t) = p_{kl}(0,t-s)$$

Les intensités de transition sont alors constantes : $\alpha_{kl}(t) = \alpha_{kl}$.

Ce cas particulier est particulièrement intéressant car on peut déduire de l'équation de Chapman-Kolmogorov (donnée par $p_{kl}(s,t) = \sum_{i \in \mathcal{S}} p_{ki}(s,u)p_{il}(u,t)$) une relation simple entre la matrice des probabilités de transition P et celle des intensités de transition Q (Cox and Miller, 1965).

En effet, de l'équation de Chapman-Kolmogorov, se déduisent des équations différentielles appelées équations arrière (*backward*) et avant (*forward*) de Kolmogorov :

$$\frac{\partial P(0,t)}{\partial t} = QP(0,t) \quad ; \quad \frac{\partial P(0,t)}{\partial t} = P(0,t)Q$$

Avec la condition initiale $P(0,0) = I$ (*i.e.* $p_{kk}(0,0) = 1$ et $p_{kl}(0,0) = 0$, $k \neq l$), la solution de l'une ou de l'autre est :

$$P(0,t) = \mathrm{e}^{Qt} \tag{I.5}$$

Remarque I.5 *On peut relâcher un peu l'hypothèse d'homogénéité qui est très forte en considérant un processus de Markov homogène par périodes et donc des intensités de transition constantes par morceaux (constantes sur la même période et différentes d'une période à l'autre). Dans ce cas, on peut encore faire le lien entre intensités et probabilités de transition grâce aux équations de Kolmogorov qui sont alors généralement résolues de façon numérique.*

Remarque I.6 *Pour les processus simples, il est possible d'avoir une expression explicite des probabilités de transition en fonction des intensités de transition en utilisant des intégrales (cf. chapitre III pour un processus illness-death). Il est alors possible de considérer un processus de Markov non homogène avec des intensités de transition beaucoup plus flexibles.*

Un ***processus de semi-Markov*** (Janssen and Limnios, 1999) est un processus dont l'évolution future dépend du temps auquel la dernière transition a eu lieu en plus

du temps présent s et de l'état en ce temps $X(s)$:

$$p_{kl}(s,t) = \mathbb{P}\Big(X(t) = l \,\Big|X(s) = k, t_k\Big) \ , \quad k,l \in \mathcal{S}$$

où t_k est le temps d'entrée dans l'état k. Par rapport à un processus de Markov, on tient compte en plus du temps passé dans l'état actuel. Les intensités de transition sont de la forme $\alpha_{kl}(t, t-t_k)$. Lorsqu'on ne tient compte que du temps passé dans l'état actuel et plus du temps proprement dit, on obtient un *processus de semi-Markov homogène* (dans le temps) avec des intensités de transition de la forme $\alpha_{kl}(t-t_k)$.

2.2 Prise en compte des facteurs de risque

Dans un modèle multi-état, les intensités de transition permettent de prendre en compte des variables explicatives de façon similaire à la fonction de risque en analyse de survie. Ainsi, le modèle qui découle du modèle à risques proportionnels de Cox est un modèle à intensités de transition proportionelles :

$$\alpha_{kl}\Big(t \,\Big|Z_{kl}^{(i)}\Big) = \alpha_{0,kl}(t)\mathrm{e}^{\beta_{kl}^T Z_{kl}^{(i)}} \tag{I.6}$$

où i est l'indice du sujet, $Z_{kl}^{(i)}$ et β_{kl} les vecteurs des variables explicatives et des coefficients de régression associés à la transition $k \to l$, $\alpha_{0,kl}$ l'intensité de transition de base associée à la transition $k \to l$.

2.3 Modèles et estimation

Dans cette partie (et plus généralement dans ce manuscrit), sauf mention contraire, nous ne considérons que des modèles de Markov, c'est-à-dire des modèles qui font l'hypothèse que le processus sous-jacent au modèle est Markovien. De plus, nous considérons des modèles de Markov non homogènes dans le temps, plus plausibles d'un point de vue épidémiologique.

Le choix de la méthode d'estimation d'un modèle multi-état et l'ajout éventuel d'hypothèses supplémentaires à l'hypothèse Markovienne dépendent beaucoup du schéma d'observation des données. En effet, on étudie un processus à temps continu mais l'observation des transitions de passage d'un état à l'autre ne se fait pas toujours en temps continu.

2.3.1 Observations en temps continu

Le cas le plus simple est celui où tous les temps de transition sont connus de façon exacte.

Intéressons-nous d'abord à un modèle multi-état sans variables explicatives. Grâce au travail fondamental d'Aalen (1975, 1978), les techniques classiques d'analyse de survie ont pu se généraliser aux modèles multi-états dans le contexte de la théorie des processus de comptage (Andersen and Borgan, 1985).

Soit $N_{kl}(t)$ le processus de comptage du nombre de transitions $k \to l$ qui ont eu lieu dans l'intervalle de temps $[0, t]$. Soit $Y_k(t)$ le processus relatif au nombre de sujets à risque au temps t^- pour la transition $k \to l$. L'indice l est omis car le nombre de sujets à risque est le même pour toutes les transitions au départ de l'état k et correspond au nombre de sujets en l'état k au temps t^-. L'estimateur de Nelson-Aalen de la fonction d'intensité cumulée pour la transition $k \to l$, $A_{kl}(t)$, est donné par :

$$\hat{A}_{kl}(t) = \int_0^t \frac{\mathrm{d}N_{kl}(u)}{Y_k(u)} = \sum_{t_i \leq t} \frac{m_{kl}(t_j)}{Y_k(t_j)}$$

où les t_j sont les temps d'évènement et m_{kl} est le nombre de transitions $k \to l$ au temps t_j.

Les incréments de l'estimateur de Nelson-Aalen peuvent fournir un estimateur des intensités de transition : $\mathrm{d}\hat{A}_{kl}(t) = \frac{\mathrm{d}N_{kl}(t)}{Y_k(t)}$. En lissant ces incréments, il est possible d'obtenir des estimations lisses des intensités de transition (Ramlau-Hansen, 1983).

En généralisant l'estimateur de Kaplan-Meier aux modèles de Makov non homogène, Aalen and Johansen (1978) ont proposé un estimateur des probabilités de transition, l'estimateur dit de Aalen-Johansen. Celui-ci peut être également vu comme le *product integral* de la matrice des estimateurs de Nelson-Aalen (Gill and Johansen, 1990).

Intéressons-nous maintenant à un modèle multi-état avec des variables explicatives qui interviennent grâce à des modèles de régression définis par l'équation I.6. Les estimateurs de Nelson-Aalen et de Aalen-Johansen se généralisent. On peut obtenir des estimations pour les paramètres de régression spécifiques à chaque transition $k \to l$ en considérant autant de modèles de Cox qu'il y a de transitions (Putter et al., 2007). Un modèle de Cox sur la transition $k \to l$ doit être estimé sur le sous-échantillon des sujets à risque pour cette transition. Par exemple, dans un modèle *illness-death*, l'échantillon entier sera utilisé pour estimer β_{01} et β_{02} mais seulement le sous-échantillon des sujets ayant fait la transition $0 \to 1$ au cours du suivi sera utilisé pour estimer β_{12}.

En pratique, le paquet R **mstate** (De Wreede et al., 2010, 2011) peut être utilisé pour analyser des modèles markoviens multi-états, avec ou sans variables explicatives, lorsque les données sont observées en temps continu. Il permet l'estimation des intensités de transition et des éventuels paramètres de régression mais aussi de faire des prévisions en estimant les probabilités de transition.

Remarque I.7 ***Sur l'estimateur de Kaplan-Meier*** *Il faut être prudent avec l'utilisation de l'estimateur de Kaplan-Meier dans un contexte multi-état. Prenons l'exemple d'un modèle à deux risques concurrents (voir figure I.2). La relation directe qui lie la fonction de risque cumulée et la fonction de survie (voir équation I.1) n'est plus valable dans ce contexte. Il en résulte que l'estimateur naïf de Kaplan-Meier de la fonction de survie calculé en censurant à droite les sujets ayant subi l'évènement concurrent (à l'évènement d'intérêt) est biaisé (Andersen et al., 2012; Putter et al., 2007). Plus précisément, il sous-estime la fonction de survie. En effet, en analyse de survie, l'évènement d'intérêt est dit « de mortalité toutes causes » (all-cause mortality), c'est-à-dire que si le suivi était suffisamment long, l'évènement surviendrait avec une probabilité 1. Ce n'est plus le cas dans un contexte de risques concurrents où la réalisation de l'évènement concurrent empêche la réalisation de l'évènement d'intérêt. La fonction de survie dépend à la fois du risque associé à l'évènement d'intérêt et du risque associé à l'évènement concurrent :* $S(t) = \mathrm{e}^{-\mathrm{A}_1(\mathrm{t})-\mathrm{A}_2(\mathrm{t})}$ *où* A_1 *est la fonction de risque cumulé « cause 1-spécifique » et* A_2 *est la fonction de risque cumulé « cause 2-spécifique ». Elle peut être estimée correctement grâce à l'estimateur de Kaplan-Meier de la fonction de survie dite « globale ». Cet estimateur de Kaplan-Meier est en fait le produit des estimateurs de Kaplan-Meier naïfs associés à chaque évènement.*

Sous une hypothèse d'indépendance des deux risques concurrents, l'estimateur de Kaplan-Meier naïf associé à la cause 1 de décès correspondrait à la survie d'une population hypothétique dans laquelle la cause 2 de décès n'existe pas (et vive versa). Cependant, une telle hypothèse n'est pas vérifiable sur la base des données et est souvent peu crédible d'un point de vue biologique.

Remarque I.8 ***Approche « stochastic process » vs. approche « latent failure times »*** *Dans ce manuscrit, nous nous plaçons toujours dans une approche multi-état avec un processus stochastique sous-jacent. Une autre approche que l'on trouve dans la littérature des modèles à risques concurrents consiste à considérer des temps d'évènement latents.*

1. *Plaçons-nous dans le modèle à deux risques concurrents de la figure I.2 et considérons les deux variables aléatoires T_1 (temps de décès par la cause 1) et T_2 (temps de décès par la cause 2) telles que :*

$$T_j = \inf_{t>0}\Big(X(t) = j\Big) \ , \quad j = 1, 2$$

dont la distribution est donnée par les fonctions « d'incidence cumulée » :

$$F_j(t) = \mathbb{P}(T \leq t, D = j)$$

où $T = min(T_1, T_2)$ et D est la cause de décès.

On remarque que F_1 et F_2 ne sont pas des fonctions de répartition : $\lim_{t\to\infty} F_j(t) = \mathbb{P}(D = j) < 1$ $(j = 1, 2)$. Ceci implique que les variables aléatoires T_1 et T_2 ne sont pas définies correctement car par exemple pour un sujet j tel que $D_j = 2$ on a : $T_1 = \infty$.

En l'absence de risques concurrents, la distribution de T_j serait donnée par :

$$\tilde{F}_j(t) = 1 - \exp\left\{-\int_0^t \alpha_j(u)\mathrm{d}u\right\}$$

(par analogie avec l'équation I.1 d'un modèle de survie).

Cette fonction est aussi la distribution de la variable latente mais définie correctement $\tilde{T}_j$ qui est le temps de décès par la cause j dans un monde dans lequel j serait la seule cause de décès. L'approche « latent failure times » consiste à imaginer l'existence des temps latents $\tilde{T}_1$, $\tilde{T}_2$ et de s'intéresser à leur distribution jointe qui est non identifiable à moins de faire certaines hypothèses fortes comme l'indépendance des risques, impossible à vérifier sur la base des données disponibles. Une autre façon de pallier ce problème de non identifiabilité est de modéliser l'association des temps latents en utilisant une copule (Zheng and Klein, 1995). Cependant, ce genre d'approches, faisant intervenir des notions assez évasives, est peu utilisé car suppose des hypothèses non vérifiables sur la base des données existantes et peut entraîner des interprétations hasardeuses (Andersen and Keiding, 2012; Tsiatis, 1975).

Remarquons qu'avec une approche multi-état avec processus sous-jacent, aucune hypothèse d'indépendance n'est requise. En effet, c'est parce que la vraisemblance se factorise que chaque risque cause-spécifique $\alpha_1(.)$, $\alpha_2(.)$ peut être analysé sépa-

rément en assimilant les décès « autre cause » à des censures à droite (Andersen et al., 2002).

2. *Considérons maintenant le modèle illness-death de la figure I.3. Dans la littérature, cette configuration illness-death peut être évoquée sous le terme de semi-competing risks (risques semi-concurrents). Les risques semi-concurrents y sont définis de la façon suivante : les sujets peuvent subir deux types d'évènements, un évènement terminal et un évènement non terminal ; l'évènement terminal censure l'évènement non terminal mais pas inversement. Cette description est généralement associée à une analyse basée sur l'estimation d'une fonction de survie jointe de deux temps d'évènements et entretient implicitement l'idée de temps d'évènement latents (latent failure times). De la même façon que lorsqu'on considère un modèle à risques concurrents, cette approche est déconseillée (Xu et al., 2010). Nous nous en tenons dans ce manuscrit à la formulation modèle illness-death qui évite toute allusion à des temps d'évènement latents.*

2.3.2 Observations en temps discret : *panel data*

Le cas le plus délicat est celui où l'observation des différents états se fait en temps discret. Les temps de transition ne sont alors pas connus de façon exacte et le chemin pris pour aller d'un état à l'autre entre deux temps d'observation consécutifs peut être inconnu avec un nombre de chemins possibles pouvant être grand, voire infini. Les modèles multi-états adaptés à de telles données sont souvent désignés dans la littérature sous le terme de *Markov models for panel data*.

Dans ce type de modèles, décrits d'abord par Kalbfleisch and Lawless (1985), l'estimation est basée sur la résolution des équations différentielles de Kolmogorov qui permet d'obtenir les probabilités de transition et donc la vraisemblance (qui s'écrit en fonction des probabilités de transition). C'est pourquoi l'hypothèse d'homogénéité dans le temps du processus est souvent faite bien qu'elle soit très forte. La vraisemblance s'écrit :

$$\prod_{i,j} L_{ij} = \prod_{i,j} p_{S(t_{ij})S(t_{i,j+1})}(t_{ij}, t_{i,j+1}) = \prod_{i,j} p_{S(t_{ij})S(t_{i,j+1})}(t_{i,j+1} - t_{ij})$$

où i est l'indice du sujet, j est l'indice correspondant aux différents temps d'observation (j est un entier allant de 1 au nombre de temps d'observation du sujet i), $S(t_{ij})$ est l'état dans lequel se trouve le sujet i au $j^{\text{ème}}$ temps d'observation.

Des solutions analytiques des équations existent cependant aussi pour des intensités de transition non constantes.

— Lorsque les intensités de transition sont constantes par morceaux (Kay, 1986), la résolution des équations peut être faite sur chaque intervalle de temps. Le modèle obtenu est très flexible mais l'hypothèse de fonctions d'intensité de transition discontinues n'est pas très plausible d'un point de vue biologique. De plus, une hypothèse de temps de séjour distribués de façon exponentielle par morceaux est inhérente à ces modèles.

— En utilisant des modèles de transformation (Hubbard et al., 2008), on peut se ramener à un processus de Markov homogène dans le temps. On peut alors supposer par exemple des distributions de Weibull pour les intensités de transition (Omar et al., 1995). Une des limitations de ces modèles est que le ratio des intensités de transition (vers des états différents) doit rester constant dans le temps. Cette hypothèse est parfois peu réaliste. Par exemple, dans un modèle *illness-death*, on peut s'attendre à ce que le taux de mortalité α_{02} croisse plus rapidement avec l'âge que le taux d'incidence α_{01}.

D'autres formes plus flexibles ont été proposées pour les intensités de transition. Titman (2011) propose d'utiliser des méthodes basées sur une solution numérique des équations de Kolmogorov et utilise des B-splines quadratiques pour modéliser les intensités de transition. Cependant, son modèle peut être rapidement limité lorsqu'on y inclut des variables explicatives car alors, autant d'équations différentielles doivent être résolues qu'il y a de valeurs différentes des variables explicatives dans le jeu de données.

En pratique, le paquet R **msm** (Jackson, 2011) peut être utilisé pour l'analyse de modèles multi-états lorsque les données sont observées en temps discret. Les intensités de transition y sont supposées constantes ou constantes par morceaux, et des variables explicatives peuvent être inclues dans des modèles de régression à intensités proportionnelles (équation I.6).

Cas particuliers

Pour les modèles progressifs ou hiérarchiques (sans retour en arrière possible), le nombre de chemins possibles entre deux temps d'observation est fini. L'écriture de la vraisemblance peut être développée et lorsque le nombre d'intégrations est faible il vaut mieux utiliser des méthodes plus flexibles basées sur la vraisemblance. Par exemple, dans un modèle *illness-death* sans rétablissement possible où les temps de décès sont observés en temps continu et les temps de maladie en temps discret, il y a un ou

deux chemins possibles entre deux temps d'observation et les temps de maladie sont seulement censurés par intervalle : lorsqu'un sujet est observé en l'état 1, on sait que le passage de 0 à 1 a eu lieu dans un intervalle de temps donné. Le modèle *illness-death* pour données censurées par intervalle, étant celui qui nous intéresse dans l'application à l'étude de la démence, est plus amplement développé dans la sous-section suivante.

2.4 Modèle *illness-death* et données censurées par intervalle

Nous nous intéressons dans cette partie au cas particulier d'un modèle illness-death (voir figure I.3) sans retour en arrière possible. Les différents états peuvent différer selon les applications mais pour plus de clarté nous admettons dans cette partie que les états 0, 1 et 2 correspondent aux états dit « sain », « malade » et « décédé ». Dans un tel modèle, les sujets sont donc initialement sains et peuvent décéder avec ou sans maladie. Nous admettons que les temps de décès (temps de transition $0 \to 2$ ou $1 \to 2$) sont observés en temps continu (connus exactement) et que les temps d'apparition de la maladie (temps de transition $0 \to 1$) sont observés en temps discret (censurés par intervalle).

Frydman (1995) et Frydman and Szarek (2009, 2010) ont proposé des estimateurs non paramétriques des intensités de transition cumulées A_{01}, A_{02}, A_{12} et des probabilités cumulées de maladie et de décès F_{01} et F_{02} (voir chapitre III) pour des modèles *illness-death* avec données censurées par intervalle. Ils n'ont cependant pas envisagé l'inclusion de variables explicatives dans leur modèle. Nous privilégions des approches paramétriques ou semi-paramétriques qui permettent d'inclure facilement des variables (par exemple grâce à des modèles à intensités proportionnelles) et d'obtenir des estimations lisses des intensités de transition.

2.4.1 Introduction

Les méthodes développées dans la présente sous-section sont basées sur la vraisemblance du modèle. Comme nous le verrons dans le chapitre III, l'écriture des probabilités de transition p_{00}, p_{01}, p_{02}, p_{11}, p_{12} en fonction des intensités de transition α_{01}, α_{02}, α_{12} est relativement simple (avec au plus une intégrale). La vraisemblance qui s'écrit en fonction des α_{kl} et des p_{kl} peut alors être écrite en fonction des α_{kl} uniquement qui sont alors estimées par maximum de vraisemblance ou de vraisemblance pénalisée. Des formes flexibles sont donc possibles pour les intensités de transition (lois paramétriques, splines) sans hypothèse supplémentaire. Afin de faciliter la lecture de cette

sous-section, les variables explicatives sont omises dans les notations des α_{kl} et des p_{kl}. Bien évidemment, il est possible d'en inclure grâce à des modèles de régression sur les α_{kl} par exemple ceux de l'équation I.6.

2.4.2 Vraisemblance

Soit a le temps à partir duquel les sujets sont sous observation. En a, tous les sujets sont non malades (en l'état 0). a peut être nul mais peut aussi être par exemple un âge correspondant à un temps de troncature à gauche. Les temps d'observation relatifs à l'apparition de la maladie (*i.e.* à la transition $0 \to 1$) pour un sujet i donné sont supposés indépendants du processus *illness-death*. Soit δ_2 un indicateur de décès (1 si le sujet est décédé, 0 sinon). Soit $\tilde{T}$ le temps de décès si $\delta_2 = 1$, le temps de dernières nouvelles si $\delta_2 = 0$. On note l et r (pour *left* et *right*) les bornes gauche et droite de l'intervalle de censure d'un sujet qui a été observé malade. Si un sujet n'a jamais été observé malade, l est le temps de dernière observation.

Si le sujet i a été vu pour la dernière fois au temps l, et qu'il n'était pas malade, sa contribution à la vraisemblance est :

$$L_i = p_{00}(a,l)\Big(p_{00}(l,\tilde{T})\alpha_{02}(\tilde{T})^{\delta_2} + p_{01}(l,\tilde{T})\alpha_{12}(\tilde{T})^{\delta_2}\Big) \tag{I.7}$$

où $p_{00}(s,t)$ et $p_{01}(s,t)$ représentent les probabilités pour un sujet non malade en s d'être respectivement toujours non malade et malade en t.

Ainsi, en l, le sujet était toujours dans l'état 0. Puis, entre l et $\tilde{T}$, il a pu rester dans l'état 0 (et décéder en $\tilde{T}$ si $\delta_2 = 1$) ou, transiter vers l'état 1 (et décéder en $\tilde{T}$ si $\delta_2 = 1$). Si le sujet n'est pas décédé et que sa date de dernière nouvelle $\tilde{T}$ coïncide avec la dernière date à laquelle il a été observé non malade l, l'expression de la contribution du sujet j se simplifie : $L_i = p_{00}(a,l)$.

Supposons maintenant que le sujet i a été observé malade en r (et non malade en l). Sa contribution à la vraisemblance est :

$$L_i = p_{00}(a,l)p_{01}(l,r)p_{11}(r,\tilde{T})\alpha_{12}(\tilde{T})^{\delta_2} \tag{I.8}$$

où $p_{11}(s,t)$ représente la probabilité pour un sujet malade en s d'être toujours malade en t (*i.e.* de ne pas décéder entre s et t).

Ainsi, en l, le sujet i était toujours dans l'état 0. Puis, entre l et r, il a transité vers l'état 1. Enfin, il est resté en 1 jusqu'au temps $\tilde{T}$, où il est décédé si $\delta_2 = 1$.

Si le sujet n'est pas décédé et que $\tilde{T}$ coïncide avec r, la contribution se simplifie : $L_i = p_{00}(a,l)p_{01}(l,r)$.

Détaillons maintenant la contribution à la vraisemblance en exprimant les p_{kl} en fonction des intensités de transition α_{kl} (les formules des p_{kl} en fonction des α_{kl} ainsi que leur explication sont disponibles dans le chapitre III à la section 2), en distinguant les différents cas de figure possibles, et en nous plaçant dans le contexte qui nous intéresse, à savoir l'étude de la démence à partir de données de cohorte. Les états 0, 1 et 2 du modèle *illness-death* correspondent respectivement aux états « non dément », « dément » et « décédé ». Les différents temps d'observation correspondent aux différentes visites de suivi. L'échelle de temps choisie est l'âge. a représente l'âge d'entrée dans la cohorte.

— Si le sujet i est toujours vivant à l'âge $\tilde{T} = C$ de dernières nouvelles et n'était pas dément à l'âge l auquel il a été vu pour la dernière fois :

$$L_i = \frac{1}{\mathrm{e}^{-A_{01}(a)-A_{02}(a)}} \left\{ \mathrm{e}^{-A_{01}(C)-A_{02}(C)} + \int_l^C \mathrm{e}^{-A_{01}(u)-A_{02}(u)} \alpha_{01}(u) \frac{\mathrm{e}^{-A_{12}(C)}}{\mathrm{e}^{-A_{12}(u)}} \, \mathrm{d}u \right\} \tag{I.9}$$

Dans la pratique, la date de dernières nouvelles concernant le statut vital correspond souvent à la date de dernière visite. Dans ce cas, la contribution s'écrit :

$$L_i = \frac{\mathrm{e}^{-A_{01}(l)-A_{02}(l)}}{\mathrm{e}^{-A_{01}(a)-A_{02}(a)}}$$

— Si le sujet i est décédé à l'âge $\tilde{T} = T$ et qu'il n'était pas dément à l'âge l auquel il a été vu pour la dernière fois :

$$L_i = \frac{1}{\mathrm{e}^{-A_{01}(a)-A_{02}(a)}} \Big\{ \mathrm{e}^{-A_{01}(T)-A_{02}(T)} \alpha_{02}(T) + \int_l^T \mathrm{e}^{-A_{01}(u)-A_{02}(u)} \alpha_{01}(u) \frac{\mathrm{e}^{-A_{12}(T)}}{\mathrm{e}^{-A_{12}(u)}} \alpha_{12}(T) \, \mathrm{d}u \Big\} \tag{I.10}$$

— Si le sujet i a été diagnostiqué malade à l'âge r et est toujours vivant à l'âge C de dernières nouvelles :

$$L_i = \frac{1}{\mathrm{e}^{-A_{01}(a)-A_{02}(a)}} \left\{ \int_l^r \mathrm{e}^{-A_{01}(u)-A_{02}(u)} \alpha_{01}(u) \frac{\mathrm{e}^{-A_{12}(C)}}{\mathrm{e}^{-A_{12}(u)}} \, \mathrm{d}u \right\} \tag{I.11}$$

où l est son âge à la visite précédant la visite de diagnostic.

— Si le sujet i a été diagnostiqué malade à l'âge r et est décédé à l'âge T :

$$L_i = \frac{1}{\mathrm{e}^{-A_{01}(a)-A_{02}(a)}} \left\{ \int_l^r \mathrm{e}^{-A_{01}(u)-A_{02}(u)} \alpha_{01}(u) \frac{\mathrm{e}^{-A_{12}(T)}}{\mathrm{e}^{-A_{12}(u)}} \alpha_{12}(T) \, \mathrm{d}u \right\} \tag{I.12}$$

2.4.3 Estimation

Nous venons de voir que la vraisemblance s'écrit de façon explicite en fonction des intensités de transition α_{kl}. Nous détaillons dans cette partie les deux méthodes d'estimation des α_{kl}, basées sur la vraisemblance, qui ont été utilisées dans cette thèse et qui sont disponibles dans le paquet R **SmoothHazard**. Dans la première, les estimations des intensités de transition de base sont de type Weibull tandis que dans la seconde ce sont des combinaisons linéaires de M-splines. Elles dépendent de variables explicatives à travers trois modèles à intensités de transition proportionnelles (voir équation I.6).

Méthode paramétrique

Nous supposons que les intensités de transition de base $\alpha_{0,kl}$ ont une forme paramétrique de type Weibull :

$$\alpha_{0,kl}(t) = a_{kl} \left(\frac{1}{b_{kl}} \right)^{a_{kl}} t^{a_{kl}-1}$$

où a_{kl} et b_{kl} sont les paramètres de forme et d'échelle relatifs à la transition $k \to l$, $(k,l) \in \{(0,1),(0,2),(1,2)\}$.

Les paramètres a_{01}, b_{01}, a_{02}, b_{02}, a_{12}, b_{12}, β_{01}^T, β_{02}^T, β_{12}^T sont estimés par maximum de vraisemblance.

Méthode semi-paramétrique

La vraisemblance, et donc la log-vraisemblance, pouvant s'écrire en fonction des intensités de transition, notons $l(\alpha_{01}, \alpha_{02}, \alpha_{12})$ la log-vraisemblance. Nous définissons une log-vraisemblance pénalisée en pénalisant la log-vraisemblance par un terme relatif à la courbure des intensités de transition :

$$pl(\alpha_{01}, \alpha_{02}, \alpha_{12}) = l(\alpha_{01}, \alpha_{02}, \alpha_{12}) - \kappa_{01} \int \alpha_{01}^{''\,2}(u)\mathrm{d}u - \kappa_{02} \int \alpha_{02}^{''\,2}(u)\mathrm{d}u - \kappa_{12} \int \alpha_{12}^{''\,2}(u)\mathrm{d}u \tag{I.13}$$

où $\kappa_{01}, \kappa_{02}, \kappa_{12}$ sont des paramètres de lissage positifs qui réalisent un compromis entre fidélité aux données et régularité.

Les estimateurs des intensités de transition de base définis par maximum de vraisemblance pénalisée $\hat{\alpha}_{kl}^0$ sont approximés à l'aide d'une base de M-splines (Ramsay, 1988).

Une famille de M-splines d'ordre q, $M_1, \cdots, M_n$ est définie sur un ensemble de nœuds : $t = (t_1 \leq t_2 \leq \ldots \leq t_{n+q})$ où $n = m + q$ est le nombre de nœuds intérieurs additionné de l'ordre q. Ces nœuds sont tels que :

- $t_1 = \cdots = t_q$;
- $t_{n+1} = \cdots = t_{n+q}$;
- $t_i < t_{i+q} \forall i$.

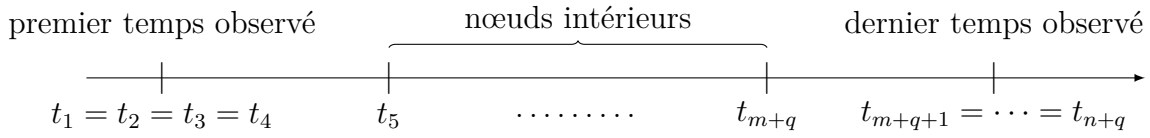
Une M-spline $M_i(\cdot|q,t)$ a les propriétés mathématiques suivantes :

- $M_i(\cdot|q,t)$ est positive ;
- $M_i(\cdot|q,t)$ est nulle en dehors de l'intervalle $[t_i; t_{i+q}]$;
- $M_i(\cdot|q,t)$ possède $q-2$ dérivées continues en $t_{q+1}, \cdots, t_{n-1}$ (nœuds intérieurs) ;
- $\int M_i(x|q,t)\mathrm{d}x = 1$.

Un algorithme récursif est utilisé pour les calculer :

$$\begin{cases} \text{Pour } q = 1, \quad M_i(x|1,t) = \dfrac{1}{t_{i+1} - t_i} \quad \text{si } t_i \leq x \leq t_{i+1} \ , \ 0 \text{ sinon} ; \\ \\ \text{Pour } q > 1, \quad M_i(x|q,t) = \dfrac{q[(x - t_i)M_i(x|q-1,t) + (t_{i+q} - x)M_{i+1}(x|q-1,t)]}{(q-1)(t_{i+q} - t_i)} \ . \end{cases}$$

Nous utilisons une famille de M-splines cubiques, c'est-à-dire d'ordre $q = 4$: $M_1, \cdots, M_n$, où $n = m_{kl} + 4$ avec m_{kl} le nombre de nœuds intérieurs choisis sur la transition $k \to l$ ($m_{kl} + 2$ est le nombre de nœuds). En général, le premier nœud choisi correspond au premier temps observé, le dernier nœud au dernier temps observé (sur la transition $k \to l$) et les autres nœuds sont placés de façon équidistante. La séquence de nœuds peut-être schématisée de la façon suivante :



Les estimateurs par maximum de (log-)vraisemblance pénalisée des intensités de transition de base, $\hat{\alpha}_{kl}^0$, sont alors approchés par une combinaison linéaire des n M-splines :

$$\tilde{\alpha}_{kl}^0(t) = \sum_{i=1}^{n} a_{kl,i} M_{kl,i}(x)$$

où les $a_{kl,i}$ sont les coefficients à estimer.

Une contrainte de positivité sur les coefficients $a_{kl,i}$ assure la positivité des $\tilde{\alpha}_{kl}$. En pratique, nous estimons des paramètres $\theta_{kl,i}$ tels que $a_{kl,i} = \theta_{kl,i}^2$ ou $a_{kl,i} = \mathrm{e}^{\theta_{kl,i}}$.

Les M-splines peuvent être intégrées pour produire une famille de splines monotones, appelées I-splines : $I_{kl,i}(x) = \int_0^x M_{kl,i}(u)\mathrm{d}u,\ \ i = 1, \cdots, n$. Les même coefficients $a_{kl,i}$ permettent ainsi d'approcher les estimateurs des intensités de transition cumulées de base, $\hat{A}_{kl}^0$, avec une combinaison linéaire de I-splines :

$$\tilde{A}_{kl}^0(x) = \sum_{i=1}^{n} a_{kl,i} I_{kl,i}(x)$$

La contrainte de positivité sur les a_i assure la croissance monotone des $\tilde{A}_{kl}$.

En pratique, cette méthode d'estimation est semblable à une méthode paramétrique : pour κ_{01}, κ_{02}, κ_{12} fixés, les paramètres de régression β_{01}^T, β_{02}^T, β_{12}^T et les paramètres des splines $a_{01,1}, \cdots, a_{01,n}$, $a_{02,1}, \cdots, a_{02,n}$, $a_{12,1}, \cdots, a_{12,n}$ sont estimés en maximisant la vraisemblance pénalisée.

Les paramètres de lissage κ_{01}, κ_{02} et κ_{12} peuvent être choisis de façon arbitraire ou en utilisant une technique automatique comme celle de la validation croisée. La *leave-one-out cross validation* consiste à utiliser une seule observation comme échantillon de validation et les observations restantes comme échantillon d'apprentissage. C'est un cas particulier de la *K-fold cross validation* où K est égal au nombre d'observations dans l'échantillon initial. La maximisation du score de *leave-one-out cross validation* est cependant coûteuse puisqu'elle nécessite pour chaque valeur de κ_{01}, κ_{02} et κ_{12} de maximiser la vraisemblance autant de fois qu'il y a d'observations dans l'échantillon initial. Nous utilisons une approximation du score de *leave-one-out cross validation* proposé initialement par O'Sullivan (1988a) pour les modèles de survie et étendu aux modèles multi-états par Commenges et al. (2007), pour lequel une seule maximisation de vraisemblance suffit. Quelques conseils pratiques sur le choix des κ_{kl} sont donnés dans le chapitre III et dans l'annexe A.

Algorithme de maximisation

L'algorithme utilisé pour maximiser la log-vraisemblance ou la log-vraisemblance pénalisée est l'algorithme de Levenberg-Marquardt (Levenberg, 1944; Marquardt, 1963) qui consiste à alterner deux algorithmes : l'algorithme de Newton-Raphson et l'algorithme du gradient (aussi connu sous le nom d'algorithme de la plus profonde descente). Pour des points éloignés de la solution, l'algorithme du gradient, plus robuste, est utilisé. Pour des points proches de la solution, il est relayé par l'algorithme de Newton-Raphson, plus rapide.

Chapitre II

Modèle de régression : estimation des effets des facteurs de risque de la démence

Le travail présenté dans ce chapitre a été motivé par l'étude des facteurs de risque de la démence à partir des données de la cohorte Paquid. La particularité de ces données provient de la concomitance de deux choses : la censure par intervalle et le risque de décès. Une approche naïve consiste à les « ignorer » afin de se ramener à des données appropriées aux techniques classiques d'analyse de survie. Le but de ce travail a été d'étudier l'impact d'une telle approche sur l'estimation des effets des facteurs de risque de démence. Nous livrons ici une version plus approfondie que celle qui a été publiée (Leffondré et al., 2013).

1 Introduction

1.1 Données Paquid

La cohorte Paquid (Personnes Âgées QUID) a pour objectif général d'étudier le processus de vieillissement cérébral chez les personnes âgées de plus de 65 ans et d'en distinguer les modalités normales et pathologiques. En particulier, des recherches sont menées afin d'explorer les facteurs de risque d'une détérioration des capacités mentales à l'origine d'une démence sénile. Pour être inclus dans la cohorte, les sujets doivent vivre à leur domicile, être âgés de 65 ans ou plus et être non déments. L'échantillon de Paquid est constitué de 3675 sujets initialement non déments, recrutés en 1988, et répartis sur 75 communes des départements de la Gironde et de la Dordogne. Un tirage aléatoire stratifié selon l'âge, le sexe et la taille des unités urbaines de résidence a été effectué afin que l'échantillon soit représentatif de la population générale. En plus d'une visite initiale, des visites à domicile ont été effectuées à 3, 5, 8, 10, 13, 15, 17

et 20 ans par des psychologues. Les sujets girondins ont en plus reçu une visite à 1 an. À chaque visite, les performances cognitives des sujets ont été mesurées à l'aide de tests psychométriques et un dépistage de démence a été effectué selon les critères du manuel diagnostique et statistique des troubles mentaux. Lorsque ce dépistage s'est avéré positif, un neurologue a ensuite réalisé une visite à domicile afin de confirmer ou d'infirmer le diagnostic de démence et de préciser son étiologie (Alzheimer, vasculaire, *etc.*).

Nous nous intéressons dans ce chapitre à l'étude des facteurs de risque associés à la survenue d'une démence à partir des données de Paquid.

1.2 Problématique

Les sujets de la cohorte étant âgés, leur risque de décès au cours du suivi est important. Ils peuvent décéder sans démence ou devenir dément et décéder par la suite. Cette configuration est celle d'un modèle *illness-death* dans lequel l'état initial (0) correspond au statut « non dément », l'état transitoire (1) au statut « dément » et l'état absorbant (2) à l'état « décédé » (voir figure II.1). Si les temps de décès et de démence étaient observés en temps continu, nous pourrions (comme nous l'avons vu dans le chapitre I), estimer à l'aide de modèles de Cox sur les transitions $0 \rightarrow 1$ et $0 \rightarrow 2$, d'une part, les effets des facteurs de risque de démence en censurant à droite les sujets décédés sans démence, d'autre part, les effets de ces facteurs sur le risque de décès des non déments en censurant à droite les sujets déments. Nous pourrions également estimer les effets des facteurs sur le risque de décès des déments avec un modèle de Cox sur la transition $1 \rightarrow 2$. Malheureusement, cette méthode est inadaptée ici car les temps de démence sont censurés par intervalle entre la visite de diagnostic et la visite précédente. La difficulté majeure vient des sujets qui décèdent alors qu'ils ont été vus non déments à leur dernière visite. On ne sait pas lesquels sont décédés sans démence et lesquels sont devenus déments entre leur dernière visite et leur décès (voir figure II.2).

L'approche naïve et la plus répandue en épidémiologie pour estimer les effets des facteurs de risque de démence dans un tel contexte consiste à se ramener à une situation permettant l'utilisation d'un modèle de Cox. En ce qui concerne les sujets qui décèdent sans avoir précédemment été diagnostiqués déments, ils sont censurés à droite à la date de dernière visite. Quant aux sujets diagnostiqués déments, la date de la maladie est imputée soit à la visite de diagnostic (Al Hazzouri et al., 2011), soit au milieu de

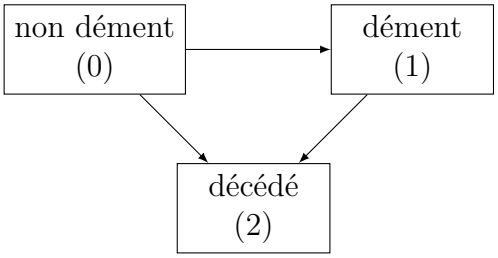
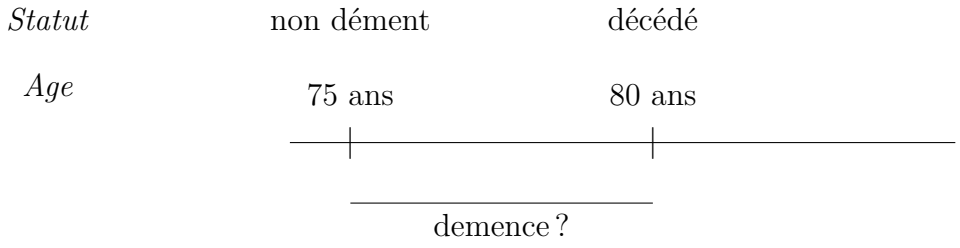


Figure II.1 – Modèle *illness-death* irréversible permettant à un sujet initialement non dément de décéder avec ou sans démence



Deux trajectoires sont possibles pour ce sujet :

Figure II.2 – Exemple d'un sujet décédé et vu non dément à sa dernière visite

l'intervalle entre la visite de diagnostic et la visite précédente (Freitag et al., 2006). L'approche que nous préconisons consiste en un modèle *illness-death* pour données censurées par intervalle.

1.3 Objectif

Dans ce travail, nous avons simulé des données de type Paquid et estimé sur la base de ces données les effets d'un facteur de risque de démence en utilisant plusieurs modèles : *i*) ne prenant en compte ni le décès ni la censure par intervalle (modèles 0 et 1) *ii*) prenant en compte uniquement la censure par intervalle (modèle 2) *iii*) prenant en compte le décès et la censure par intervalle (modèles 4 et 5). L'objectif est de comparer selon ces modèles les estimations de l'effet du facteur sur le risque de démence lorsqu'on fait varier son effet sur les risques de décès des déments et des non déments. Les données ont été simulées à partir d'un modèle *illness-death* avec des intensités de transition de base paramétrisées selon des lois de Weibull (modèle 3). Ce même modèle a été utilisé pour estimer l'effet du facteur de risque. Il sert de référence puisque nous nous attendons à ce qu'il fournisse les meilleures estimations. Nous avons également utilisé un modèle de survie qui est aussi bien spécifié (l'intensité de transition de base vers la démence, α_{01}^0, est paramétrisée selon une loi de Weibull) et qui tient compte de la censure par intervalle mais qui traite les décédés sans diagnostic de démence comme des censurés à droite (modèle 2). La comparaison des modèles 2 et 3 permet de s'apercevoir de ce qu'apporte la prise en compte du risque de décès en plus de la prise en compte seule de la censure par intervalle. Les méthodes d'estimation des autres modèles sont semi-paramétriques : aucune hypothèse n'est faite sur la distribution des intensités de transition de base. Les modèles 0 et 1 sont des modèles de Cox ne prenant en compte ni la censure par intervalle ni le risque de décéder. Ce sont les plus utilisés dans la pratique. Le modèle 4 tient compte à la fois de la censure par intervalle et du risque de décès.

Nous commençons par présenter les différents modèles qui seront utilisés. Puis, nous expliquons les simulations effectuées et tentons d'en expliquer les résultats. Ces simulations sont volontairement simples (un seul facteur de risque, effet fort de ce facteur) afin de faciliter la compréhension des mécanismes que nous voulons mettre en lumière et de dégager des messages clairs. Dans une dernière partie, nous utilisons les données de Paquid pour estimer l'effet d'un facteur avec les différents modèles afin d'observer dans quelle mesure ces mécanismes peuvent impacter les résultats obtenus

à partir de données issues d'un monde réel plus complexe.

2 Modélisations

L'âge étant le facteur de risque de démence le plus important, le choix de l'âge comme échelle de temps fait sens. De plus, en le prenant en compte de cette manière plutôt qu'en l'incluant dans le modèle en tant que variable explicative, nous évitons d'éventuels problèmes de non proportionnalité. Nous évitons également de faire une hypothèse de log-linéarité et donc de supposer que quel que soit l'âge, le rapport des risques instantanés est constant pour une augmentation de un an. Les sujets de Paquid ne sont ni devenus déments ni décédés entre leur naissance et leur âge d'entrée dans l'étude. Nous présentons donc dans cette section des modèles se généralisant à des données tronquées à gauche. Nous prenons en compte les facteurs de risque en utilisant des modèles de régressions *i*) sur la fonction de risque pour les modèles de survie *ii*) sur les fonctions d'intensité de transition pour les modèles *illness-death*. Le paramètre d'intérêt correspond à l'effet des facteurs de risque de démence et est noté β pour les modèles de survie et β_{01} pour les modèles *illness-death*.

2.1 Modèle 0 et 1 : Modèle de Cox standard

Les modèles 0 et 1 sont des modèles à risques proportionnels de Cox :

$$\lambda(t|Z) = \lambda_0(t)\mathrm{e}^{\beta^T Z} \tag{II.1}$$

où Z est le vecteur des variables explicatives, β est le vecteur des coefficients de régression, λ_0 est la fonction de risque de base.

L'estimation du paramètre de régression β se fait de façon standard par maximisation de la vraisemblance partielle de Cox dont nous rappelons l'expression :

$$L_{Cox} = \prod_{i=1}^{D} L_{Cox,i} = \prod_{i=1}^{D} \frac{\mathrm{e}^{\beta^T Z_i}}{\sum\limits_{j \in R(T_i)} \mathrm{e}^{\beta^T Z_j}} \tag{II.2}$$

où D est le nombre d'évènements (de déments) et $R(T_i)$ l'effectif à risque au temps T_i^-, c'est-à-dire juste avant que le sujet i subisse l'évènement.

N'ayant pas la connaissance des temps exacts d'évènements, c'est-à-dire des âges

de démence, ceux-ci sont imputés :

— (modèle 0) à l'âge à la visite de diagnostic ;

— (modèle 1) à l'âge au milieu de l'intervalle dont la borne supérieure correspond à la visite de diagnostic et la borne inférieure à la visite précédente.

Les sujets décédés qui ont été vus non déments à leur dernière visite sont censurés à droite à leur âge de dernière visite.

2.2 Modèle 2 : Modèle paramétrique de survie pour données censurées par intervalle

Le modèle 2 suppose une distribution de Weibull de la variable « âge de démence ». La fonction de risque de base s'écrit :

$$\lambda_0(t) = b\left(\frac{1}{c}\right)^b t^{b-1}$$

où b et c sont les paramètres de forme et d'échelle.

Le modèle à risques proportionnels de l'équation II.1 permet de prendre en compte les facteurs de risque de démence.

À la différence des modèle 0 et 1, le modèle 2 prend en compte la censure par intervalle. En effet, un sujet i diagnostiqué dément à l'âge r et qui avait l'âge l à la visite précédente contribue à la vraisemblance comme suit :

$$L_i = \frac{1}{S(a|Z)}\int_l^r S(u|Z)\lambda(u|Z)\mathrm{d}u$$

où a est l'âge d'entrée dans l'étude et $S(t|Z) = \mathrm{e}^{-\Lambda(t|Z)}$ est la fonction de survie avec Λ la fonction de risque cumulé.

Les paramètres du modèle a, b, et β sont estimés par la méthode du maximum de vraisemblance.

Comme dans les modèles 0 et 1, les sujets décédés qui ont été vus non déments à leur dernière visite sont censurés à droite à leur âge de dernière visite.

En résumé, les différences du modèle 2 par rapport aux modèles 0 et 1 sont que :

— le modèle 2 est paramétrique avec une paramétrisation de type Weibull de la fonction de risque de base λ_0 tandis que pour les modèles 0 et 1 aucune forme n'est spécifiée pour λ_0 ;

— chez les sujets diagnostiqués déments, le modèle 2 tient compte de l'incertitude

relative à l'âge d'apparition de la maladie.

2.3 Modèle 3 : Modèle paramétrique *illness-death* pour données censurées par intervalle

Le modèle 3 suppose des distributions de Weibull sous-jacentes au trois intensités de transition de base. Pour $kl = 01, 02, 12$:

$$\alpha_{0,kl}(t) = a_{kl} \left(\frac{1}{b_{kl}}\right)^{a_{kl}} t^{a_{kl}-1}$$

où a_{kl} et b_{kl} sont les paramètres de forme et d'échelle relatifs aux transitions $0 \to 1$, $0 \to 2$ et $1 \to 2$.

Des modèles à intensités de transition proportionnelles permettent d'inclure des facteurs de risque sur chacune des transitions. Pour $kl = 01, 02, 12$:

$$\alpha_{kl}(t|Z_{kl}) = \alpha_{0,kl}(t)\mathrm{e}^{\beta_{kl}^T Z_{kl}} \tag{II.3}$$

où Z_{kl} est le vecteur des variables explicatives sur la transition $k \to l$, β_{kl} le vecteur des coefficients de régression spécifique à la transition $k \to l$ et $\alpha_{0,kl}$ la fonction d'intensité de transition de base spécifique à la transition $k \to l$.

α_{01} s'interprète comme le taux d'incidence de la démence, α_{02} comme le taux de mortalité des non déments et α_{12} comme le taux de mortalité des déments. Les paramètres du modèle a_{kl}, b_{kl}, β_{kl} sont estimés par la méthode du maximum de vraisemblance.

À la différence du modèle 2, le modèle 3 prend en compte le risque de décès. Un sujet i décédé en t qui a été vu non dément à sa dernière visite, à l'âge l, n'est pas censuré à droite en l mais contribue à la vraisemblance comme suit :

$$\begin{aligned} L_i \;=\; & \frac{\mathrm{e}^{-A_{01}(l|Z)-A_{02}(l|Z)}}{\mathrm{e}^{-A_{01}(a|Z)-A_{02}(a|Z)}} \Bigg(\underbrace{\mathrm{e}^{-A_{01}(t|Z)-A_{02}(t|Z)}\alpha_{02}(t|Z)}_{(\star)} + \\ & \underbrace{\int_l^t \mathrm{e}^{-A_{01}(u|Z)-A_{02}(u|Z)}\alpha_{01}(u|Z)\frac{\mathrm{e}^{-A_{12}(t|Z)}}{\mathrm{e}^{-A_{12}(u|Z)}}\alpha_{12}(u|Z)\mathrm{d}u}_{(\star\star)} \Bigg) \end{aligned}$$

où les A_{kl} sont les intensités de transition cumulées.

Le sujet i est soit décédé sans démence à l'âge t $(\star)$, soit devenu dément à un âge

compris entre l et t puis décédé en t ($\star\star$).

2.4 Modèle 4 : Modèle semi-paramétrique *illness-death* pour données censurées par intervalle

Le modèle 4 ne diffère du modèle 3 que dans la méthode d'estimation des paramètres. Ici, les intensités de transition de base sont approximées par des combinaisons linéaires de M-splines, et les coefficients des splines ainsi que les paramètres de régression sont déterminés en maximisant la vraisemblance pénalisée. On se référera au chapitre I pour de plus amples explications sur cette méthode. La différence de ce modèle par rapport au modèle 3 est qu'il est plus flexible, ne spécifiant pas de forme paramétrique pour les intensités de transition de base.

3 Simulations

3.1 Schéma de simulation

3.1.1 Génération des données

Plusieurs scénarios correspondant à différents effets du facteur sur le risque de démence, de décès des non déments et de décès des déments ont été étudiés. Pour chaque scénario, 500 jeux de données de 2000 sujets chacun ont été générés. Les simulations ont été faites de façon à se conformer autant que possible aux données de Paquid.

L'âge d'entrée est la réalisation d'une distribution uniforme sur l'intervalle $[65, 70]$. Les âges de démence et de décès suivent des distributions de Weibull dont les paramètres a_{kl} et b_{kl} *i*) sont proches des estimations $\hat{a}_{kl}$ et $\hat{b}_{kl}$ obtenues avec le modèle 3 sans variables explicatives sur les données de Paquid *ii*) sont tels que $\alpha_{12,0}(t) > \alpha_{02,0}(t) > \alpha_{01,0}(t), \forall t$. Une variable explicative binaire dont le taux d'exposition est de 40% est introduite selon les modèles de régression de l'équation II.3. Les effets β_{01}, β_{02}, β_{12} sont différents selon les scénarios :

— (scénario 1) effet sur le risque de démence α_{01} seulement ;

— (scénario 2) effet sur les risques de décès α_{02} et α_{12} seulement ;

— (scénario 3-6) effets à la fois sur le risque de démence et sur les risques de décès.

Nous avons choisi des valeurs de β_{01}, β_{02}, β_{12} donnant lieu à des scénarios assez simples pour faciliter l'interprétation des résultats mais qui restent pertinentes d'un point de vue épidémiologique.

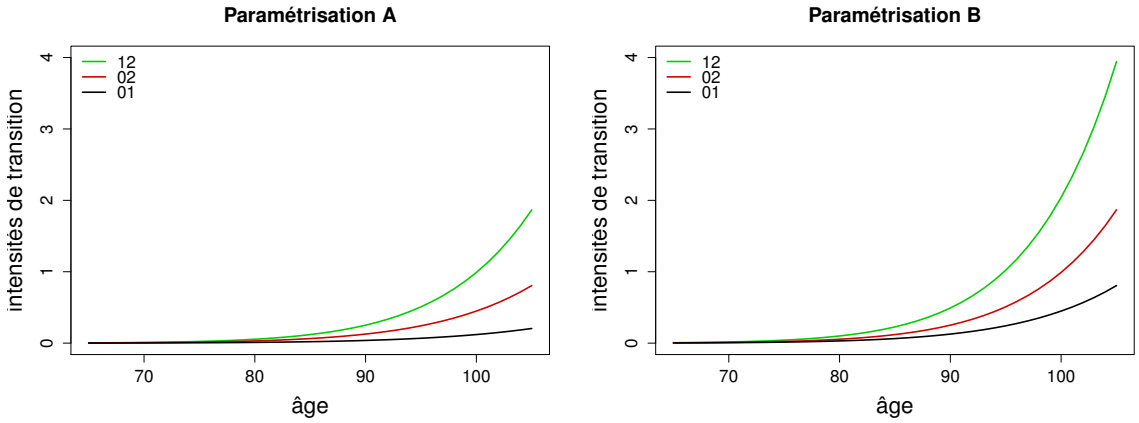


Figure II.3 – Courbes des intensités de transition de base correspondant aux lois de Weibull choisies pour les paramétrisations A et B. Les paramètres de A (risques moyens) sont : $(b_{01}, c_{01}) = (12, 100)$; $(b_{02}, c_{02}) = (13, 91)$; $(b_{12}, c_{12}) = (14, 87)$. Les paramètres de B (risques forts) sont : $(b_{01}, c_{01}) = (13, 91)$; $(b_{02}, c_{02}) = (14, 87)$; $(b_{12}, c_{12}) = (14.5, 83)$.

Un âge de censure à droite ainsi que des âges de visite ont été générés selon une distribution uniforme. Le temps entre deux visites varie de 2 à 3 ans et est de 2.5 ans en moyenne.

En plus de faire varier les effets du facteur sur les trois intensités de transition dans les différents scénarios, nous avons fait varier le nombre de cas de démences et de décès survenant au cours du suivi à travers deux paramétrisations différentes des intensités de transition de base. La figure II.3 indique les paramètres des lois de Weibull choisis ainsi que les courbes des intensités de transition de base correspondantes.

— (Paramétrisation A) Taux moyens de démence et de décès
  Elle correspond à la partie gauche du tableau II.1 des résultats. On obtient en moyenne 50% de décès et 11% de démence. Bien sûr, ces pourcentages sont légèrement modifiés selon les différents scénarios.

— (Paramétrisation B) Taux forts de démence et de décès
  Elle correspond à la partie droite du tableau II.1 des résultats. Les paramètres des lois de Weibull générant les intensités de transition de base ont été modifiés de façon à ce que le nombre de démences et de décès soit plus grand. On obtient en moyenne 76% de décès et 25% de démences.

Remarquons que les pourcentages de la paramétrisation B sont semblables à ceux de Paquid avec un suivi de 20 ans. Le pourcentage de démences non diagnostiquées parce que le décès est survenu avant la visite de diagnostic est à peu près identique dans les deux paramétrisations (15% et 16% des déments).

3.1.2 Statistiques calculées

Chacun des 500 jeux de données a été analysé avec les modèles 0-4. Pour chaque modèle, ont été calculés :

- la moyenne empirique des estimations : $\overline{\hat{\beta}_{01}} = \frac{1}{n}\sum_{i=1}^{500}\beta_{01,i}$;
- le biais relatif $\frac{\overline{\hat{\beta}_{01}} - \beta_{01}}{\beta_{01}} \times 100$ où β_{01} est le « vrai » effet sur $0 \rightarrow 1$: un biais relatif positif indique une surestimation de la valeur absolue de l'effet (surestimation de β_{01} lorsque $\beta_{01} > 0$, sous-estimation de β_{01} lorsque $\beta_{01} < 0$) tandis qu'un biais relatif négatif indique une sous-estimation de la valeur absolue de l'effet (sous-estimation de β_{01} lorsque $\beta_{01} > 0$, surestimation de β_{01} lorsque $\beta_{01} < 0$) ;
- la racine carrée de l'erreur quadratique moyenne (RMSE) : $\sqrt{\text{biais}(\hat{\beta}_{01})^2 + var(\hat{\beta}_{01})}$ qui nous renseigne sur la précision de l'estimation ;
- le taux de couverture, c'est-à-dire la proportion d'intervalles de confiance à 95%, $[\hat{\beta}_{01} \pm 1.96 \times s(\hat{\beta}_{01})]$, qui contiennent la vraie valeur β_{01}. Un « bon » taux de couverture doit être proche de 95% ; il s'en écarte si l'estimation est biaisée et/ou si l'écart-type $s(\hat{\beta}_{01})$ est sous-estimé ou surestimé.

3.2 Résultats

Les principaux résultats des simulations sont résumés dans le tableau II.1. Les commentaires des différents scénarios étant similaires pour les deux paramétrisations, nous nous concentrons dans un premier temps sur la première qui correspond à des taux moyens de démence et de décès (partie gauche du tableau II.1). Pour plus de clarté, les résultats du modèle 0 ne sont pas présents et sont commentés dans un deuxième temps à l'aide du tableau II.2. Enfin, nous examinons l'impact d'une augmentation des intensités de transition de base sur les résultats (comparaison de la partie gauche (paramétrisation A) et de la partie droite (paramétrisation B) du tableau).

Scénario 1

Le premier scénario correspond à une situation avec un facteur de risque de démence qui n'a aucun effet sur le décès. Tous les modèles donnent une estimation satisfaisante.

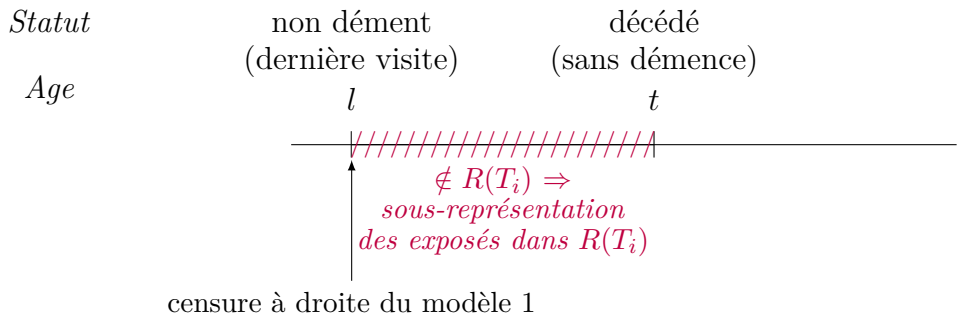
Scénario 2

À l'inverse du scénario 1, le scénario 2 correspond à une situation où un facteur de risque de décès (de même effet pour les déments et les non déments) n'a aucun effet sur le risque de démence. La colonne du biais relatif n'est pas remplie puisque le dénominateur β_{01} serait nul. Les estimations des deux modèles de survie sont plus biaisées que celles des deux modèles *illness-death*. De même, les deux RMSE des modèles de survie sont plus grandes que celles des modèles *illness-death*. Notons que comme dans ce scénario l'effet du facteur sur la démence est nul ($\beta_{01} = 0$), la quantité $1 - \frac{\text{taux de couverture}}{100}$ correspond à l'erreur de type I du test de Wald. Une erreur de type I plus grande que 5% indique une tendance à rejeter l'hypothèse nulle « $H_0 : \beta_{01} = 0$ » trop souvent. Dans ce scénario, le risque de conclure à tort à un effet significatif du facteur est légèrement plus grand en utilisant un modèle de survie qu'en utilisant un modèle *illness-death*.

Scénario 3

Dans le scénario 3, l'exposition au facteur augmente le risque de démence ($\beta_{01} = 0.5$) et de décès avec un effet beaucoup plus fort sur les non déments ($\beta_{02} = 2$) que sur les déments ($\beta_{12} = 0.5$). Les modèles 1 et 2 de survie fournissent des estimations biaisées de β_{01} (biais relatifs de 26.1 et 27.3 contre -0.4 et -4.2 pour les modèles *illness-death* 3 et 4), des RMSE grandes (239 et 243 contre 196 et 195), et des taux de couverture qui s'écartent de 95 % (90.6 et 89.4 contre 95 et 96.2).

Regardons de plus près ce qui se passe avec l'estimation du modèle de Cox pour tenter de comprendre de façon intuitive les mécanismes menant à cette surestimation de l'effet du facteur sur le risque de démence. Le facteur ayant un effet très important sur le décès des non déments, une proportion importante de sujets exposés (tels que $z = 1$) décède sans démence. Considérons un tel sujet. Soit t son âge de décès et l son âge de dernière visite. Dans un schéma classique d'observation en temps continu de l'âge de démence (sans censure par intervalle), ce sujet serait censuré à droite en t. Dans notre cas, nous savons que le sujet était encore non dément en l mais le fait que le sujet ne soit pas devenu dément entre l et t nous est inconnu. Donc avec le modèle 1, nous censurons à droite ce sujet en l et ce sujet, au lieu de contribuer à l'effectif à risque jusqu'à t, n'y contribue que jusqu'à l. Pour toutes les contributions à la vraisemblance $L_{Cox,i}$ des sujets déments entre l et t (plus exactement, dont l'âge de démence a été imputé entre l et t), le sujet ne fait pas partie de l'effectif à risque $R(T_i)$ (voir équation II.2) :

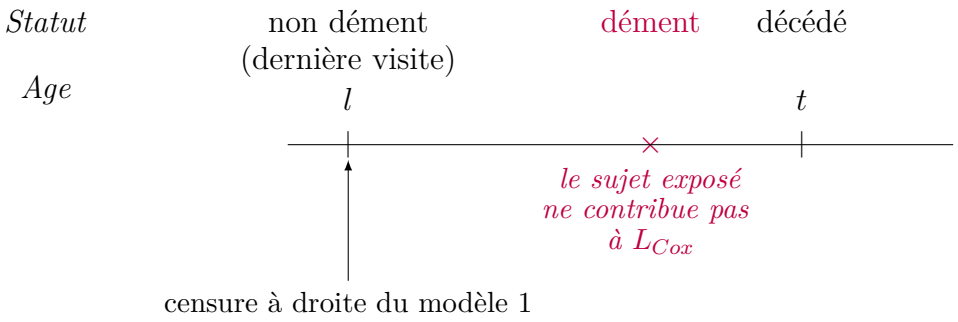


À cause d'une censure à droite prématurée des décédés sans démence, les exposés ont donc tendance à être sous-représentés dans l'effectif à risque (par symétrie, les non exposés sont surreprésentés). Cela a pour conséquence de gonfler la probabilité de devenir dément pour un exposé et inversement d'amoindrir celle d'un non exposé. L'effet du facteur sur le risque de démence est donc surestimé ($\hat{\beta}_{01} > \beta_{01}$). Notons que dans un scénario similaire mais avec un effet protecteur du facteur sur le risque de démence ($\beta_{01} < 0$), la surestimation de β_{01} correspondrait en fait à une sous-estimation de l'effet protecteur (cela a été vérifié sur des simulations non présentées ici).

Scénario 4

Dans le scénario 4, l'exposition au facteur augmente le risque de démence ($\beta_{01} = 0.5$) et de décès avec cette fois un effet beaucoup plus fort sur les déments ($\beta_{12} = 2$) que sur les non déments ($\beta_{02} = 0.5$). Les modèles de survie 1 et 2 fournissent des estimations très biaisées de β_{01} (biais relatifs de -85.4 et -83.8 contre -0.8 et -1.3 pour les modèles *illness-death* 3 et 4), et donc des RMSE très grandes (456 et 449 contre 175 et 176) et des taux de couverture extrêmement loin de 95% (25.0 et 27.2 contre 96.4 et 95.6).

Ici aussi, tentons de comprendre intuitivement sur le modèle 1 de Cox les mécanismes occasionnant ces biais. Le facteur ayant un effet très important sur le décès des déments, les sujets exposés, après être devenus déments, ont tendance à décéder rapidement. Parmi les décédés sans démence à leur dernière visite, trop de sujets exposés devenus déments avant leur décès ont été « manqués » par le modèle. Considérons un sujet exposé, vu non dément à son âge de dernière visite l, décédé en t et devenu dément entre l et t. Alors ce sujet ne contribuera pas à la vraisemblance :



Les exposés qui deviennent déments sont à haut risque de décéder et sont donc « manqués » plus souvent que les autres. *A posteriori*, cela veut dire qu'on censure à droite des sujets exposés qui ne sont pas représentatifs de ceux toujours à risque puisqu'il ont un risque plus élevé de devenir dément que les autres. On comprend bien que cela induit une sous-estimation de l'effet du facteur sur le risque de démence. Dans un scénario similaire mais avec un effet protecteur du facteur sur le risque de démence ($\beta_{01} < 0$), la sous-estimation de β_{01} correspondrait à une surestimation de l'effet protecteur.

Scénario 5

Rappelons que ce sont les sujets qui décèdent alors qu'ils étaient non déments à leur dernière visite qui représentent la principale source de biais lorsqu'on utilise un modèle de survie pour estimer l'effet d'un facteur de risque de démence. Parmi eux, il y en a qui décèdent sans démence. Lorsque le facteur a un effet positif important sur le risque de décès sans démence ($\beta_{02} >> 0$), les sujets exposés auraient tendance à être sous-représentés dans les effectifs à risque et l'effet du facteur sur le risque de démence surestimé (scénario 3). Les autres deviennent déments entre leur dernière visite et leur décès. Lorsque le facteur a un effet positif important sur le risque de décès des déments ($\beta_{12} >> 0$), les exposés auraient plus tendance à être « manqués » et ne contribueraient pas assez à la vraisemblance, entraînant une sous-estimation de l'effet du facteur sur le risque de démence.

Dans le scénario 5, l'exposition au facteur augmente le risque de démence de la même façon que dans les scénarios 3 et 4 ($\beta_{01} = 0.5$) et augmente fortement les risques de décéder des non déments *et* des déments ($\beta_{02} = \beta_{12} = 2$). De cette façon les deux phénomènes de surestimation et de sous-estimation de β_{01} agissent simultanément. On voit que c'est celui de sous-estimation qui l'emporte (biais relatifs des modèles de survie de -22.0 et -20.8 contre 1.2 et -0.5 pour les modèles *illness-death*). Ce résultat

était prévisible car les biais étaient plus importants dans le scénario 4 (-85.4 et -83.8) que dans le scénario 3 (26.1 et 27.3). Ce scénario 5 nous amène à penser qu'il est possible dans la réalité que ces phénomènes de surestimation et de sous-estimation se compensent de façon à ce que le biais de $\hat{\beta}_{01}$ « s'annule ». Cependant, il paraît difficile de prévoir dans quelle mesure ces deux phénomènes vont agir et donc s'il y aura un biais.

Scénario 6

Nous avons vu dans les scénarios 3 et 4 qu'un fort effet positif du facteur sur le risque de décéder des non déments conduit à une surestimation de β_{01} et qu'un fort effet positif du facteur sur le risque de décéder des déments conduit à une sous-estimation de β_{01}. En considérant des effets négatifs sur le décès et en déroulant des raisonnements similaires, nous aboutissons à l'assertion : un fort effet négatif du facteur sur le risque de décéder des non déments conduit à une sous-estimation de β_{01} tandis qu'un fort effet négatif du facteur sur le risque de décéder des déments conduit à une surestimation de β_{01}. Des effets aussi forts que ceux des scénarios 3 et 4 ($\beta_{01} = 2$ correspond à un *hazard ratio* de $e^{\beta_{01}} \simeq 7.4$) sont rares en pratique. Les effets du scénario 6 sont plus raisonnables. Ils correspondent à des *hazard ratio* de $e^{0.5} \simeq 1.6$ et $e^{-0.5} \simeq 0.6$. Mais à l'inverse des autres scénarios, les effets du facteur sur le risque de décéder des déments et des non déments sont contraires ($\beta_{02} = -0.5 < 0$ et $\beta_{12} = 0.5 > 0$) allant tous les deux dans le sens d'une sous-estimation de β_{01}. L'effet du facteur sur le risque de démence est négatif (facteur protecteur). De cette façon, ce facteur fictif agit dans le même sens que le facteur du niveau d'étude, primordial dans une étude sur le risque de démence : il protège contre le risque de démence et de décès des non déments, il augmente le risque de décès des déments (voir la section 4).

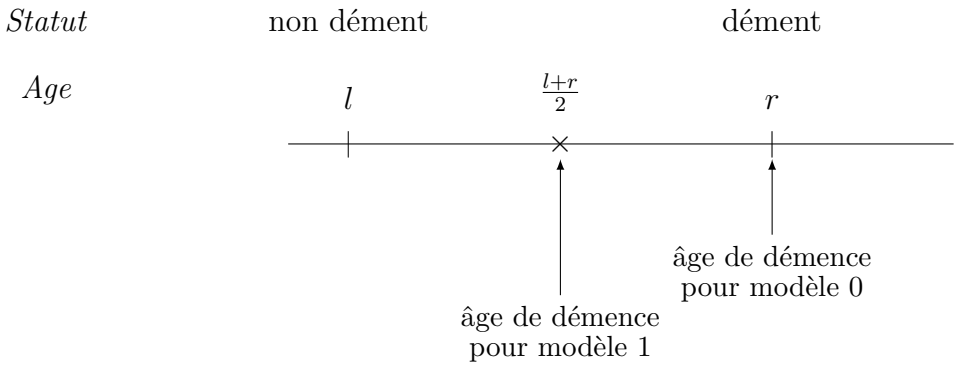
Nous constatons (tableau II.1) que des effets qui ne sont pas excessivement élevés sur les risques de décès mais qui sont opposés (et donc vont dans le même sens d'une sous-estimation ou d'une surestimation de β_{01}) conduisent à une estimation biaisée de l'effet du facteur sur le risque de démence. β_{01} est sous-estimée par les modèles de survie (biais relatifs de 21.8 et 24.3 contre 0.1 et 6.8 pour les modèles *illness-death*) ce qui revient à dire que l'effet protecteur du facteur sur le risque de démence est surestimé.

					Paramétrisation A des $\alpha_{kl,0}$				Paramétrisation B des $\alpha_{kl,0}$			
Scénario	β_{01}	β_{02}	β_{12}	Modèle	$\overline{\hat{\beta}_{01}}$	biais relatif	RMSE ×1000	taux (%) de couverture	$\overline{\hat{\beta}_{01}}$	biais relatif	RMSE ×1000	taux (%) de couverture
1	1	0	0	1	1.01	1.1	121	95.8	0.99	-0.6	88	92.4
				2	1.01	1.5	122	95.8	1.00	0.1	88	93.4
				3	1.01	1.0	121	95.4	0.99	-0.8	89	93.8
				4	1.03	3.3	126	95.8	1.01	-0.7	89	93.4
2	0	1	1	1	-0.07	–	195	94.6	-0.07	–	151	92.6
				2	-0.06	–	192	94.4	-0.06	–	147	92.8
				3	-0.00	–	189	96.0	-0.01	–	136	93.6
				4	-0.01	–	189	95.6	-0.03	–	139	93.2
3	0.5	2	0.5	1	0.63	26.1	239	90.6	0.66	31.6	214	79.0
				2	0.64	27.3	243	89.4	0.66	31.7	213	78.8
				3	0.50	-0.4	196	95.0	0.50	-0.3	142	95.0
				4	0.48	-4.2	195	96.2	0.43	-12.9	154	93.6
4	0.5	0.5	2	1	0.07	-85.4	456	25.0	0.06	-88.4	458	3.2
				2	0.08	-83.8	449	27.2	0.07	-86.8	450	3.8
				3	0.50	-0.8	175	96.4	0.55	9.8	140	91.0
				4	0.49	-1.3	176	95.6	0.51	3.0	132	93.8
5	0.5	2	2	1	0.39	-22.0	243	93.6	0.35	-29.4	215	85.4
				2	0.40	-20.8	238	93.2	0.36	-28.8	212	86.4
				3	0.51	1.2	229	95.2	0.50	0.1	166	95.8
				4	0.50	-0.5	231	95.6	0.44	-11.8	174	94.4
6	-0.5	-0.5	0.5	1	-0.61	21.8	210	90.0	-0.60	20.6	150	84.4
				2	-0.62	24.3	217	88.4	-0.63	26.4	172	79
				3	-0.50	0.1	180	95.0	-0.46	-7.6	118	94.0
				4	-0.47	6.8	183	93.2	-0.45	-10.2	123	93.0

Tableau II.1 – Résultats des simulations pour le modèle 1 (Cox avec imputation de l'âge de démence au milieu de l'intervalle de censure), le modèle 2 (survie avec censure par intervalle, paramétrisation Weibull), le modèle 3 (*illness-death* avec censure par intervalle, paramétrisation Weibull) et le modèle 4 (*illness-death* avec censure par intervalle, semi-paramétrique). Pour chaque scénario, les statistiques calculées sur 500 réplications de 2000 sujets sont de gauche à droite : moyenne, biais relatif, racine carrée de l'erreur quadratique moyenne (RMSE) ×1000, taux de couverture (%).

Modèle 0

Les résultats du modèle 0 ont été retirés du tableau II.1 afin de ne pas perturber la compréhension de l'origine des biais des modèles 1 et 2 (voir en particulier les commentaires des scénarios 3 et 4). En effet, on pressent que le fait d'imputer systématiquement l'âge d'apparition de la démence à l'âge à la visite de diagnostic va induire un autre genre de biais, qui va toujours aller dans le même sens. Il est apparu dans les résultats que ce biais est assez important. Il peut d'une certaine manière s'ajouter aux biais précédemment expliqués, ou les compenser. Le tableau II.2 réunit les résultats propres aux deux modèles de Cox afin de tenter de comprendre ce qui se joue dans le modèle 0 *par rapport au modèle 1*. La principale différence entre les deux modèles est que la contribution à la vraisemblance d'un sujet dément en r (modèle 0) n'est pas la même que celle d'un sujet dément en $\frac{l+r}{2}$ (modèle 1).



Cette différence concerne le dénominateur de l'équation II.2 puisque l'effectif à risque au temps r n'est pas le même que celui au temps $\frac{l+r}{2}$. En effet, dans l'intervalle $[\frac{l+r}{2}, r]$, on « perd » des sujets. Si ceux-ci n'étaient que de « vrais » censurés à droite, la proportion d'exposés et de non exposés ne varierait pas entre $\frac{l+r}{2}$ et r. Mais il y a aussi des décédés, et surtout des décédés sans démence (voir la répartition des pourcentages de démences et de décès dans les paramétrisations A et B). Dans les scénarios 2, 3, 4 et 5, ces décédés sont plus exposés que les autres ($\beta_{01} = 1, 2, 0.5, 2$). Les exposés seraient donc moins bien représentés dans l'effectif à risque du modèle 0 que dans celui du modèle 1, conduisant à une estimation de β_{01} plus grande dans le modèle 0 que dans le modèle 1. Du reste, on remarque que *i*) les différences entre le biais relatif du modèle 0 et celui du modèle 1 sont les plus grandes dans les scénarios 3 et 5, lorsque l'effet du facteur sur le risque de décès des non déments est le plus grand ($\beta_{02} = 2$) *ii*) le modèle 0 est aussi bon que le modèle 1 en ce qui concerne l'estimation de β_{01} lorsque le facteur n'a pas d'effet sur les risques de décéder ($\beta_{02} = \beta_{12} = 0$), en particulier celui des non

déments (scénario 1). Dans le scénario 6, les décédés sont moins exposés que les autres ($\beta_{02} = -0.5$). Les exposés sont donc plus représentés dans l'effectif à risque du modèle 0 et celui-ci aurait donc tendance à fournir une estimation de β_{01} plus petite que celle du modèle 1. La différence entre les biais relatifs des modèles 0 et 1 est cependant plus petite que dans le scénario 4 dans lequel $\beta_{02} = 0.5$. Cela pourrait être dû *i*) au taux d'exposition du facteur de 40% : enlever des exposés dans une population qui en comptait 40% à la base aurait plus d'impact que d'en ajouter *ii*) au fait que parmi les décédés qu'on « perd » entre $\frac{l+r}{2}$ et r, il y a probablement quelques déments non diagnostiqués qui sont plus exposés que les autres au facteur ($\beta_{12} = 0.5$).

Augmentation des risques de base

Regardons maintenant le tableau II.1 (on peut également regarder le tableau II.2) pour voir ce qui se passe lorsqu'on augmente les risques de base de démence et de décès (en choisissant les paramètres des lois de Weibull de façon à ce que les courbes des intensités de transition de base $\alpha_{0,01}(t)$, $\alpha_{0,02}(t)$, $\alpha_{0,12}(t)$ soient au-dessus de celles de la paramétrisation A, $\forall t$). Comme on pouvait s'y attendre, les RMSE sont globalement plus petites dans la partie droite du tableau grâce à une plus grande précision des estimations, ou plus exactement, à une plus petite variance des estimations. En revanche, les biais pour les scénarios 3, 4, 5 et 6 (avec un effet du facteur sur les trois transitions) sont un peu plus importants. La plus grande différence entre les deux paramétrisations concerne les taux de couverture qui se sont fortement dégradés pour les modèles de survie. Ces dégradations sont dues à des biais plus important mais probablement aussi à une sous-estimation de l'écart-type $s(\hat{\beta}_{01})$.

Dans les deux paramétrisations, ce sont les modifications des lois de Weibull qui ont eu pour conséquence de modifier le nombre de démences et de décès. Cependant, plusieurs paramètres n'ont pu être contrôlés, par exemple la proportion des décédés par rapport aux déments, et ont pu jouer un rôle dans les différences observées dans les parties gauche et droite du tableau. Peut-être aurait-il était aussi intéressant de faire varier le nombre de démences et de décès en modifiant uniquement la quantité d'information disponible, par exemple, en modifiant la durée du suivi.

Remarques générales

— Le modèle 3 est performant vis-à-vis de l'estimation de β_{01} sur les simulations présentées mais il faut garder à l'esprit que celui-ci est bien spécifié puisque les

données ont aussi été simulées selon le modèle 3. On peut cependant penser qu'il est raisonnable dans beaucoup d'applications de faire l'hypothèse d'une distribution de Weibull pour chaque intensité de transition de base.

— Le modèle 4 est plus flexible que le modèle 3 dans le sens qu'aucune loi paramétrique n'est spécifiée pour les intensités de transition de base. Cependant, il nécessite de choisir un nombre de nœuds (et de les placer) et, ce qui est plus délicat, de choisir un paramètre de lissage qui réalise un compromis entre fidélité aux données et régularité. Dans la pratique, ce choix est fait en utilisant des techniques de validation croisée ou par tâtonnements, en faisant tourner un même modèle plusieurs fois avec des paramètres de lissages différents. Ces deux méthodes ne sont pas envisageables dans un travail de simulations. Des paramètres proches de ceux que nous utilisons pour Paquid ont été choisis au préalable et tous les jeux de données générés ont été analysés en utilisant ces paramètres. Ils étaient probablement plus ou moins bien adaptés selon les cas, et peut-être est-ce pour cette raison que le modèle 4 est parfois moins performant qu'on aurait pu l'espérer, par rapport au modèle 3.

— Il est important de remarquer que le modèle 2, qui spécifie une distribution paramétrique du risque de démence de base et qui prend en compte la censure par intervalle, n'est pas plus performant que le modèle 1 de Cox. Cela montre que lorsqu'on s'intéresse à l'effet d'un facteur de risque de démence dans un contexte *illness-death*, la censure par intervalle ne pose pas tant problème chez les sujets déments pour qui l'on ne connaît pas exactement l'âge d'apparition de la maladie mais chez les sujets décédés sans diagnostic de démence pour qui l'on ne sait pas s'ils sont devenus déments ou pas entre leur dernière visite et leur décès.

— D'autres simulations qui ne sont pas présentées ici ont montré, de façon attendue, que les modèles de survie font d'autant plus défaut que l'on augmente la durée entre les visites (*i.e.* la taille des intervalles de censure). Dans Paquid, cette durée est en moyenne plus grande que dans nos simulations car dans les faits, nous ne connaissons pas le statut des sujets à tous les temps de visite préalablement planifiés car il arrive que les sujets « manquent » des visites.

					Paramétrisation A des $\alpha_{kl,0}$				Paramétrisation B des $\alpha_{kl,0}$			
Scénario	β_{01}	β_{02}	β_{12}	Modèle	$\overline{\hat{\beta}_{01}}$	biais relatif	RMSE ×1000	taux (%) de couverture	$\overline{\hat{\beta}_{01}}$	biais relatif	RMSE ×1000	taux (%) de couverture
1	1	0	0	0	1.01	1.1	121	95.6	0.99	-0.6	89	93.0
				1	1.01	1.1	121	95.8	0.99	-0.6	88	92.4
2	0	1	1	0	0.03	–	187	93.8	0.06	–	147	91.8
				1	-0.07	–	195	94.6	-0.07	–	151	92.6
3	0.5	2	0.5	0	0.83	65.8	388	65.0	0.90	80.1	42	21.0
				1	0.63	26.1	239	90.6	0.66	31.6	214	79.0
4	0.5	0.5	2	0	0.14	-72.1	396	40.6	0.17	-66.1	352	20.4
				1	0.07	-85.4	456	25.0	0.06	-88.4	458	3.2
5	0.5	2	2	0	0.59	17.5	241	92.8	0.60	20.8	194	90.0
				1	0.39	-22.0	243	93.6	0.35	-29.4	215	85.4
6	-0.5	-0.5	0.5	0	-0.64	27.5	226	87.2	-0.63	26.8	172	78.0
				1	-0.61	21.8	210	90.0	-0.60	20.6	150	84.4

Tableau II.2 – Résultats des simulations pour les modèles 0 (Cox avec imputation de l'âge de démence à l'âge à la visite de diagnostic) et 1 (Cox avec imputation de l'âge de démence au milieu de l'intervalle de censure). Pour chaque scénario, les statistiques calculées sur 500 réplications de 2000 sujets sont de gauche à droite : moyenne, biais relatif, racine carrée de l'erreur quadratique moyenne (RMSE) ×1000, taux de couverture (%).

	Total ($n = 3675$)	Hommes ($n = 1542$)	Femmes ($n = 2133$)
âge d'entrée (moy,s)	75.3 (6.8)	74.6 (6.4)	75.7 (7.0)
démences observées (n, %)	832 (22.6)	247 (16.0)	585 (27.4)
âge de diagnostic (moy,s)	86.3 (5.8)	84.5 (5.9)	87.0 (5.6)
décès (n, %)	2937 (79.9)	1311 (85.0)	1626 (76.2)
âge de décès (moy,s)	86.3 (6.9)	84.6 (6.8)	87.7 (6.6)
cep (n, %)	2396 (65.2)	1081 (70.1)	1315 (61.6)

Tableau II.3 – Description des sujets de Paquid par sexe (n, effectif ; moy, moyenne ; s, écart-type)

4 Application

Nous avons vu comment se comportaient les modèles 0-4 sur simulations au regard de l'estimation de l'effet d'un facteur sur le risque de démence. Dans cette section, nous allons voir comment ils se comportent sur données réelles pour estimer l'effet du niveau d'études sur le risque de démence. Nous avons choisi pour cette application le niveau d'études car c'est un facteur qui est connu pour être associé au risque de démence, et qui peut aussi être associé au risque de décès. La variable utilisé est binaire : 1 si obtention du certificat d'études primaires (cep) ; 0 sinon. L'âge, qui est fortement associé aux risques de démence et de décès, a été choisi comme temps de base. Pour chaque modèle, nous avons analysé de façon séparée les hommes et les femmes. En effet, l'hypothèse de proportionnalité ne serait pas valide pour la variable sexe car les risques de démence et de décès en fonction de l'âge semblent avoir un comportement différent chez les hommes et chez les femmes (Commenges et al., 1998; Fratiglioni et al., 1997).

Parmi les 3675 sujets de Paquid initialement non déments, 663 n'ont eu aucune visite au cours du suivi. Ceux-ci n'apportent aucune information aux modèles de survie (modèles 0, 1, 2) qui ont donc été appliqués à l'échantillon des 3012 sujets restants. En revanche, les 3675 sujets ont été utilisés pour les modèles *illness-death* (modèles 3 et 4) car pour tous les sujets sans visite de suivi, nous avions une date de décès ou une date de dernières nouvelles strictement supérieure à la date d'entrée dans la cohorte. 832 sujets ont été diagnostiqués déments dont 639 (76.8%) qui sont décédés (voir le tableau II.3 pour un descriptif succinct des sujets par sexe). Parmi les 2843 (3675-832) qui n'ont jamais été diagnostiqués déments, 2298 (80.8%) sont décédés et l'on ne sait pas s'ils sont devenus déments entre leur dernière visite et leur décès. Une description du temps écoulé entre leur dernière visite et leur décès est disponible dans

Description du temps écoulé entre la dernière visite et le décès des 1702 sujets suivis				Description du temps écoulé entre la date d'entrée dans la cohorte et le décès des 596 sujets sans visite de suivi			
min	max	moy	med	min	max	moy	med
0.015	18.92	3.09	1.92	0.017	20.91	5.97	3.51

Tableau II.4 – Description du temps écoulé entre la dernière visite et le décès des 2298 sujets décédés qui n'ont pas été diagnostiqués déments. Abréviations : min, minimum ; max, maximum ; moy, moyenne ; med, médiane.

		Hommes			Femmes		
Modèle	Transition	$\hat{\beta}$	HR	p-valeur	$\hat{\beta}$	HR	p-valeur
0	$0 \to 1$	-0.67	0.51	<0.001	-0.38	0.69	<0.001
1	$0 \to 1$	-0.74	0.48	<0.001	-0.44	0.65	<0.001
2	$0 \to 1$	-0.75	0.47	<0.001	-0.45	0.64	<0.001
3	$0 \to 1$	-0.56	0.57	<0.001	-0.42	0.66	<0.001
	$0 \to 2$	-0.29	0.75	0.001	-0.06	0.95	0.56
	$1 \to 2$	0.35	1.42	0.01	0.05	1.05	0.56
4	$0 \to 1$	-0.55	0.58	<0.001	-0.37	0.69	<0.001
	$0 \to 2$	-0.25	0.78	0.004	-0.08	0.97	0.44
	$1 \to 2$	0.34	1.40	0.02	0.09	1.09	0.33

Tableau II.5 – Estimations de l'effet du cep sur le risque de démence (modèles 0-4) et sur les risques de décès des déments et des non déments (modèles 3 et 4). Paquid suivi 20 ans.

le tableau II.4 ; elle distingue les 596 sujets sans suivi et les 1702 sujets avec suivi.

Le tableau II.5 résume les estimations des paramètres de régression associés à la variable cep et obtenues en utilisant les modèles 0 à 4.

Nous sommes dans une configuration similaire au scénario 6 des simulations : l'obtention du cep réduit le risque de démence ($\beta_{01} < 0$) et le risque de décès des non déments ($\beta_{02} < 0$) tandis qu'il augmente le risque de décès des déments ($\beta_{12} > 0$). D'après tous les modèles, l'effet du cep est significatif sur la transition $0 \to 1$ pour les hommes et pour les femmes. Les résultats sont cohérents avec les commentaires

Description des intervalles de censure			
min	max	moy	med
0.91	21.25	3.73	2.74

Tableau II.6 – Description de la durée en années des intervalles de censure (*i.e.* de la durée entre la visite de diagnostic et la visite précédente) pour les 832 sujets diagnostiqués déments. Abréviations : min, minimum ; max, maximum ; moy, moyenne ; med, médiane.

des simulations. Les modèles de survie 1 et 2 semblent sous-estimer β_{01}, *i.e.* minimiser l'effet protecteur du cep. Cette sous-estimation est manifeste chez les hommes, pour qui le cep a des effets significatifs sur les risques de décès estimés à -0.29 (modèle 3) et -0.25 (modèle 4) sur la transition $0 \rightarrow 2$ et à 0.35 (modèle 3) et 0.34 (modèle 4) sur la transition $1 \rightarrow 2$. Chez les femmes, pour qui l'effet du cep sur les risques de décès est non significatif, elle est moins évidente.

Comme pour les simulations, nous remarquons que la prise en compte de la censure par intervalle seule n'a pas d'effet sur l'estimation de β_{01} puisque les modèles 1 et 2 donnent des $\hat{\beta}_{01}$ très proches. Un modèle de survie avec une paramétrisation de type Weibull qui diffère du modèle 2 par le fait de ne pas prendre en compte la censure par intervalle (en utilisant la même imputation que le modèle 1) a aussi été utilisé. Les estimations $\hat{\beta}_{01}$ (non présentées ici) sont identiques à celle du modèle 2 au centième près.

Les résultats du modèle 0 ne vont pas, quant à eux, dans le même sens que les simulations du scénario 6. Il se pourrait que l'explication donnée ne tienne plus ici à cause d'une variation de la taille des intervalles de censure beaucoup plus importante que dans les simulations (voir tableau II.6). Le fait d'imputer l'âge de démence à la visite de diagnostic peut complètement modifier l'ordre des temps d'évènement et donc l'ordre de contribution des sujets déments. Les sujets déments dont l'intervalle de censure est grand ont tendance à contribuer plus tard que les autres, et inversement. Regardons par exemple le schéma ci-après.

L'imputation de l'âge de démence des trois sujets se fait au milieu de l'intervalle en utilisant le modèle 1 ($\times$) ou à la visite de diagnostic en utilisant le modèle 0 ($\otimes$). Sur cet exemple, l'ordre de contribution des trois sujets serait A, B, C pour le modèle 1, et C, B, A pour le modèle 0. Finalement, au vu des simulations et de l'application, une explication claire des biais engendrés par l'imputation de l'apparition de la démence à la visite de diagnostic ne ressort pas. Le sens et l'importance du biais paraissent difficilement prévisibles.

5 Conclusion

Nous nous sommes placés dans un contexte *illness-death* où les temps associés à la transition $0 \to 1$ (âges de démence) sont censurés par intervalle. Les sujets qui transitent vers l'état 2 (décédé) alors qu'ils étaient en l'état 0 (non dément) la dernière fois qu'ils ont été vus ont pu le faire directement (décéder sans démence) ou en passant par l'état 1 (développer une démence avant de mourir). Lorsqu'on s'intéresse aux facteurs de risque de démence, une approche consiste à censurer à droite les sujets dont le statut dément ou non dément est inconnu à la date à laquelle ils ont été vus pour la dernière fois, afin de rendre possible l'utilisation d'un modèle de survie classique, souvent un modèle de Cox avec estimation des effets des facteurs par maximum de vraisemblance partielle. Les âges de démence des sujets diagnostiqués déments sont alors imputés, le plus souvent au milieu de l'intervalle de censure (entre la visite de diagnostic et la visite précédente) comme dans le modèle 1, parfois à la borne droite de l'intervalle de censure (âge à la visite de diagnostic) comme dans le modèle 0. On peut aussi utiliser un modèle de survie qui prend en compte la censure par intervalle, comme le modèle 3. Nous avons vu grâce à un travail de simulations que cette approche peut induire des biais dans l'estimation d'un facteur de risque de démence lorsque ce facteur est aussi associé au risque de décès des non déments et/ou des déments.

En particulier, en utilisant les modèles 1 et 2 :

- Un fort effet positif du facteur sur le risque de décès des non déments ($\beta_{02} >> 0$) peut entraîner une surestimation de l'effet du facteur sur le risque de démence ($\hat{\beta}_{01} > \beta_{01}$). Inversement, un fort effet négatif sur la transition $0 \to 2$ pourrait entraîner une sous-estimation de β_{01}.
- Un fort effet positif du facteur sur le risque de décès des déments ($\beta_{12} >> 0$) peut entraîner une sous-estimation de l'effet du facteur sur le risque de démence ($\hat{\beta}_{01} < \beta_{01}$). Inversement, un fort effet négatif sur la transition $1 \to 2$ pourrait entraîner une surestimation de β_{01}.
- Un effet négatif du facteur sur le risque de décès des non déments ($\beta_{02} < 0$) associé à un effet positif sur le risque de décès des déments ($\beta_{12} > 0$) peut entraîner une sous-estimation de β_{01}. Inversement, un effet positif sur la transition $0 \to 2$ associé à un effet négatif sur la transition $1 \to 2$ pourrait entraîner une surestimation de β_{01}.

Ces observations ont étés confirmées par une application sur le jeu de données Paquid avec le facteur du niveau d'études. Une application plus poussée dans laquelle

plusieurs facteurs sont étudiés a été faite dans Leffondré et al. (2013) et les résultats vont dans le même sens. Notons que dans une situation où le facteur n'a pas d'effet sur le risque de démence ($\beta_{01} = 0$) mais en a sur le risque de décès, un biais dans l'estimation de β_{01} pourrait conduire à conclure à tort que cet effet est significatif.

L'imputation de l'âge de démence à la visite de diagnostic du modèle 0 entraîne d'autres sources de biais plus difficilement identifiables que les précédentes. En utilisant un tel modèle, il semble difficile de prévoir si $\hat{\beta}_{01}$ sera biaisé, et le cas échéant, quel sera l'ordre de grandeur et le sens (positif ou négatif) du biais.

La prise en compte de la censure par intervalle au sein d'un modèle de survie (modèle 2) ou l'imputation de l'âge de démence au milieu de l'intervalle de censure (modèle 1) donnent des résultats similaires en ce qui concerne l'estimation de β_{01}. Dans un « vrai » contexte d'analyse de survie (*all-cause mortality*), la littérature mentionne qu'une telle imputation peut induire des biais dans l'effet du facteur de risque et peut sous-estimer son écart-type (Law and Brookmeyer, 1992) lorsque les données sont doublement censurées par intervalle. Nous aurions pu penser que nous allions obtenir des résultats similaires avec des données « simplement » censurées par intervalle. Mais dans notre contexte *illness-death*, la non prise en compte de la censure par intervalle qui concerne les sujets déments semble peser très peu par rapport à la non prise en compte des deux trajectoires possibles pour les sujets décédés sans diagnostic de démence ($0 \rightarrow 1 \rightarrow 2$ ou $0 \rightarrow 2$).

Ce travail vient compléter celui de Joly et al. (2002). Ils ont montré que lorsqu'on ignore le risque de décéder des déments en censurant à droite tous les décédés sans diagnostic de démence, le taux d'incidence de la démence, c'est-à-dire l'estimation de l'intensité de transition vers la démence α_{01}, est sous-estimée. Le biais est nul lorsque les intensités de transition vers le décès α_{02} et α_{12} sont égales. Or, le taux de mortalité des déments est plus élevé que celui des non déments ($\alpha_{12} > \alpha_{02}$) ce qui induit un biais. En fait, plus grand est le différentiel entre les deux taux de mortalité, plus grand est le biais lorsqu'on estime α_{01}. Nous avons montré que l'estimation d'un facteur de risque de démence peut être biaisée lorsque celui-ci a aussi un effet sur le risque de décès.

En conclusion, lorsque l'on étudie la démence en ignorant la censure par intervalle et le risque de décès des déments, l'incidence de la démence est sous-estimée et les effets des facteurs de risque de démence peuvent être biaisés.

Chapitre III

Prévisions dans un modèle *illness-death*

Nous avons vu dans les chapitres précédents que dans le contexte de l'étude de la démence, le modèle qui paraît le plus approprié est un modèle multi-état particulier, le modèle *illness-death* qui prend en compte les données censurées par intervalle. Nous y avons aussi vu comment estimer les intensités de transition et les effets de facteurs de risque. Nous exposons dans ce chapitre d'autres quantités, pertinentes d'un point de vue clinique ou épidémiologique, dont l'estimation est relativement directe et qui permettent de faire des prévisions pour des sujets ayant des caractéristiques données.

Introduction

Souvent, les fonctions d'intérêt d'un modèle multi-état sont les intensités de transition. Cependant, d'autres quantités ayant une interprétation plus naturelle et permettant d'accéder à d'autres types d'information ont suscité l'intérêt. Par exemple, les probabilités de transition ont une interprétation plus naturelle que les intensités de transition qui sont des probabilités de transition instantanées. Pour les modèles multi-états avec un schéma d'observations en temps continu, Aalen and Johansen (1978) en ont proposé un estimateur non paramétrique en étendant l'estimateur de Kaplan-Meier aux modèles multi-états Markoviens non homogènes. Meira-Machado et al. (2006) ont quant à eux proposé un estimateur non paramétrique des probabilités de transition pour des modèles non markoviens. Dans les modèles multi-états avec un schéma d'observation en temps discret, des estimations des probabilités de transition peuvent également être calculées en considérant un processus de Markov homogène ou homogène par période (Kalbfleisch and Lawless, 1985) puisqu'une relation simple lie alors les probabilités de transition aux intensités de transition (voir équation I.5).

Concernant les modèles à risques concurrents, des auteurs se sont intéressés aux

fonctions dites d'incidence cumulée et à des méthodes de régression des effets des facteurs non pas sur le risque instantané mais directement sur ces incidences cumulées. Fine and Gray (1999) ont proposé un modèle à risques proportionnels sur les fonctions d'incidence cumulée, puis d'autres formes de régression ont été proposées pour faire le lien entre les variables explicatives et les fonctions d'incidence cumulée (Fine, 2001; Gerds et al., 2012). Andersen and Klein (2007); Andersen and Perme (2010); Klein and Andersen (2005) ont utilisé une approche de régression basée sur des pseudo-valeurs. Dans ces différentes approches, les estimateurs sont non paramétriques de type Kaplan-Meier, et s'appliquent pour un schéma d'observations en temps continu.

Remarque III.1 ***Incidence cumulée*** *Dans un contexte de risques concurrents, on appelle souvent « incidence cumulée » (cumulative incidence) la probabilité cumulée de subir l'un des évènements en présence des évènements concurrents. Elle est parfois aussi désignée dans la littérature sous les termes de « fonction de sous-répartition » (subdistribution function) (Klein and Andersen, 2005) ou de « risque absolu » (absolute risk) (Benichou, 2005).*

Remarque III.2 ***Estimation de l'incidence cumulée*** *Dans les contextes de risques concurrents, la probabilité cumulée de subir l'un des évènements en présence de risques concurrents a souvent été estimée à tort par* $1-\hat{S}_{KM}(.)$ *(Gooley et al., 1999) où* $\hat{S}_{KM}(.)$ *est l'estimateur de Kaplan-Meier naïf de la fonction de survie calculé en censurant à droite les sujets subissant le ou les évènement(s) concurrent(s). De même que la fonction de survie (parfois dite globale) dépend du risque de subir non seulement l'évènement d'intérêt mais aussi les évènements concurrents (remarque I.7), la probabilité cumulée de subir l'évènement d'intérêt dépend aussi des risques associés aux évènements concurrents. En utilisant un estimateur de Kaplan-Meier de la fonction de survie calculé en censurant à droite les sujets subissant le ou les évènement(s) concurrent(s), l'incidence cumulée sera surestimée tandis qu'en utilisant l'estimateur de Kaplan-Meier de la fonction de survie globale, on obtient un estimateur approprié (Kalbfleisch and Prentice, 2011, p.255).*

Frydman (1995) et Frydman and Szarek (2010) se sont intéressés aux pendants des incidences cumulées dans un modèle *illness-death*, les probabilités cumulées de maladie et de décès sans maladie, et en ont proposé des estimateurs non paramétriques. Ceux-ci peuvent être calculés lorsque les temps associés à la maladie sont censurés par intervalle. Remarquons qu'ils se placent exactement dans le modèle qui nous intéresse, un modèle

illness-death où les données associées à la transition $0 \to 1$ sont censurées par intervalle. Ils ne se sont cependant pas intéressés à l'effet de variables explicatives.

D'autres auteurs se sont intéressés à l'estimation d'espérances de vie dans ce modèle, en particulier à la façon dont l'espérance de vie totale se subdivise entre l'espérance de vie sans maladie et l'espérance de vie avec maladie. Ils utilisent généralement des chaînes de Markov (processus de Markov à temps discret) et intègrent l'âge à leur modèle en tant que variable explicative (Izmirlian et al., 2000; Lièvre et al., 2003; van den Hout and Matthews, 2009). van den Hout and Matthews (2008) ont aussi estimé des espérances de vie en considérant des intensités de transition qui ne sont pas constantes ou constantes par morceaux mais paramétrées selon une loi de Weibull.

Le présent travail a pour objectif de présenter les quantités d'intérêt dans un modèle *illness-death* et de les estimer en plus des intensités de transition, à partir de données censurées par intervalle et en tenant compte de facteurs individuels (variables explicatives). Il a donné lieu à une publication (Touraine et al., 2013b).

La section 1 rappelle brièvement le modèle et apporte des précisions sur les notations utilisées par la suite. Nous détaillons dans la section 2 les quantités d'intérêt dans le modèle et les exprimons en fonction des intensités de transition. La méthode d'estimation de ces quantités est présentée dans la section 3. Elle est relativement simple puisqu'elle consiste à substituer dans les expressions de la section 2 les intensités de transition par leurs estimations. En pratique, les estimations de ces quantités peuvent être calculées grâce au paquet R **SmoothHazard**. Il a été utilisé à cet effet sur les données de Paquid dans la section 4.

1 Modèle

Nous considérons le modèle *illness-death* de la figure III.1. On note $X = \{X(t), t \geq 0\}$ le processus sous-jacent au modèle, à temps continu et, sauf mention contraire, markovien non homogène. Les états possibles de X sont notés 0, 1 et 2 et correspondent aux différents états dans lesquels les sujets peuvent se trouver, respectivement : non malade, malade et décédé. Bien sûr, ceux-ci peuvent différer selon le contexte. Les sujets sont supposés être initialement non malades : $X(0) = 0$. Dans le cas où les sujets ne sont pas observés depuis $t = 0$ mais depuis un temps a tel que $a > 0$, nous supposons également que $X(a) = 0$, c'est-à-dire qu'au début du suivi tous les sujets sont non malades.

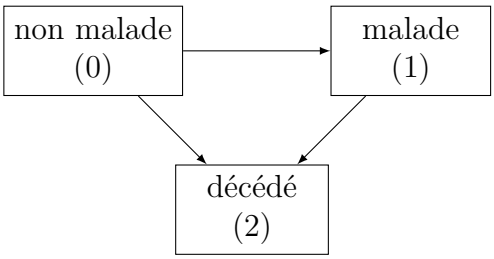


Figure III.1 – Modèle *illness-death* irréversible permettant à un sujet initialement non malade de décéder avec ou sans maladie

Les trois intensités de transition dépendent du temps (processus non homogène) et des modèles de régression de ces intensités sur des variables explicatives permettent de prendre en compte dans le modèle l'impact de facteurs individuels sur les risques de transition. Plus précisément, nous posons trois modèles à intensités de transition proportionnelles. Pour $kl = 01, 02, 12$,

$$\alpha_{kl}\Big(t \;\Big|\; Z_{kl}^{(i)}\Big) = \alpha_{0,kl}(t)\mathrm{e}^{\beta_{kl}^T Z_{kl}^{(i)}} \tag{III.1}$$

où $\alpha_{0,kl}$ est l'intensité de transition de base pour la transition $k \to l$, $Z_{kl}^{(i)}$ le vecteur des variables explicatives pour le sujet i et la transition $k \to l$, β_{kl} le vecteur des coefficients de régression pour la transition $k \to l$.

Dans la section suivante, des quantités dépendant des intensités de transition sont détaillées. Notre objectif est d'estimer ces quantités afin de faire des prévisions à un temps s pour un sujet ayant certaines caractéristiques (définies par la valeur des variables explicatives). Par conséquent, les quantités présentées dépendent (du moins, peuvent dépendre) de variables explicatives, de la même façon que les intensités de transition. Le conditionnement relatif aux variables explicatives a cependant été supprimé dans les notations afin de les alléger.

2 Quantités d'intérêt

Dans cette section, les quantités d'intérêt dans un modèle *illness-death* sont détaillées et exprimées en fonction des intensités de transition.

2.1 Probabilités de transition

Nous avons vu que les équations de Kolmogorov lient les intensités de transition et les probabilités de transition dans tout modèle multi-état, à condition de faire une hypothèse d'homogénéité dans le temps. Dans un modèle *illness-death*, qui est un modèle multi-état simple, il n'est pas nécessaire de supposer que les intensités de transition sont constantes dans le temps pour les lier aux probabilités de transition. $\forall s, t$, les probabilités $p_{kl}(s,t)$ peuvent être exprimées directement en fonction des intensités $\alpha_{kl}(.)$ (voir par exemple Putter et al. (2007)).

On définit $A_{kl}(s,t)$ la fonction d'intensité de transition cumulée entre les états k et l, et entre les dates s et t : $A_{kl}(s,t) = \int_s^t \alpha_{kl}(u)\mathrm{d}u$.

Commençons par détailler les probabilités relatives à un sujet en l'état 1 (malade) au temps s.

Pour un tel sujet, la probabilité de rester en l'état 1 (*i.e.* de ne pas décéder) jusqu'au temps t est :

$$p_{11}(s,t) = e^{-A_{12}(s,t)}$$

et la probabilité d'être en l'état 2 au temps t (de décéder entre s et t) est :

$$\begin{aligned} p_{12}(s,t) &= \int_s^t p_{11}(s,u)\alpha_{12}(u)\mathrm{d}u \\ &= 1 - p_{11}(s,t) \end{aligned}$$

Passons maintenant aux probabilités relatives à un sujet en l'état 0 (vivant et non malade) au temps s.

Pour un tel sujet, la probabilité d'être toujours vivant et non malade au temps t (*i.e.* de ne transiter ni vers l'état 1 ni vers l'état 2 entre s et t) est :

$$p_{00}(s,t) = e^{-A_{01}(s,t)-A_{02}(s,t)}$$

La probabilité d'être malade au temps t est :

$$p_{01}(s,t) = \int_s^t p_{00}(s,u)\alpha_{01}(u)p_{11}(u,t)\mathrm{d}u$$

En effet, pour être malade au temps t, il faut transiter vers l'état 1 entre s et u, où $s < u < t$, puis rester en l'état 1 jusqu'à t.

On déduit la probabilité d'être décédé en t :

$$p_{02}(s,t) = 1 - p_{00}(s,t) - p_{01}(s,t)$$

Remarquons qu'on peut transiter de l'état 0 vers l'état 2 directement ou en passant par l'état 1. La probabilité d'être décédé en t peut être formulée comme la somme de deux probabilités :

$$p_{02}(s,t) = p^0_{02}(s,t) + p^1_{02}(s,t)$$

où $p^0_{02}(s,t)$ est la probabilité de décéder sans maladie et $p^1_{02}(s,t)$ est la probabilité de décéder après avoir été malade. La première s'écrit :

$$p^0_{02}(s,t) = \int_s^t p_{00}(s,u)\alpha_{02}(u)\mathrm{d}u$$

La seconde s'écrit :

$$p^1_{02}(s,t) = \int_s^t p_{00}(s,u)\alpha_{01}(u)p_{12}(u,t)\mathrm{d}u$$

En effet, pour accomplir la trajectoire $0 \to 1 \to 2$ entre s et t, il faut atteindre l'état 1 entre s et u où $s < u < t$, puis atteindre l'état 2 entre u et t.

2.2 Probabilités cumulées

Définissons la variable T_0 comme le temps de sortie de l'état 0 du processus $X(t)$, c'est-à-dire la durée jusqu'à ce qu'un sujet devienne malade ou décède :

$$T_0 = inf\{t; X(t) \neq 0\}$$

Remarque III.3 *Contrairement aux variables de temps d'évènement latents considérés dans l'approche « latent failure times » (voir remarque I.8), la variable T_0 est correctement définie.*

Soit δ l'indicateur de la cause de sortie de l'état 0 : $\delta = 1$ si le sujet est tombé malade $(0 \to 1)$; $\delta = 2$ si le sujet est décédé $(0 \to 2)$.

Considérons les fonctions $F_{01}(s,t)$ et $F_{02}(s,t)$ qui sont des probabilités cumulées associées à T_0, correspondant respectivement au risque de maladie et au risque de

décès sans maladie :

$$F_{01}(s,t) = \mathbb{P}(T_0 \leq t, \delta = 1 | T_0 > s) \quad ; \quad F_{02}(s,t) = \mathbb{P}(T_0 \leq t, \delta = 2 | T_0 > s)$$

$F_{01}(s,t)$ et $F_{02}(s,t)$ peuvent être exprimées en fonction des probabilités de transition (qui elles-mêmes peuvent être exprimées en fonction des intensités de transition) :

$$\begin{aligned} F_{01}(s,t) &= \int_s^t p_{00}(s,u)\alpha_{01}(u)\mathrm{d}u \\ &= p_{01}(s,t) + p_{02}^1(s,t) \end{aligned}$$

$$F_{02}(s,t) = p_{02}^0(s,t)$$

En effet, $F_{01}(s,t)$ correspond à la probabilité pour un sujet non malade au temps s de devenir malade à un temps u compris entre s et t. Notons bien qu'entre u et t un tel sujet peut rester en l'état 1 ou transiter vers l'état 2. $F_{02}(s,t)$ correspond à la probabilité pour un sujet non malade au temps s de décéder entre s et t sans devenir malade auparavant.

On peut définir le risque absolu de quitter l'état non malade (pour cause de décès ou de maladie) comme la somme de F_{01} et F_{02}. On obtient :

$$F_{0\bullet}(s,t) = F_{01}(s,t) + F_{02}(s,t) = \mathbb{P}(T_0 \leq t | T_0 > s) = p_{02}(s,t) + p_{01}(s,t)$$

Les probabilités cumulées $F_{01}(s,t)$ et $F_{02}(s,t)$ sont utiles pour faire des prévisions à l'horizon t. Nous pouvons aussi faire des prévisions à l'horizon de la vie entière en faisant tendre t vers l'infini : $F_{01}(s,\infty) = \lim_{t\to+\infty} F_{01}(s,t)$ et $F_{02}(s,\infty) = \lim_{t\to+\infty} F_{02}(s,t)$. Nous avons :

$$F_{01}(s,\infty) = p_{02}^1(s,\infty) \quad ; \quad F_{02}(s,\infty) = p_{02}^0(s,\infty)$$

où $F_{01}(s,\infty)$ est le risque de maladie au cours de la vie entière (*lifetime risk*) tandis que $F_{02}(s,\infty) = 1 - F_{01}(s,\infty)$ est la probabilité de ne pas être malade au cours de la vie entière.

2.3 Espérances de vie

Dans cette sous-section, nous définissons un certain nombre de quantités de type espérances de vie résiduelles. Nous utilisons des notations un peu différentes de celles que l'on peut trouver dans la littérature. En particulier, nous nous servons de la variable T_0 définie plus haut et de la variable T que l'on définit comme étant le temps d'entrée en l'état 2 (*i.e.* le temps de décès)) :

$$T = inf\{t; X(t) = 2\}$$

Remarque III.4 *Comme T_0, la variable T est correctement définie. En fait, un modèle illness-death peut être défini par les deux variables de survie T_0 et T (Andersen and Keiding, 2012). $T_0 = T$ correspond à une transition $0 \to 2$; $T_0 < T$ correspond à une transition $0 \to 1$ en T_0 et une transition $1 \to 2$ en T.*

La plupart du temps lorsqu'on s'intéresse à des espérances de vie, l'échelle de temps choisie est l'âge, c'est-à-dire le délai depuis la naissance. Nous supposons ici que tel est le cas. Bien sûr, lorsque le délai à partir de l'entrée en l'état 0 ne correspond pas à un âge, les expressions ci-après restent valables. Par exemple, si l'état 0 correspond à la maladie, l'état 1 à une complication de la maladie et l'état 2 au décès, il pourrait être plus pertinent de travailler en délai depuis le début de la maladie et d'introduire l'âge dans le modèle en tant que variable explicative.

L'espérance de vie non malade d'un sujet d'âge s (et non malade) est :

$$\mathbb{E}\Big(T_0 - s | X(s) = 0\Big) = \int_s^{+\infty} p_{00}(s,u)\mathrm{d}u \tag{III.2}$$

Elle correspond au temps qu'il peut encore espérer passer en l'état 0.

L'espérance de vie (au sens commun du terme) d'un sujet non malade d'âge s est :

$$\mathbb{E}\Big(T - s | X(s) = 0\Big) = \int_s^{+\infty} \Big(p_{00}(s,u) + p_{01}(s,u)\Big)\mathrm{d}u \tag{III.3}$$

Elle représente le temps qu'un sujet non malade d'âge s peut encore espérer vivre. Dans la littérature, cette espérance de vie (équation III.3) est souvent appelée espérance de vie totale. Elle se subdivise entre l'espérance de vie non malade (équation III.2) et l'espérance de vie malade $\Big(\int_s^{+\infty} p_{01}(s,u)\mathrm{d}u\Big)$.

Considérons maintenant un sujet malade. À l'âge s, son espérance de vie s'écrit :

$$\mathbb{E}\Big(T - s | X(s) = 1\Big) = \int_s^{+\infty} p_{11}(s,u)\mathrm{d}u \tag{III.4}$$

Elle correspond au temps qu'il peut encore espérer passer en l'état 1.

Nous définissons maintenant l'espérance de vie d'un sujet vivant d'âge s mais dont le statut de maladie n'est pas connu, c'est-à-dire qu'il peut être non malade (en l'état 0) ou malade (en l'état 1). Cette espérance peut être intéressante dans un contexte de censure par intervalle. Supposons par exemple qu'un sujet est vivant à l'âge s (âge de dernières nouvelles) et qu'à l'âge $s^\star$ de sa dernière visite de suivi, il était non malade ($s^\star < s$). Soit π_0 et π_1 les probabilités que ce sujet soit respectivement en l'état 0 (non malade) et en l'état 1 (malade) à l'âge s. Ces probabilités s'écrivent comme suit :

$$\pi_0 = \frac{p_{00}(s^*,s)}{p_{00}(s^*,s) + p_{01}(s^*,s)} \quad ; \quad \pi_1 = \frac{p_{01}(s^*,s)}{p_{00}(s^*,s) + p_{01}(s^*,s)}$$

L'espérance de vie de ce sujet à l'âge s, ne sachant pas s'il est malade ou non s'écrit :

$$\mathbb{E}\Big(T - s\Big|X(s^*) = 0, X(s) \in \{0,1\}\Big) = \pi_0\ \mathbb{E}\Big(T - s|X(s) = 0\Big) + \pi_1\ \mathbb{E}\Big(T - s|X(s) = 1\Big) \tag{III.5}$$

Enfin, nous définissons une dernière quantité qui a la particularité de ne pas dépendre de l'âge, contrairement aux espérances de vie précédemment définies. Elle « résume » l'espérance de vie d'un sujet malade et correspond au temps moyen passé dans l'état 1 :

$$\frac{\displaystyle\int_{s_0}^{+\infty} \mathbb{E}\Big(T - s|X(s) = 1\Big)\mathrm{d}F_{01}(s)}{\displaystyle\int_{s_0}^{+\infty} \mathrm{d}F_{01}(s)} = \frac{\displaystyle\int_{s_0}^{+\infty} p_{00}(0,s)\alpha_{01}(s)\mathbb{E}\Big(T - s|X(s) = 1\Big)\mathrm{d}s}{\displaystyle\int_{s_0}^{+\infty} p_{00}(0,s)\alpha_{01}(s)\mathrm{d}s} \tag{III.6}$$

C'est l'espérance de vie à l'âge d'apparition de la maladie, moyennée sur cet âge et pondérée par la probabilité de tomber malade à cet âge. s_0 est généralement l'âge à partir duquel les sujets sont suivis et donc susceptibles de développer la maladie. En effet, dans la plupart des applications, la maladie étudiée touche des sujets d'un certain âge et les sujets sont donc suivis à partir de l'âge où ils sont susceptibles de développer la maladie. Par exemple, les sujets de Paquid intègrent la cohorte à

65 ans ou plus. Pour calculer l'espérance de vie des sujets déments dans Paquid, on poserait donc $s_0 = 65$. Dans le contexte évoqué précédemment où les états 0, 1 et 2 du processus correspondraient aux états de maladie, complication de la maladie et décès et où l'échelle de temps serait le délai depuis la maladie, on aurait $s_0 = 0$. Ainsi, la valeur s_0 varie selon l'application et l'échelle de temps choisie.

Remarque III.5 ***Sur l'hypothèse markovienne***

En supposant que le processus sous-jacent au modèle est markovien, nous supposons que les intensités de transition du modèle dépendent uniquement du temps (et des variables explicatives). Cela signifie que le risque de décéder d'un sujet malade dépend uniquement du temps. Par exemple, si l'échelle de temps choisie est l'âge, deux sujets malades de 80 ans ont le même risque de décès, même si l'un est tombé malade à 70 ans et l'autre à 79 ans. Une telle hypothèse peut être considérée comme étant trop forte.

Si l'on suppose plutôt que le processus est semi-markovien homogène, le risque de décéder d'un sujet malade dépend uniquement du temps depuis lequel il est malade. Remarquons que l'hypothèse semi-markovienne dans un modèle illness-death ne concerne que la transition $1 \to 2$*. Les risques de maladie et de décès d'un sujet non malade ne varient pas puisque les sujets étant initialement non malades (*$X(0) = 0$*), le temps passé dans l'état 0 et le temps courant sont une seule et même chose.*

Cependant, l'hypothèse d'homogénéité d'un modèle de semi-Markov peut être elle aussi considérée comme trop forte. Par exemple, deux sujets malades depuis un an ont le même risque de décéder, même si l'un a 70 ans et l'autre 80 ans. Une alternative serait de considérer un processus de semi-Markov homogène mais avec le temps d'apparition de la maladie en variable explicative sur la transition $1 \to 2$ *ou, un processus de Markov (non homogène) avec le temps passé dans l'état 1 en variable explicative sur la transition* $1 \to 2$*. Ces alternatives sont facilement envisageables s'il n'y a pas de censure par intervalle. En présence de censure par intervalle, on ne connaît pas exactement le temps d'apparition de la maladie et le temps passé dans l'état malade. Pour considérer l'un de ces temps comme une variable explicative, on doit alors l'intégrer ou l'imputer.*

Dans un modèle de semi-Markov homogène

Supposons que le processus X est semi-markovien et homogène. Dans un tel modèle, $\alpha_{12}(.)$ ne dépend plus du temps courant mais du temps passé en l'état 1. Les quantités présentées dans cette section, excepté les espérances de vie III.4, III.5 et III.6, peuvent toujours s'écrire en fonction des intensités de transition $\alpha_{01}(.)$, $\alpha_{02}(.)$ et $\alpha_{12}(.)$. Les

expressions qui diffèrent dans le cas semi-markovien sont p_{11} et celles faisant intervenir p_{11}.

Détaillons la nouvelle expression de p_{11} et des quantités qui ne peuvent être obtenues par simple substitution de p_{11}, en distinguant le cas où les temps de transition $0 \to 1$ sont censurés par intervalle et celui où ils ne le sont pas.

— Sans censure par intervalle, le temps d'apparition de la maladie est connu exactement. Considérons un sujet malade et notons t_0 la date à laquelle la maladie est survenue. La probabilité pour ce sujet de ne pas être décédé à la date t sachant qu'il n'est pas décédé à la date s, avec $t_0 < s < t$, s'écrit :

$$\begin{aligned} p_{11}(s,t) &= \mathbb{P}\Big(X(t) = 1 \,\Big|X(s) = 1, T_0 = t_0\Big) \\ &= \mathbb{P}\Big(X(t-t_0) = 1 \,\Big|X(s-t_0) = 1\Big) \\ &= \mathrm{e}^{-A_{12}(s-t_0,t-t_0)} \end{aligned}$$

L'espérance de vie d'un sujet malade depuis $s-t_0$ (qui remplace l'équation III.4) est :

$$\mathbb{E}\Big(T - s|X(s) = 1, T_0 = t_0\Big) = \int_{s-t_0}^{+\infty} p_{11}(s-t_0,u)\mathrm{d}u \tag{III.7}$$

— Avec censure par intervalle, on sait seulement que la maladie est apparue dans un certain intervalle de temps. Considérons un sujet malade et notons l et r les bornes gauches et droites de l'intervalle de censure pour ce sujet. La probabilité pour ce sujet de ne pas être décédé à la date t sachant qu'il n'est pas décédé à la date $s > r$ s'écrit :

$$p_{11}(s,t) = \int_l^r \mathrm{e}^{-A_{12}(s-u,t-u)}\mathrm{d}u$$

Son espérance de vie (qui remplace l'équation III.4) :

$$\mathbb{E}\Big(T - s|X(s) = 1, T_0 \in [l,r]\Big) = \int_l^r \int_{s-v}^{+\infty} p_{11}(s-v,u)\mathrm{d}u \; \mathrm{d}v \tag{III.8}$$

Par substitution de cette dernière espérance dans l'équation III.5, on obtient l'espérance de vie d'un sujet vivant dont on ne connaît pas le statut de maladie. L'équivalent de l'espérance III.6 serait le temps de séjour moyen passé en l'état

1 :

$$\int_0^{+\infty} p_{11}(0,u)\mathrm{d}u$$

3 Estimation

3.1 Estimation des quantités d'intérêt

La méthode pour estimer les quantités de la section précédente est immédiate. Elle consiste :

- dans un modèle sans variables explicatives : à estimer les intensités de transition, puis, à calculer les quantités d'intérêt en remplaçant les α_{kl} par les $\hat{\alpha}_{kl}$;
- dans un modèle avec variables explicatives : à estimer les intensités de transition de base $\alpha_{0,01}$, $\alpha_{0,02}$, $\alpha_{0,12}$ et les paramètres de régression β_{01}, β_{02}, β_{12}, puis, à calculer les quantités d'intérêt en remplaçant les α_{kl} par les $\hat{\alpha}_{kl}(.|Z_{kl})$.

Nous proposons d'obtenir les $\hat{\alpha}_{kl}$ ou les $\hat{\alpha}_{0,kl}$ et les $\hat{\beta}_{kl}$ par l'une des deux méthodes suivantes :

- (paramétrique) par la méthode du maximum de vraisemblance (MV) en supposant que les $\alpha_{0,kl}$ sont régies par une distribution de Weibull ;
- (semi-paramétrique) par la méthode du maximum de vraisemblance pénalisée (MVP) en approximant les $\hat{\alpha}_{0,kl}$ par des M-splines.

Ces méthodes sont celles qui sont détaillées à la fin du chapitre I, utilisées dans l'application de la section 4 et qui sont disponibles dans le paquet R **SmoothHazard**. D'autres méthodes pourraient être utilisées pourvu qu'elles fournissent des estimations des intensités de transition qui permettent les calculs d'intégrales.

Nous ne détaillons pas ici l'écriture de la vraisemblance dans un modèle *illness-death* pour données censurées par intervalle car cela a déjà été fait dans le chapitre I (voir équations I.7 et I.8). Remarquons que ces deux méthodes sont aussi valables dans un modèle *illness-death* sans censure par intervalle et que dans ce cas particulier **SmoothHazard** peut encore être utilisé afin d'estimer les $\alpha_{0,kl}$ et les β_{kl}, puis les quantités d'intérêt. Dans ce cas particulier de schéma d'observations en temps continu, détaillons la contribution à la vraisemblance L_i. Notons $\tilde{T}_0 = T_0 \wedge C$ et $\tilde{T} = T \wedge C$ où C est une variable de censure. Notons de plus δ_1 l'indicateur de maladie (1 si malade, 0 sinon), et δ_2 l'indicateur de décès (1 si décédé, 0 sinon). La contribution s'écrit :

$$L_i = p_{00}(a,\tilde{T}_0)\alpha_{01}(\tilde{T}_0)^{\delta_1}p_{11}(\tilde{T}_0,\tilde{T})^{\delta_1}\alpha_{12}(\tilde{T})^{\delta_1\delta_2}\alpha_{02}(\tilde{T}_0)^{\delta_2(1-\delta_1)}$$

Dans un modèle de semi-Markov homogène

Ces méthodes peuvent encore être appliquées si l'on considère un modèle de semi-Markov homogène au lieu d'un modèle de Markov non homogène. Supposons que l'on se place dans cette situation et détaillons la vraisemblance ; d'abord, pour un schéma d'observations en temps continu (*i.e.* sans censure par intervalle) puis, pour un schéma d'observations des dates de décès en temps continu et des dates de maladie en temps discret (*i.e.* avec censure par intervalle).

— Pour un schéma d'observations en temps continu, la contribution à la vraisemblance s'écrit :

$$L_i = p_{00}(a, \tilde{T}_0)\alpha_{01}(\tilde{T}_0)^{\delta_1} p_{11}(0, \tilde{T} - \tilde{T}_0)^{\delta_1}\alpha_{12}(\tilde{T} - \tilde{T}_0)^{\delta_1\delta_2}\alpha_{02}(\tilde{T}_0)^{\delta_2(1-\delta_1)}$$

— Dans le cas de données censurées par intervalle, considérons d'abord un sujet qui n'a pas été observé malade. Soit l sa date de dernière observation. La contribution à la vraisemblance de ce sujet est :

$$L_i = p_{00}(a, l)\left(p_{00}(l, \tilde{T})\alpha_{02}(\tilde{T})^{\delta_2} + \int_l^{\tilde{T}} p_{00}(l, u)\alpha_{01}(u)p_{11}(0, \tilde{T} - u)\alpha_{12}(\tilde{T} - u)^{\delta_2}\mathrm{d}u\right) \tag{III.9}$$

L'intégrale est due au fait que si le sujet a séjourné dans l'état 1 entre l et $\tilde{T}$, on ne connaît pas le temps exact de transition $0 \to 1$ et donc on ne connaît pas exactement la durée passée dans cet état, comprise entre $\tilde{T} - l$ et 0.

Considérons maintenant un sujet qui a été vu malade en r et vu non malade pour la dernière fois en l. Sa contribution à la vraisemblance est :

$$L_i = p_{00}(a, l)\int_l^r p_{00}(l, u)\alpha_{01}(u)p_{11}(0, \tilde{T} - u)\alpha_{12}(\tilde{T} - u)^{\delta_2}\mathrm{d}u \tag{III.10}$$

La durée passée dans l'état 1 n'est ici pas connue exactement ; elle est comprise entre $\tilde{T} - l$ et $\tilde{T} - r$.

3.2 Intervalles de confiance et bandes de confiance

Des intervalles de confiance des quantités d'intérêt peuvent être calculés à un temps fixé (ou à des temps fixés pour les quantités dépendant de deux temps comme les probabilités de transition). Nous utilisons une technique de simulation, qui a déjà été utilisée dans d'autres contextes (van den Hout and Matthews, 2008, 2010) et a récemment été formellement justifiée (Mandel, 2013). Cette technique est basée sur l'hypothèse

selon laquelle l'estimateur du vecteur des paramètres du modèle est asymptotiquement normal.

Pour plus de généralité, considérons un modèle avec variables explicatives. Soit $\hat{\theta}$ le vecteur des paramètres estimés du modèle. Ce vecteur contient les estimations des paramètres liés aux intensités de transition de base, c'est-à-dire des paramètres définissant les trois lois de Weibull (méthode paramétrique) ou les estimations des paramètres correspondant aux coefficients des splines (méthode semi-paramétrique). Il contient de plus, les estimations des paramètres de régression $\hat{\beta}_{kl}$. Notons $\hat{V}_{\hat{\theta}}$ la matrice de variance-covariance de $\hat{\theta}$. Considérons une distribution normale multivariée de moyenne $\hat{\theta}$ et de matrice de variance-covariance $\hat{V}_{\hat{\theta}}$. La technique consiste à générer aléatoirement n vecteurs de paramètres $\theta^{(1)}, \ldots, \theta^{(n)}$ à partir de cette distribution. Chaque $\theta^{(q)}$ permet de calculer les intensités de transition en n'importe quel(s) point(s) ; notons-les $\alpha_{kl}^{(q)}(.)$. Notons $Q(s,t)$ la quantité d'intérêt dont on souhaite calculer un intervalle de confiance. On la calcule n fois à l'aide des $\alpha_{kl}^{(1)}(.), \ldots, \alpha_{kl}^{(n)}(.)$. On obtient les valeurs $Q(s,t)^{(1)}, \ldots, Q(s,t)^{(n)}$ qui reflètent la fluctuation d'échantillonnage (Aalen et al., 1997) et on les classe par ordre croissant. Nous utilisons alors les $2.5^{\text{èmes}}$ et $97.5^{\text{èmes}}$ percentiles comme bornes inférieure et supérieure de l'intervalle de confiance de $Q(s,t)$. Cette procédure pouvant être répétée quel que soit s (resp. quel que soit t), il est aussi possible d'obtenir des bandes de confiance point par point (*pointwise confidence bands*) pour la courbe associée à la fonction $Q(.,t)$ (resp. $Q(s,.)$).

3.3 Un large éventail de prévisions

Les estimations des intensités de transition $\hat{\alpha}_{kl}(.)$ nous permettent d'estimer un grand nombre de quantités d'intérêt qui dépendent d'un seul temps s ou de deux temps s et t. Cela induit un large éventail de prévisions possibles. On peut tout d'abord estimer des quantités d'intérêt à des temps s et t fixés. Par exemple, on peut s'intéresser aux probabilités que des sujets non malades de 70, 75 et 80 ans connaissent la maladie dans leur vie et comparer ces probabilités selon les caractéristiques de ces sujets. D'un point de vue épidémiologique, il peut aussi être intéressant de regarder comment les différentes quantités évoluent avec l'âge. Pour les quantités dépendant de s et de t, on peut alors fixer s (respectivement t) et faire varier t (respectivement s). On peut également faire évoluer simultanément s et t. Par exemple, on peut tracer la courbe de la probabilité qu'ont des sujets non malades de connaître la maladie dans les 5 ans à venir selon leur âge. Un large éventail de prévisions est donc possible. Nous essayons

dans la section suivante d'en donner un aperçu.

3.4 Mises en garde

Les quantités de la section précédente sont valables pour tout $s \in \mathbb{R}$ et pour tout $t \in \mathbb{R}$. Cependant les données disponibles et la méthode d'estimation utilisée (paramétrique ou semi-paramétrique) ne nous permettent pas toujours de les estimer $\forall s$ et $\forall t$.

3.4.1 Méthode d'estimation utilisée

Si la méthode d'estimation utilisée est paramétrique, les $\hat{\alpha}_{kl}(.)$ sont définies en tout point. Il est donc techniquement possible de calculer n'importe quelle quantité d'intérêt $\forall s$ et $\forall t$. Si elle est semi-paramétrique, les $\hat{\alpha}_{kl}(.)$ sont uniquement définies entre le premier nœud et le dernier nœud.

Il n'est donc possible d'estimer que des quantités qui dépendent d'un s supérieur aux premiers nœuds des transitions concernées. Par exemple, pour estimer $p_{11}(s,t)$ (qui est une fonction de $\alpha_{12}(.)$) il faut que s soit supérieur au premier nœud de la transition $1 \to 2$; et pour estimer $p_{00}(s,t)$ (qui est une fonction de $\alpha_{01}(.)$ et $\alpha_{02}(.)$) il faut que s soit supérieur au plus grand des premiers nœuds des transitions $0 \to 1$ et $0 \to 2$.

De la même façon, il n'est possible d'estimer que des quantités qui dépendent d'un t inférieur aux derniers nœuds des transitions concernées. Par exemple, pour estimer $p_{11}(s,t)$, il faut que t soit inférieur au dernier nœud de la transition $1 \to 2$; et pour estimer $p_{00}(s,t)$, il faut que t soit inférieur au plus petit des derniers nœuds des transitions $0 \to 1$ et $0 \to 2$. Les quantités qui peuvent poser vraiment problème sont celles où t tend vers l'infini. Dans certains cas, elles peuvent être estimées mais cela nous renvoie à la question des données disponibles.

3.4.2 Données disponibles

Les données disponibles doivent aussi être prises en compte dans l'estimation des quantités d'intérêt. Les intensités de transition sont estimées sur la base des durées observées. Il faut donc avoir conscience que même si les intensités peuvent être calculées en dehors des temps d'observation (comme dans la méthode paramétrique où elles sont calculables en tout temps), leur valeur en dehors de ces temps d'observation relève plus d'une extrapolation que d'une estimation. Il n'est donc pas souhaitable d'estimer des quantités pour lesquelles s est trop inférieur au premier temps de suivi ou t est trop

supérieur au dernier temps d'observation. Ainsi, dans une étude où le temps de suivi est court, nous déconseillons d'estimer des quantités de type espérances de vie, même si cela est possible avec une méthode d'estimation paramétrique. En revanche, lorsque le temps de suivi est suffisant, il est alors envisageable d'estimer des quantités faisant intervenir l'infini, même avec une méthode d'estimation semi-paramétrique. En effet, lorsque le temps de suivi est long, beaucoup de sujets sont décédés (sont entrés dans l'état absorbant 2) et on peut considérer qu'au temps de décès le plus élevé t_{max} (ou légèrement après), tous les sujets seraient décédés. Dans les quantités faisant intervenir l'infini, l'infini peut alors être remplacé par t_{max}.

4 Illustration sur les données de Paquid

La cohorte Paquid, déjà présentée dans le chapitre II, a été utilisée afin d'illustrer le type de prévisions qu'il est possible de réaliser. De la même façon que précédemment, l'âge a été choisi comme temps de base et les échantillons des hommes et des femmes ont été analysés séparément. Nous avons considéré deux modèles : l'un sans variables explicatives et l'autre avec le certificat d'études primaires (cep) comme variable explicative sur les trois transitions. Les paramètres de chaque modèle ont été estimés avec chacune des méthodes d'estimation proposées plus haut :

i) la méthode paramétrique consistant à maximiser la vraisemblance et à spécifier les intensités de transition de base avec une distribution de Weibull ;

ii) la méthode semi-paramétrique consistant à maximiser la vraisemblance pénalisée et à approcher les estimateurs des intensités de transition de base par des M-splines.

Pour chaque modèle et chaque méthode d'estimation, les intensités de transition puis les quantités d'intérêt ont été estimées. Les intervalles et bandes de confiance point par point ont été calculés en simulant $n = 2000$ vecteurs de paramètres. Toutes ces analyses ont été faites à l'aide du paquet R **SmoothHazard**.

Méthode semi-paramétrique

Choix des paramètres de lissage

Les paramètres de lissage de la méthode semi-paramétrique ont été choisis de la façon suivante. Le modèle sans variable avec estimation semi-paramétrique a d'abord été appliqué sur les échantillons des hommes et des femmes avec une option de recherche des paramètres de lissage par validation croisée (plus précisément par *approximate leave-one-out cross validation*). Les paramètres de lissage obtenus étaient $\kappa_{01} = 1439325$,

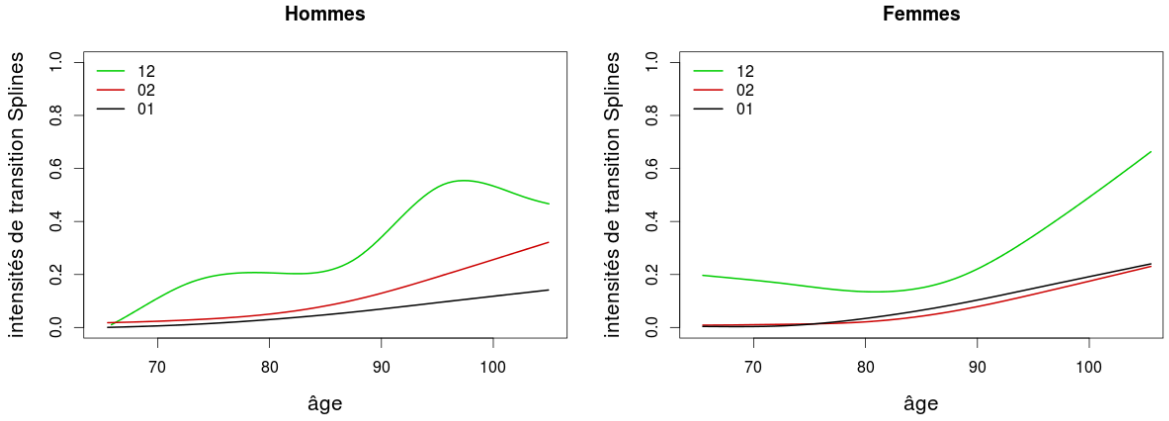


Figure III.2 – Modèle sans variable explicative : intensités de transition $\hat{\alpha}_{01}$, $\hat{\alpha}_0$, $\hat{\alpha}_{12}$ de type splines et estimées par MVP avec recherche automatique du paramètre de lissage. Les paramètres obtenus sont $\kappa_{01} = 1439325$, $\kappa_{02} = 204287$, $\kappa_{12} = 860$ pour les hommes et $\kappa_{01} = 545160$, $\kappa_{02} = 575460$, $\kappa_{12} = 14798$ pour les femmes.

$\kappa_{02} = 204287$ et $\kappa_{12} = 860$ pour les hommes, et $\kappa_{01} = 545160$, $\kappa_{02} = 575460$ et $\kappa_{12} = 14798$ pour les femmes. Nous avons tracé les courbes des intensités de transition et celle associée à la transition $1 \to 2$ ne paraissait pas suffisamment lisse, en particulier chez les hommes (voir figure III.2). Nous avons alors appliqué à plusieurs reprises le même modèle mais en faisant varier des paramètres de lissage fixés, en particulier en augmentant progressivement κ_{12}. Nous avons finalement retenu les paramètres de lissage suivant : $\kappa_{01} = 1\ 000\ 000$, $\kappa_{02} = 200\ 000$, $\kappa_{12} = 50\ 000$ pour les hommes, et $\kappa_{01} = 500\ 000$, $\kappa_{02} = 500\ 000$, $\kappa_{12} = 50\ 000$ pour les femmes, qui correspondent aux courbes des intensités de transition de la partie basse de la figure III.3. Les mêmes paramètres de lissage ont été utilisés pour le modèle avec le cep en variable explicative.

Choix du nombre de nœuds

Le nombre de nœuds choisi est de 7 sur les trois transitions. La méthode que nous utilisons pour choisir le nombre de nœuds est la suivante. Nous commençons par estimer les paramètres du modèle sans variables avec un nombre faible de nœuds (nous avons commencé avec 5 sur chaque transition) et recherche du paramètre de lissage par validation croisée. Nous estimons à nouveau les paramètres du modèle en augmentant progressivement le nombre de nœuds et nous nous arrêtons lorsque l'allure des courbes des intensités de transition est stable. En effet, nous augmentons dans un premier temps la flexibilité des fonctions d'intensité de transition en ajoutant des nœuds. Mais

à partir d'un certain nombre de nœuds, la flexibilité n'est plus augmentée car c'est le paramètre de lissage qui contrôle l'équilibre entre l'ajustement aux données et le degré de lissage. Il est alors inutile d'augmenter le nombre de nœuds car plus il y a de nœuds, plus il y a de paramètres à estimer et donc, plus on augmente le temps de calcul.

Choix de la place des nœuds

La forme des fonctions splines en général n'étant pas très sensible au placement des nœuds, tant qu'il y a suffisamment d'observations entre chaque nœud, nous les avons placés de façon équidistante. Les premiers nœuds des transitions $0 \to 1$ et $0 \to 2$ ont été placés au minimum des âges a d'entrée dans la cohorte. Le premier nœud de la transitions $1 \to 2$ a été placé à l'âge correspondant à la plus petite valeur l (borne gauche de l'intervalle de censure des sujets déments ou dernier âge de suivi des sujets non déments). Les derniers nœuds des trois transitions ont été placés 5 ans après le maximum de tous les âges de décès ce qui correspond à des âges d'un peu plus de 110 ans pour les hommes et d'un peu plus de 111 ans pour les femmes. Nous avons supposé qu'à ces âges, la probabilité d'être décédé (*i.e.* dans l'état absorbant 2) est de 1. Nous avons ainsi pu estimé les probabilités cumulées à l'échelle de la vie entière ($F_{01}(.,\infty)$ et $F_{02}(.,\infty)$) et les espérances de vie. Pour vérifier que cette hypothèse est acceptable, nous avons vérifié que les probabilités à l'échelle de la vie entière se sommaient approximativement à 1 lorsqu'on remplaçait l'infini par le dernier nœud choisi pour la méthode semi-paramétrique. Notons que les intensités de transition sur les figures III.2 et III.3 n'ont pas été tracées sur tout le domaine sur lequel elles ont été estimées (*i.e.* sur le domaine allant du premier au dernier nœud associés à la transition $k \to l$) mais seulement jusqu'aux âges pour lesquels des temps d'évènements ont été observés.

4.1 Modèle sans variables explicatives

Considérons dans un premier temps un modèle sans variables explicatives afin de bien appréhender les différentes quantités d'intérêt.

4.1.1 Estimation des intensités de transition

Les intensités de transition estimées avec les méthodes paramétrique et semi-paramétrique sont tracées sur la figure III.3.

Nous remarquons des différences d'une part entre les hommes et les femmes (*i.e.*

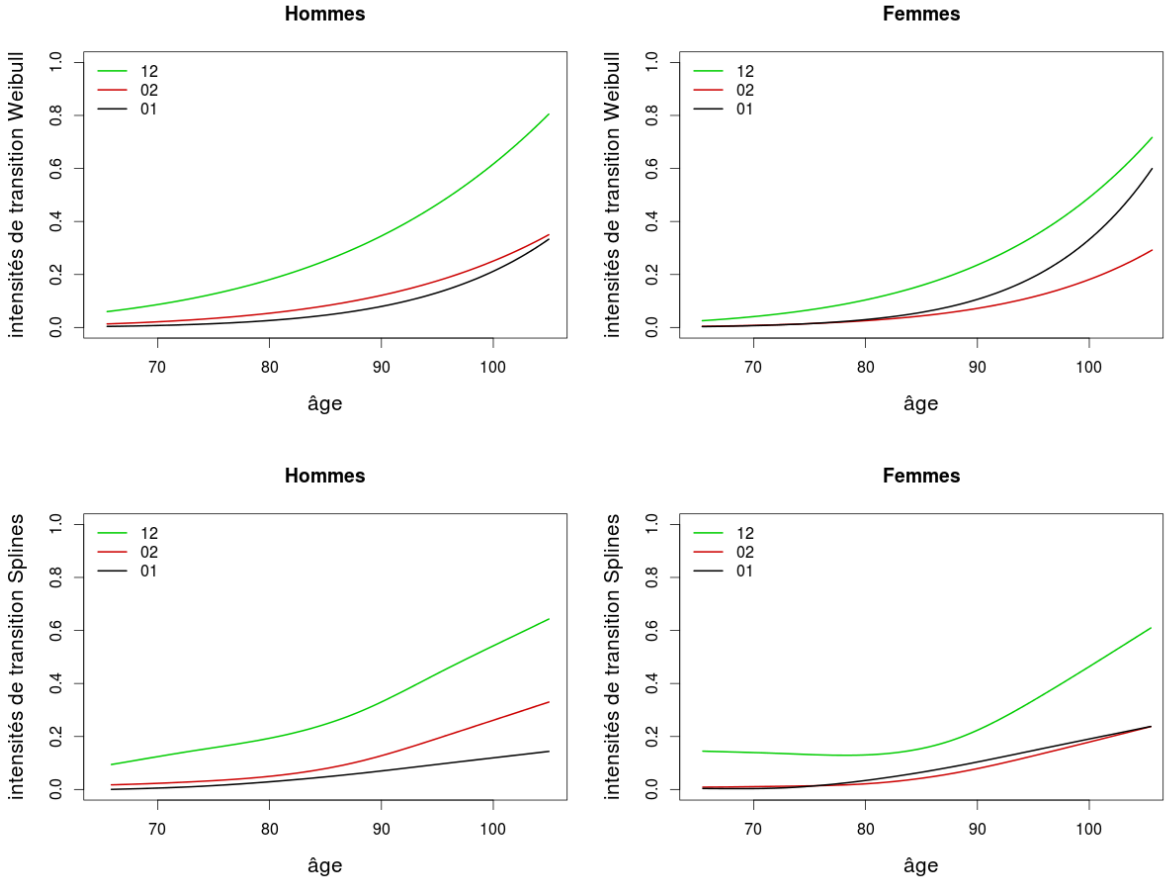


Figure III.3 – Modèle sans variables explicatives : estimations des intensités de transition $\hat{\alpha}_{01}$, $\hat{\alpha}_0$, $\hat{\alpha}_{12}$ *i*) de type Weibull et estimées par MV (partie haute), *ii*) de type splines et estimées par MVP (partie basse).

entre la partie gauche et la partie droite de la figure) et, d'autre part, entre les méthodes d'estimation (*i.e.* entre la partie haute et la partie basse de la figure).

— Les courbes de $\hat{\alpha}_{02}$ (rouges) et $\hat{\alpha}_{12}$ (vertes) ont tendance à être plus élevées chez les hommes que chez les femmes. En d'autres termes, le risque instantané de décès des hommes non déments (resp. déments) a tendance à être plus élevé que celui des femmes non démentes (resp. démentes) du même âge. En revanche, les courbes de $\hat{\alpha}_{01}$ (noires) sont plus élevées chez les femmes : le risque instantané de démence est plus élevé chez les femmes que chez les hommes du même âge. Enfin, on remarque que les hommes non malades ont un risque de décès toujours plus élevé que le risque de démence (courbe rouge au-dessus de la noire), tandis que chez les femmes ce serait plutôt le contraire, du moins à partir de 80 ans.

— Les allures des courbes sont assez différentes selon qu'on utilise la méthode d'estimation paramétrique (partie haute de la figure) ou semi-paramétrique (partie basse de la figure), en particulier sur les dernières années. Les prévisions que nous allons maintenant présenter utilisent ces intensités de transition estimées. Il sera intéressant de voir à quel point ces différences impactent les prévisions.

4.1.2 Prévisions

Les estimations des intensités de transition ont été utilisées pour estimer les quantités d'intérêt présentées.

Probabilités de transition

Nous avons d'abord estimé des probabilités de transition. Le tableau III.1 (resp. le tableau III.2) montre les probabilités de transition estimées entre 70 et 80 ans, 70 et 90 ans, 80 et 90 ans et 80 et 100 ans, calculées avec les $\hat{\alpha}_{kl}$ estimées par la méthode paramétrique (resp. semi-paramétrique). Notons que $p^0_{02}(s,t)$ et $p^1_{02}(s,t)$ ne sont pas des probabilités de transition à proprement parler mais que ces quantités apportent une information plus précise que la probabilité de transition $p_{02}(s,t)$ qui peut être retrouvée en en faisant la somme.

Tableau III.1 – Modèle sans variables explicatives : estimations des probabilités de transition $\hat{p}_{ij}(s,t)$ avec intervalles de confiance basées sur le modèle paramétrique chez les hommes (♂) et les femmes (♀).

(s,t)		$\hat{p}_{00}$	$\hat{p}_{01}$	$\hat{p}^0_{02}$	$\hat{p}^1_{02}$	$\hat{p}_{11}$	$\hat{p}_{12}$
(70,80)	♂	0.60 [0.58 ;0.63]	0.07 [0.06 ;0.09]	0.28 [0.25 ;0.31]	0.05 [0.03 ;0.07]	0.28 [0.22 ;0.51]	0.72 [0.49 ;0.78]
	♀	0.73 [0.71 ;0.75]	0.10 [0.09 ;0.11]	0.13[0.11 ;0.15]	0.04 [0.03 ;0.05]	0.50 [0.43 ;0.62]	0.50 [0.38 ;0.57]
(70,90)	♂	0.16 [0.14 ;0.18]	0.05 [0.05 ;0.08]	0.56 [0.53 ;0.60]	0.23 [0.18 ;0.26]	0.02 [0.01 ;0.09]	0.98 [0.91 ;0.99]
	♀	0.25 [0.23 ;0.27]	0.14 [0.13 ;0.16]	0.34 [0.31 ;0.37]	0.26 [0.23 ;0.29]	0.10 [0.07 ;0.16]	0.90 [0.84 ;0.93]
(80,90)	♂	0.27 [0.24 ;0.29]	0.08 [0.07 ;0.11]	0.47 [0.44 ;0.50]	0.19 [0.15 ;0.21]	0.08 [0.06 ;0.18]	0.92 [0.82 ;0.94]
	♀	0.35 [0.33 ;0.36]	0.17 [0.16 ;0.19]	0.28 [0.26 ;0.31]	0.20 [0.18 ;0.22]	0.20 [0.17 ;0.26]	0.80 [0.74 ;0.83]
(80,100)	♂	0.01 [0.01 ;0.02]	0.01 [0.01 ;0.02]	0.62 [0.57 ;0.66]	0.36 [0.31 ;0.41]	0.00 [0.00 ;0.01]	1.00 [0.99 ;1.00]
	♀	0.01 [0.01 ;0.02]	0.03 [0.03 ;0.05]	0.41 [0.37 ;0.45]	0.54 [0.50 ;0.58]	0.01 [0.00 ;0.01]	0.99 [0.99 ;1.00]

Tableau III.2 – Modèle sans variables explicatives : estimations des probabilités de transition $\hat{p}_{ij}(s,t)$ avec intervalles de confiance basées sur le modèle semi-paramétrique chez les hommes (♂) et les femmes (♀).

(s,t)		$\hat{p}_{00}$	$\hat{p}_{01}$	$\hat{p}^0_{02}$	$\hat{p}^1_{02}$	$\hat{p}_{11}$	$\hat{p}_{12}$
(70,80)	♂	0.60 [0.56 ;0.62]	0.07 [0.05 ;0.08]	0.28 [0.25 ;0.31]	0.06 [0.04 ;0.08]	0.20 [0.08 ;0.36]	0.80 [0.64 ;0.92]
	♀	0.75 [0.72 ;0.77]	0.08 [0.07 ;0.10]	0.13 [0.11 ;0.15]	0.04 [0.03 ;0.06]	0.26 [0.12 ;0.44]	0.74 [0.56 ;0.88]
(70,90)	♂	0.16 [0.14 ;0.17]	0.05 [0.04 ;0.06]	0.55 [0.52 ;0.59]	0.23 [0.20 ;0.27]	0.02 [0.01 ;0.03]	0.98 [0.97 ;0.99]
	♀	0.25 [0.22 ;0.26]	0.15 [0.14 ;0.17]	0.33 [0.30 ;0.36]	0.27 [0.24 ;0.30]	0.05 [0.02 ;0.09]	0.95 [0.91 ;0.98]
(80,90)	♂	0.27 [0.24 ;0.29]	0.08 [0.05 ;0.12]	0.46 [0.42 ;0.50]	0.19 [0.16 ;0.23]	0.08 [0.05 ;0.12]	0.92 [0.88 ;0.95]
	♀	0.33 [0.31 ;0.35]	0.18 [0.17 ;0.20]	0.27 [0.24 ;0.30]	0.22 [0.20 ;0.25]	0.20 [0.16 ;0.24]	0.80 [0.76 ;0.84]
(80,100)	♂	0.01 [0.01 ;0.02]	0.01 [0.01 ;0.01]	0.63 [0.58 ;0.67]	0.35 [0.30 ;0.40]	0.00 [0.00 ;0.00]	1.00 [1.00 ;1.00]
	♀	0.02 [0.02 ;0.03]	0.03 [0.02 ;0.04]	0.41 [0.37 ;0.45]	0.54 [0.50 ;0.58]	0.00 [0.00 ;0.01]	0.99 [0.99 ;1.00]

La première remarque que l'on peut faire en regardant ces deux tableaux est que globalement les valeurs ne varient que très peu selon la méthode d'estimation utilisée pour les intensités de transition. Cela n'était pas forcément prévisible au vu des courbes de la figure III.3. Cependant, l'estimation de α_{12} plus haute avec la méthode semi-paramétrique chez les femmes entre 70 et 80 ans (côté droit de la figure III.3) induit des estimations de $\hat{p}_{11}(70, 80)$ et $\hat{p}_{12}(70, 80)$ très différentes chez les femmes. La deuxième remarque que l'on peut faire est que les estimations des probabilités de transition sont cohérentes avec les différences entre hommes et femmes précédemment observées sur ces mêmes courbes : un risque de démence plus élevé chez les femmes et des risques de décès (des non déments et des déments) plus élevés chez les hommes.

Concentrons-nous par exemple sur le tableau III.1 avec $(s, t) = (70, 90)$. Un homme (resp. une femme) qui est non dément(e) à 70 ans a une probabilité :

— de 0.16 (resp. 0.25) d'être toujours vivant(e) et non dément(e) à 90 ans ;
— de 0.05 (resp. 0.14) d'être dément(e) à 90 ans ;
— de 0.79 (resp. 0.60) d'être décédé(e) à 90 ans.

En revanche, une femme non démente à 70 ans a une probabilité légèrement plus importante qu'un homme d'être décédée entre 70 et 90 ans tout en ayant été démente entre 70 ans et son décès (0.26 *vs* 0.23). Mais sa probabilité d'être décédée sans démence est beaucoup moins importante que celle d'un homme (0.56 *vs* 0.34).

Les estimations des probabilités de transition peuvent aussi être visualisées sur des courbes. Par exemple si l'on s'intéresse aux probabilités relatives aux sujets non déments à 70 ans, on peut tracer leur probabilité d'être toujours non déments, déments, décédés sans démence et décédés avec démence à l'âge t, en faisant varier t de 71 à 100 ans (voir figure III.4). Ces courbes confirment le fait que les probabilités de transition ne diffèrent que très peu selon qu'on estime préalablement les intensités de transition avec la méthode paramétrique (partie haute de la figure) ou semi-paramétrique (partie basse de la figure).

Dans la suite de cette sous-section, nous avons choisi de tracer sur les mêmes graphiques les courbes relatives aux hommes et les courbes relatives aux femmes afin d'axer les commentaires sur les différences hommes femmes. D'autres choix auraient pu être faits. Par exemple concernant les probabilités cumulées, auxquelles nous allons nous intéresser maintenant, nous aurions pu mettre en perspective les différences entre probabilité de démence et probabilité de décès sans démence.

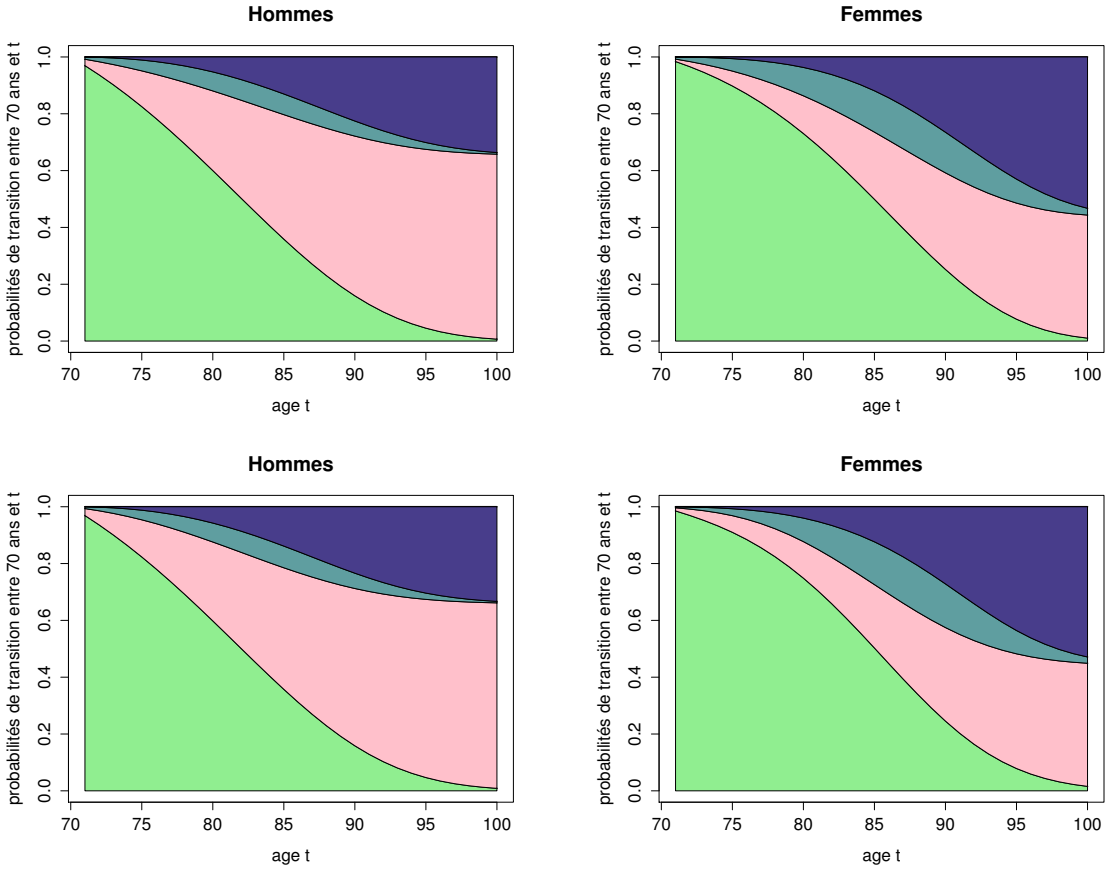


Figure III.4 – Modèle sans variables explicatives : estimations des probabilités de transition entre 70 ans et t, $71 \leq t \leq 100$ calculées avec les $\hat{\alpha}_{kl}(.)$ $i)$ de type Weibull et estimées par MV (partie haute), $ii)$ de type splines et estimées par MVP (partie basse). De bas en haut : $p_{00}(70,t)$ (vert), $p^0_{02}(70,t)$ (rose), $p_{01}(70,t)$ (bleu), $p^1_{02}(70,t)$ (violet).

Probabilités cumulées à un horizon fini

Les estimations des probabilités cumulées (relatives aux sujets non déments) $F_{01}(s,t)$ et $F_{02}(s,t)$ entre les âges 70 et 80 ans, 70 et 90 ans, 80 et 90 ans, 80 et 100 ans figurent dans le tableau III.3. Les valeurs de la moitié haute du tableau peuvent être retrouvées sur la figure III.5 qui montre les estimations des probabilités cumulées entre les âges 70 et t, avec t variant entre 71 ans et 100 ans. Plus précisément, cette figure montre les probabilités pour un sujet non dément à 70 ans de devenir dément entre 70 ans et t (courbe de gauche) et de décéder sans démence entre 70 ans et t (courbe de droite).

Les estimations des probabilités $F_{01}(70,t)$ et $F_{02}(70,t)$ s'écrivant respectivement $\hat{p}_{01}(70,t) + \hat{p}^1_{02}(70,t)$ et $\hat{p}^0_{02}(70,t)$, nous remarquons sans surprise que les estimations utilisant les $\hat{\alpha}_{kl}(.)$ de type Weibull avec MV (méthode paramétrique) et celles utilisant

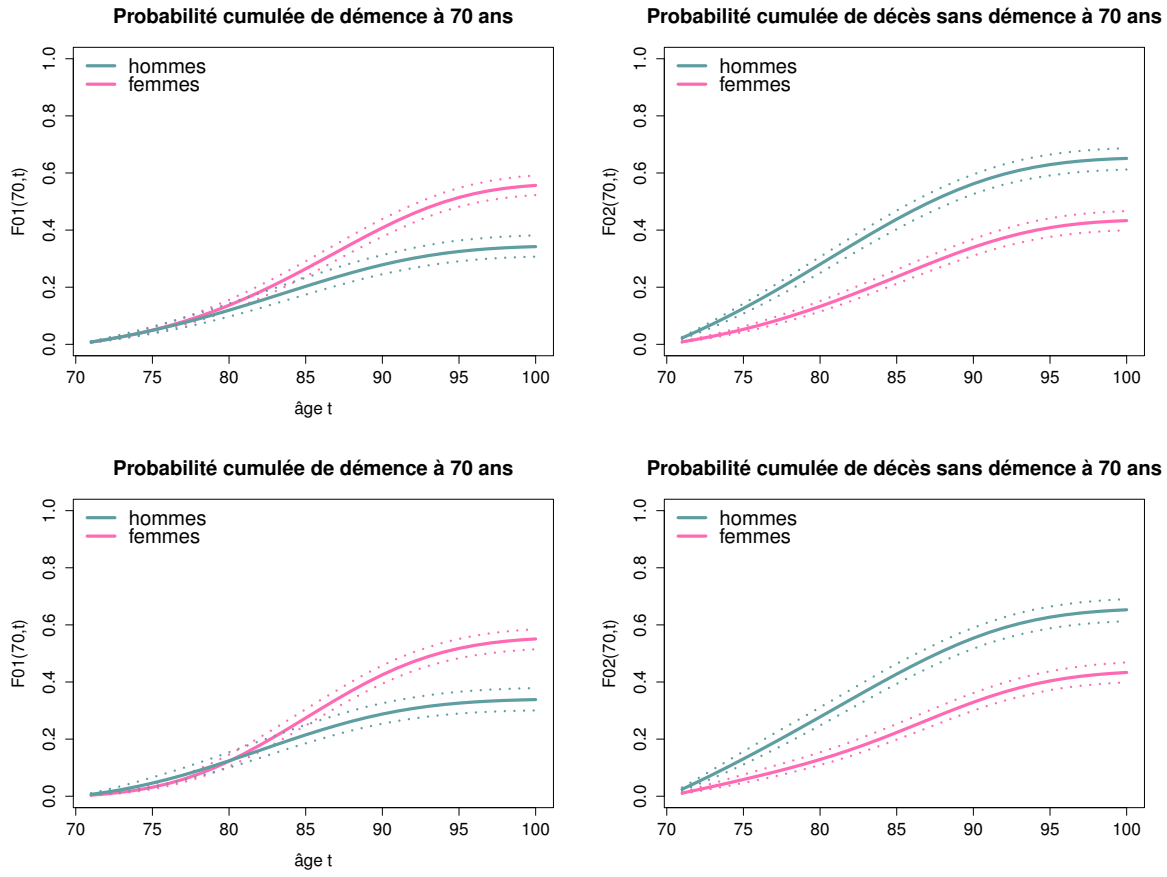


Figure III.5 – Modèle sans variables explicatives : estimations des probabilités cumulées entre 70 ans et t, $71 \leq t \leq 100$ ($F_{01}(70,t)$ à gauche, $F_{02}(70,t)$ à droite) avec leurs bandes de confiance, calculées avec les $\hat{\alpha}_{kl}(.)$ i) de type Weibull et estimées par MV (partie haute), ii) de type splines et estimées par MVP (partie basse).

Tableau III.3 – Modèle sans variables explicatives : estimations des probabilités cumulées $\hat{F}_{01}(s,t)$ et $\hat{F}_{02}(s,t)$ avec intervalles de confiance basées sur la méthode paramétrique (à gauche) semi-paramétrique (à droite) chez les hommes (♂) et les femmes (♀).

		méthode paramétrique		méthode semi-paramétrique	
(s,t)		$\hat{F}_{01}$	$\hat{F}_{02}$	$\hat{F}_{01}$	$\hat{F}_{02}$
(70,80)	♂	0.12 [0.10 ;0.14]	0.28 [0.25 ;0.31]	0.12 [0.10 ;0.15]	0.28 [0.25 ;0.31]
	♀	0.14 [0.12 ;0.16]	0.13 [0.11 ;0.15]	0.12 [0.10 ;0.15]	0.13 [0.11 ;0.15]
(70,90)	♂	0.28 [0.25 ;0.31]	0.56 [0.53 ;0.60]	0.29 [0.25 ;0.33]	0.55 [0.52 ;0.59]
	♀	0.41 [0.38 ;0.44]	0.34 [0.31 ;0.37]	0.43 [0.39 ;0.46]	0.33 [0.30 ;0.36]
(80,90)	♂	0.27 [0.23 ;0.30]	0.47 [0.44 ;0.50]	0.27 [0.24 ;0.31]	0.46 [0.42 ;0.50]
	♀	0.37 [0.35 ;0.40]	0.28 [0.26 ;0.31]	0.40 [0.37 ;0.43]	0.27 [0.24 ;0.30]
(80,100)	♂	0.37 [0.32 ;0.42]	0.62 [0.57 ;0.66]	0.36 [0.31 ;0.41]	0.63 [0.58 ;0.67]
	♀	0.57 [0.53 ;0.61]	0.41 [0.37 ;0.45]	0.57 [0.53 ;0.61]	0.41 [0.37 ;0.45]

les $\hat{\alpha}_{kl}(.)$ de type splines avec MVP (méthode semi-paramétrique) sont très proches.

Notons bien que les probabilités cumulées de démence (partie gauche de la figure) et de décès sans démence (partie droite de la figure) sont aussi visibles sur la figure III.4. L'aire sous leur courbe est identique à respectivement l'aire en rose et la réunion des aires bleues et violettes de la figure III.4. Cependant, sur la figure III.5 où les femmes et les hommes sont représentés sur les mêmes graphiques, et où l'on a tracé les bandes de confiance point par point des courbes, les différences hommes-femmes apparaissent nettement. Concentrons-nous sur la partie gauche de la figure (probabilité cumulée de démence à 70 ans). À 70 ans, la probabilité de devenir dément dans les 1, 2, 3, etc., 10 années à venir semble être la même pour les hommes et pour les femmes puisqu'entre 71 et 80 ans les courbes ne se détachent pas vraiment. En revanche, la probabilité de devenir dément dans les 17, 18, 19, etc., 30 années à venir est bien plus élevée chez les femmes que chez les hommes puisqu'entre 87 et 100 ans les courbes se détachent franchement. On peut expliquer cette différence en partie parce que le risques de décès des femmes est plus faible que celui des hommes.

La figure III.6 représente aussi les probabilités cumulées de démence et de décès sans démence mais apporte un autre genre d'information en faisant varier simultanément s et t. Plus précisément, elle représente les estimations des probabilités de démence et de décès sans démence dans les 5 années à venir pour des sujets d'âge s, s variant de 66 ans à 95 ans. Le tableau III.4 résume celles qui correspondent à des abscisses sur la courbe de 75, 85 et 95 ans. Il faut bien voir qu'en abscisse de la figure III.6, on a l'âge s, c'est-à-dire l'âge *auquel* on fait une prévision (le présent) tandis qu'en abscisse de la figure III.5, on a l'âge t, c'est-à-dire l'âge *pour lequel* on fait une prévision (le

futur). Les figures III.5 et III.6, bien qu'elles aient des abscisses différentes, coïncident quant à l'âge t pour lequel on fait la prévision. On aurait tout aussi bien pu tracer la figure III.6 avec la même abscisse que la figure III.5 en écrivant $F_{01}(t-5,t)$ et $F_{02}(t-5,t)$ en ordonnée. On constate au vu du tableau III.4 et de la figure III.6 que les estimations sont d'autant plus éloignées selon qu'on utilise les $\hat{\alpha}_{kl}(.)$ estimées avec la méthode paramétrique ou semi-paramétrique, qu'on se place à un âge s qui est grand. Par exemple, la probabilité de démence dans les 5 années à venir est de 0.55 pour une femme âgée de 95 ans lorsqu'on utilise les $\hat{\alpha}_{kl}(.)$ de type Weibull tandis qu'elle est de 0.42 lorsqu'on utilise les $\hat{\alpha}_{kl}(.)$ de type splines. Cette différence est compréhensible au vu des intensités de transition tracées à la figure III.3. Entre 95 et 100 ans, le risque instantané de démence $\alpha_{01}(.)$ (courbes noires) est plus élevé lorsqu'il est estimé avec la méthode paramétrique (partie haute) plutôt qu'avec la méthode semi-paramétrique (partie basse). On constate également que les différences hommes-femmes sont moins importantes que sur la figure III.5 mais toujours présentes.

Tableau III.4 – Modèle sans variables explicatives : estimations des probabilités cumulées $\hat{F}_{01}(s,t)$ et $\hat{F}_{02}(s,t)$ avec intervalles de confiance basées sur la méthode paramétrique (à gauche) et semi-paramétrique (à droite) chez les hommes (♂) et les femmes (♀).

		méthode paramétrique		méthode semi-paramétrique	
(s,t)		$\hat{F}_{01}$	$\hat{F}_{02}$	$\hat{F}_{01}$	$\hat{F}_{02}$
(75,80)	♂	0.09 [0.07 ;0.10]	0.19 [0.17 ;0.20]	0.09 [0.08 ;0.11]	0.18 [0.16 ;0.20]
	♀	0.10 [0.09 ;0.11]	0.09 [0.08 ;0.10]	0.10 [0.09 ;0.12]	0.08 [0.06 ;0.09]
(85,90)	♂	0.21 [0.18 ;0.24]	0.35 [0.31 ;0.38]	0.20 [0.17 ;0.24]	0.35 [0.32 ;0.39]
	♀	0.29 [0.27 ;0.31]	0.21 [0.19 ;0.23]	0.30 [0.27 ;0.33]	0.21 [0.19 ;0.24]
(95,100)	♂	0.37 [0.29 ;0.48]	0.48 [0.39 ;0.56]	0.26 [0.19 ;0.35]	0.55 [0.47 ;0.63]
	♀	0.55 [0.48 ;0.61]	0.32 [0.26 ;0.38]	0.42 [0.36 ;0.49]	0.38 [0.32 ;0.44]

Probabilités cumulées à l'horizon de la vie entière

Faisons maintenant tendre l'âge futur t vers l'infini pour estimer la probabilité d'un sujet non dément d'âge s de devenir dément au cours de sa vie, $F_{01}(s,\infty)$. Nous rappelons que la probabilité de décéder sans démence est $F_{02}(s,\infty) = 1 - F_{01}(s,\infty)$. La figure III.7 montre les estimations de $F_{01}(s,\infty)$ pour s allant de 66 à 100 ans, et le tableau III.5 en résume quelques-unes. Nous constatons qu'ici aussi, plus on avance en âge, plus les estimations diffèrent selon les estimations $\hat{\alpha}_{kl}(.)$ utilisées. De la même façon que précédemment, les probabilités issues des $\hat{\alpha}_{kl}(.)$ de type Weibull ont tendance à être plus grandes que celles issues des $\hat{\alpha}_{kl}(.)$ de type splines. Par exemple, la probabilité

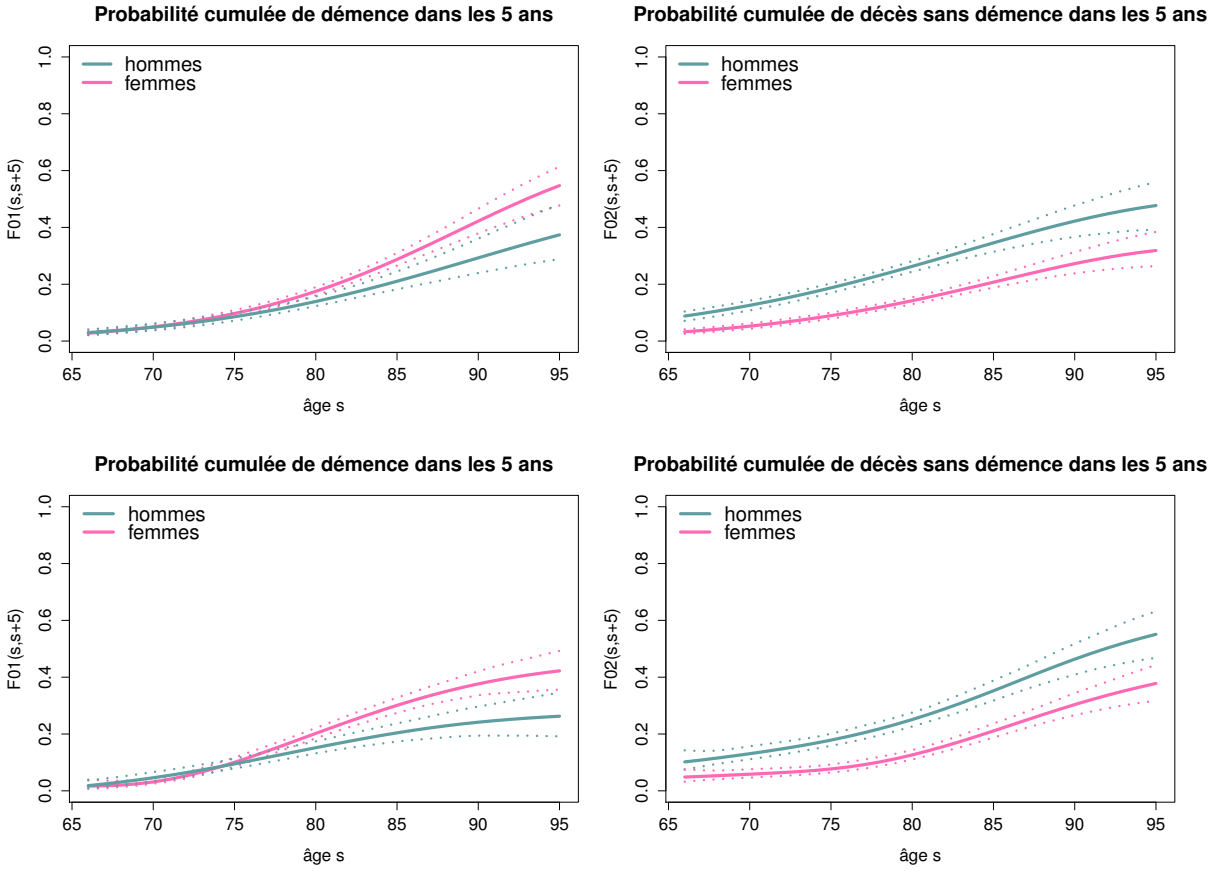


Figure III.6 – Modèle sans variables explicatives : estimations des probabilités cumulées entre s et $s+5$, $66 \leq t \leq 95$ ($\hat{F}_{01}(s, s+5)$ à gauche, $\hat{F}_{02}(s, s+5)$ à droite) avec leurs bandes de confiance, calculées avec les $\hat{\alpha}_{kl}(.)$ i) de type Weibull et estimées par MV (partie haute), ii) de type splines et estimées par MVP (partie basse).

qu'une femme non démente à 90 ans soit démente au cours de sa vie est de 0.62 si l'on en croit l'estimation paramétrique, et de 0.54 si l'on en croit l'estimation semi-paramétrique. De plus, d'après l'estimation paramétrique, les probabilités augmentent légèrement avec l'âge allant pour les hommes de 0.35 à 70 ans jusqu'à 0.42 à 90 ans (et pour les femmes de 0.56 à 70 ans jusqu'à 0.62 à 90 ans). D'après l'estimation semi-paramétrique, les probabilités restent relativement stables, autour de 0.35 pour les hommes et autour de 0.56 pour les femmes ; elles sont les plus élevées entre 75 et 80 ans.

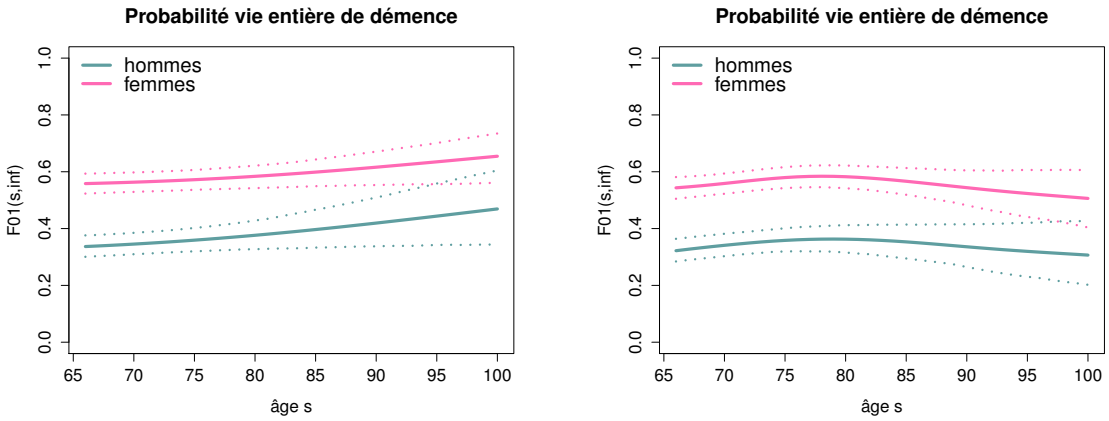


Figure III.7 – Modèle sans variables explicatives : estimations des probabilités de démence au cours de la vie entière à l'âge s, $\hat{F}_{01}(s,\infty)$, $66 \leq s \leq 100$, avec leurs bandes de confiance, calculées avec les $\hat{\alpha}_{kl}(.)$ $i)$ de type Weibull et estimées par MV (à gauche), $ii)$ de type splines et estimées par MVP (à droite).

Tableau III.5 – Modèle sans variables explicatives : estimations de $F_{01}(s,\infty)$ la probabilité de démence au cours de la vie entière (*lifetime risk of dementia*) avec intervalles de confiance basées sur la méthode paramétrique (à gauche) et semi-paramétrique (à droite).

		méthode paramétrique	méthode semi-paramétrique
s		$\hat{F}_{01}$	$\hat{F}_{01}$
70	♂	0.35 [0.31 ;0.38]	0.34 [0.30 ;0.38]
	♀	0.56 [0.53 ;0.60]	0.56 [0.52 ;0.59]
80	♂	0.38 [0.33 ;0.43]	0.36 [0.32 ;0.41]
	♀	0.58 [0.54 ;0.62]	0.58 [0.54 ;0.62]
90	♂	0.42 [0.34 ;0.51]	0.34 [0.26 ;0.42]
	♀	0.62 [0.55 ;0.67]	0.54 [0.48 ;0.60]

Espérances de vie

Intéressons-nous enfin à estimer des quantités de type espérances de vie. Sur la figure III.8 sont tracées l'espérance de vie sans démence des sujets non déments d'âge s, l'espérance de vie des sujets non déments d'âge s et l'espérance de vie des sujets déments d'âge s, avec s variant de 66 à 100 ans. Certaines valeurs sont résumées dans le tableau III.6 dans lequel figure également l'espérance de vie « moyenne » d'un sujet dément (équation III.6).

Tableau III.6 – Modèle sans variables explicatives : estimations des espérances de vie à 70, 80, 90 ans et espérance de vie moyenne d'un dément avec IC chez les hommes (♂) et les femmes (♀).

s		$\hat{\mathbb{E}}(S\|_{X(s)=0})$	$\hat{\mathbb{E}}(T\|_{X(s)=0})$	$\hat{\mathbb{E}}(T\|_{X(s)=1})$
méthode paramétrique				
70	♂	12.42 [12.02 ;12.85]	13.74 [13.39 ;14.37]	7.24 [6.48 ;10.65]
	♀	14.81 [14.44 ;15.17]	17.54 [17.20 ;17.99]	10.72 [9.66 ;12.70]
80	♂	7.09 [6.80 ;7.36]	8.20 [7.95 ;8.70]	4.27 [3.90 ;5.87]
	♀	8.13 [7.89 ;8.35]	10.39 [10.15 ;10.71]	6.15 [5.76 ;7.0]
90	♂	3.73 [3.43 ;4.03]	4.60 [4.32 ;5.07]	2.51 [2.27 ;3.51]
	♀	3.86 [3.64 ;4.09]	5.51 [5.26 ;5.85]	3.36 [3.14 ;3.78]
$\forall s$		espérance de vie moyenne des déments		
	♂		4.17 [3.72 ;5.71]	
	♀		5.15 [4.76 ;5.85]	
méthode semi-paramétrique				
70	♂	12.41 [11.89 ;12.81]	13.71 [13.20 ;14.15]	6.16 [3.88 ;8.51]
	♀	14.99 [14.52 ;15.31]	17.59 [17.10 ;17.95]	7.05 [4.44 ;9.85]
80	♂	7.12 [6.80 ;7.42]	8.26 [7.94 ;8.60]	4.25 [3.69 ;4.88]
	♀	8.02 [7.75 ;8.28]	10.43 [10.14 ;10.72]	5.99 [5.35 ;6.62]
90	♂	3.87 [3.52 ;4.19]	4.62 [4.27 ;5.02]	2.63 [2.28 ;3.08]
	♀	4.08 [3.82 ;4.33]	5.58 [5.31 ;5.89]	3.47 [3.19 ;3.76]
$\forall s$		espérance de vie moyenne des déments		
	♂		3.94 [3.39 ;4.58]	
	♀		4.73 [4.28 ;5.22]	

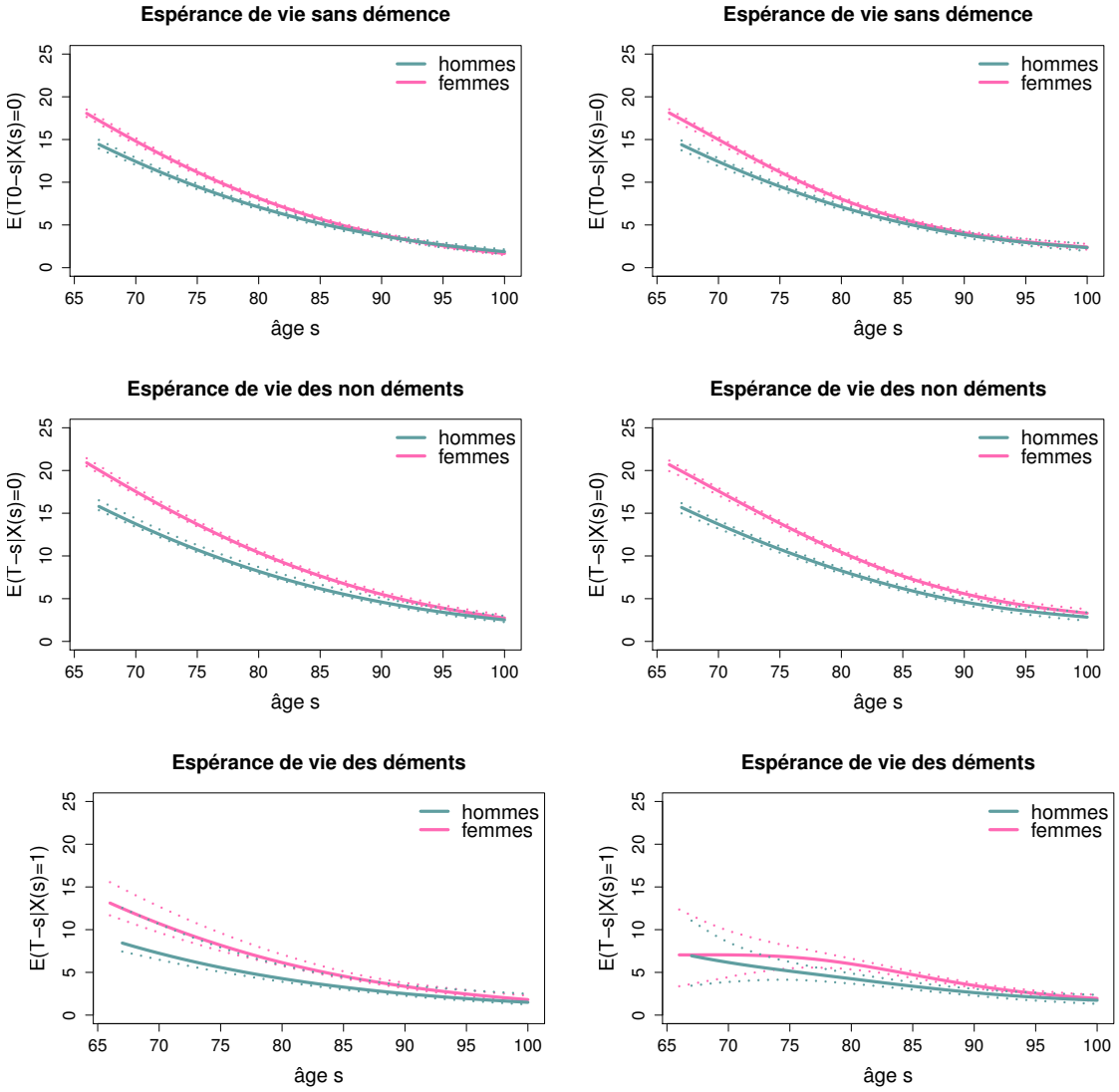


Figure III.8 – Modèle sans variables explicatives : estimations des espérances de vie à l'âge s, $66 \leq s \leq 100$, avec leurs bandes de confiance, calculées avec les $\hat{\alpha}_{kl}(.)$ i) de type Weibull et estimées par MV (à gauche), ii) de type splines et estimées par MVP (à droite). De haut en bas : espérance de vie sans démence (équation III.2), espérance de vie des non déments (équation III.3), espérance de vie des déments (équation III.4).

4.2 Modèle avec une variable explicative

Considérons maintenant un modèle avec une variable explicative, le certificat d'études primaires (cep), afin d'illustrer simplement le fait que les prévisions précédentes peuvent être faites en fonction de facteurs d'exposition. Pour voir les estimations (paramétriques et semi-paramétriques) relatives à l'effet du cep sur chaque transition, on pourra se référer au chapitre précédent (tableau II.5). En résumé, le cep apparaît comme un facteur protecteur de la démence et du décès des non déments et comme un facteur de risque du décès des déments. L'effet du cep est significatif sur les trois transitions chez les hommes et seulement sur la transition $0 \to 1$ chez les femmes. De façon attendue, les différences entre les intensités de transition correspondant au groupe de sujets n'ayant pas obtenu le cep et celles correspondant au groupe l'ayant obtenu sont les plus importantes lorsque l'effet du cep est significatif (voir figure III.9).

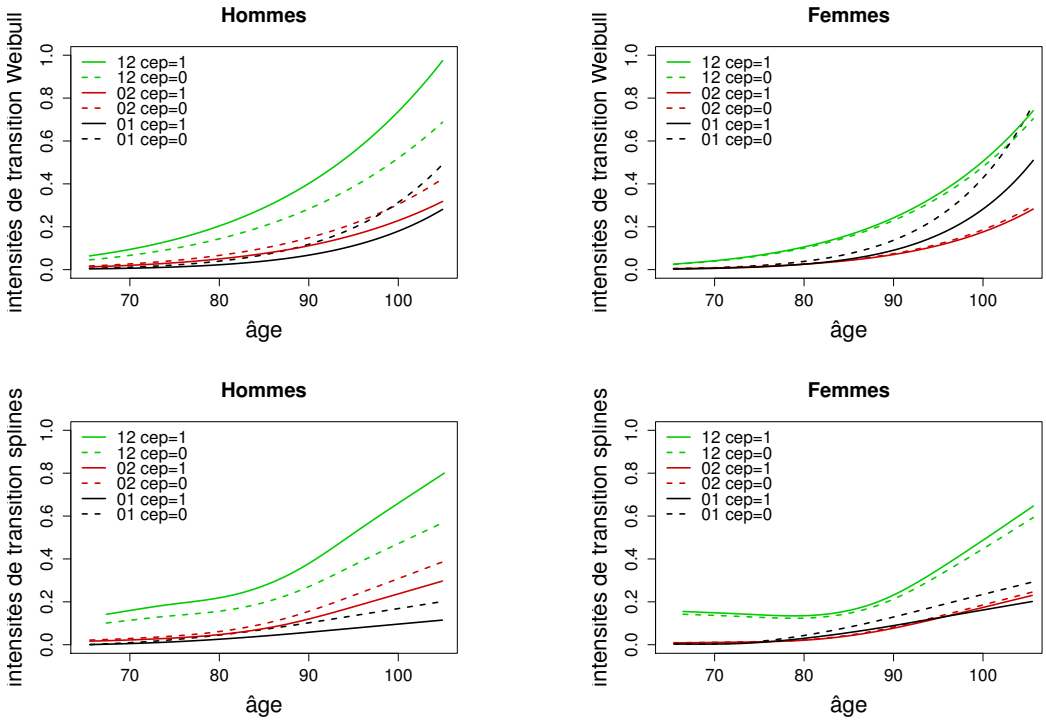


Figure III.9 – Modèle avec la variable cep : estimations des intensités de transition α_{01}, α_{02}, α_{12} *i*) de type Weibull et estimées par MV (partie haute), *ii*) de type splines et estimées par MVP (partie haute).

Nous avons estimé les mêmes quantités que précédemment lorsqu'on a considéré un modèle sans variables mais relativement au groupe d'appartenance (détention ou non du cep). Pour ne pas surcharger ce document, les tableaux et figures qui n'illustrent

pas les commentaires ont été placés dans l'annexe B. Les tableaux B.1, B.2, B.3 et B.4 résument les probabilités de transition obtenues, la figure B.1 et le tableau B.5 les probabilités cumulées à un horizon fini, la figure III.10 et le tableau III.7 les probabilités cumulées sur 5 ans, la figure B.2 et le tableau B.6 les probabilités cumulées sur la vie entière, et enfin, la figure III.11 et les tableaux III.8 et III.9 les espérances de vie.

En ce qui concerne les différences selon la méthode d'estimation utilisée pour les $\hat{\alpha}_{kl}$, les remarques qu'on peut faire sont similaires à celles déjà faites précédemment. Il en est de même pour les différences selon le sexe.

Concernant les différences entre les sujets ayant obtenu leur cep et ceux ne l'ayant pas obtenu, elles sont cohérentes avec les estimations des effets de ce facteur $\hat{\beta}_{01}$, $\hat{\beta}_{02}$, $\hat{\beta}_{12}$. L'effet du cep est plus important sur la transition $0 \to 1$ que sur les autres. Les plus grandes différences concernent donc les quantités qui dépendent de l'estimation du risque instantané de démence $\hat{\alpha}_{01}$. Par exemple, la probabilité d'une femme (resp. d'un homme) non dément(e) à 75 ans de le devenir dans les 5 années à venir est de 0.27 (resp. 0.18) si elle (il) a son cep, 0.36 (resp. 0.27) sinon (méthode semi-paramétrique, tableau III.7) ; l'espérance de vie sans démence d'une femme (resp. d'un homme) non dément(e) à 70 ans est de 15.54 ans (resp. 13.08 ans) si elle (il) a son cep, 14.07 (resp. 10.69) sinon (méthode semi-paramétrique, tableau III.9). Cette différence d'espérance de vie est plus importante chez les hommes que chez les femmes. Cela est dû au fait que le cep a un effet protecteur sur le risque de décès sans démence (transition $0 \to 2$) significatif seulement chez les hommes. De même, l'effet du cep étant significatif sur la transition $1 \to 2$ seulement chez les hommes, certaines quantités ne sont réellement différentes que chez les hommes. Par exemple, l'espérance de vie d'un homme dément à 80 ans est de 3.86 ans s'il a son cep, 5.08 ans sinon ; l'espérance de vie d'une femme démente à 80 ans est de 5.84 ans si elle a son cep, 6.23 ans sinon (méthode semi-paramétrique, tableau III.9).

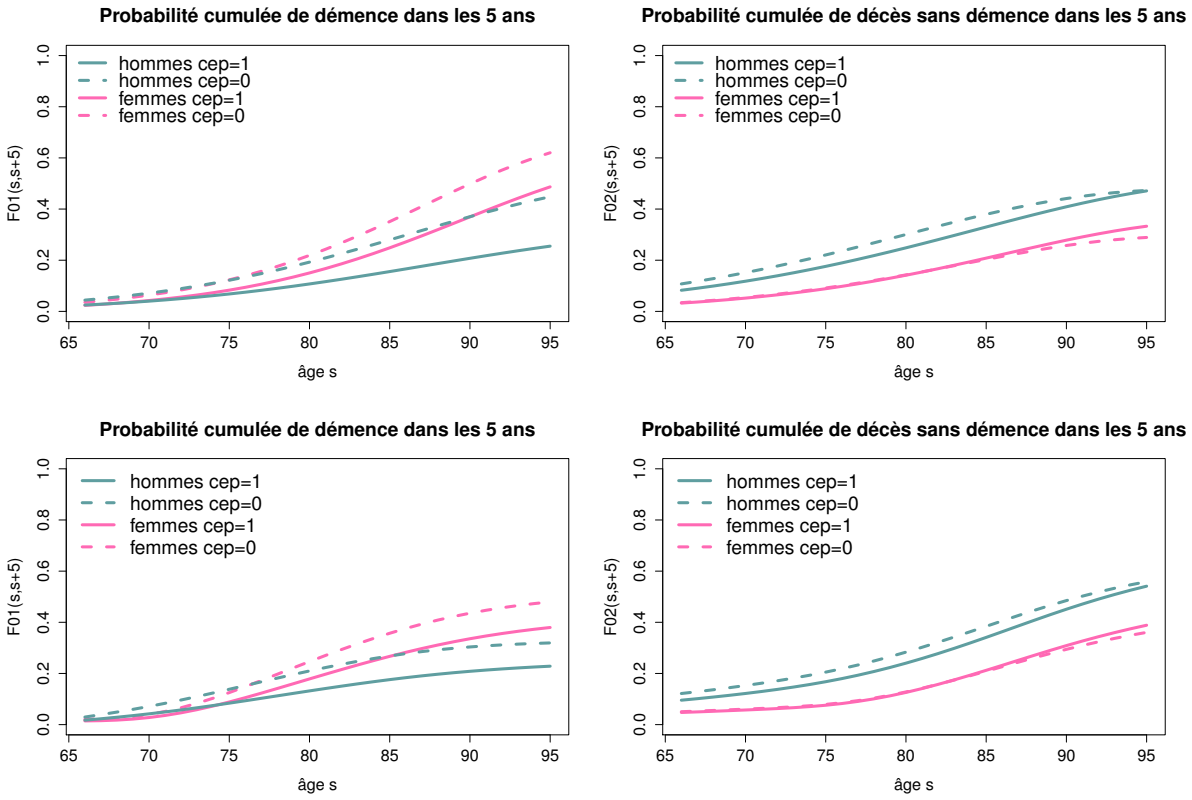


Figure III.10 – Modèle avec le cep : estimations des probabilités cumulées entre s et $s+5$, $66 \leq t \leq 95$ ($F_{01}(s, s+5)$ à gauche, $F_{02}(s, s+5)$ à droite) calculées avec les $\alpha_{kl}(.)$ i) de type Weibull et estimées par MV (partie haute), ii) de type splines et estimées par MVP (partie basse).

Tableau III.7 – Modèle avec le cep : estimations des probabilités cumulées $\hat{F}_{01}(s,t)$ et $\hat{F}_{02}(s,t)$ avec intervalles de confiance basées sur la méthode paramétrique (à gauche) semi-paramétrique (à droite).

		méthode paramétrique		méthode semi-paramétrique	
(s,t)	cep	$\hat{F}_{01}$	$\hat{F}_{02}$	$\hat{F}_{01}$	$\hat{F}_{02}$
hommes					
(75,80)	oui	0.07 [0.06 ;0.09]	0.18 [0.16 ;0.19]	0.08 [0.07 ;0.10]	0.17 [0.15 ;0.19]
	non	0.12 [0.09 ;0.15]	0.22 [0.19 ;0.25]	0.14 [0.11 ;0.18]	0.21 [0.17 ;0.24]
(85,90)	oui	0.19 [0.15 ;0.22]	0.33 [0.29 ;0.37]	0.18 [0.14 ;0.21]	0.34 [0.30 ;0.38]
	non	0.28 [0.23 ;0.34]	0.38 [0.33 ;0.43]	0.27 [0.22 ;0.32]	0.38 [0.33 ;0.44]
(95,100)	oui	0.34 [0.25 ;0.45]	0.47 [0.38 ;0.56]	0.23 [0.16 ;0.31]	0.54 [0.46 ;0.62]
	non	0.45 [0.340.56 ;]	0.47 [0.38 ;0.57]	0.32 [0.24 ;0.40]	0.56 [0.48 ;0.64]
femmes					
(75,80)	oui	0.08 [0.07 ;0.10]	0.09 [0.08 ;0.10]	0.09 [0.07 ;0.10]	0.08 [0.06 ;0.09]
	non	0.12 [0.10 ;0.14]	0.09 [0.08 ;0.11]	0.13 [0.10 ;0.15]	0.08 [0.06 ;0.10]
(85,90)	oui	0.25 [0.23 ;0.28]	0.21 [0.18 ;0.23]	0.27 [0.24 ;0.30]	0.21 [0.18 ;0.24]
	non	0.35 [0.31 ;0.39]	0.20 [.017 ;0.24]	0.36 [0.32 ;0.40]	0.21 [0.18 ;0.24]
(95,100)	oui	0.50 [0.43 ;0.57]	0.33 [0.27 ;0.40]	0.38 [0.32 ;0.45]	0.39 [0.32 ;0.46]
	non	0.62 [0.54 ;0.69]	0.29 [0.23 ;0.36]	0.48 [0.41 ;0.55]	0.36 [0.30 ;0.43]

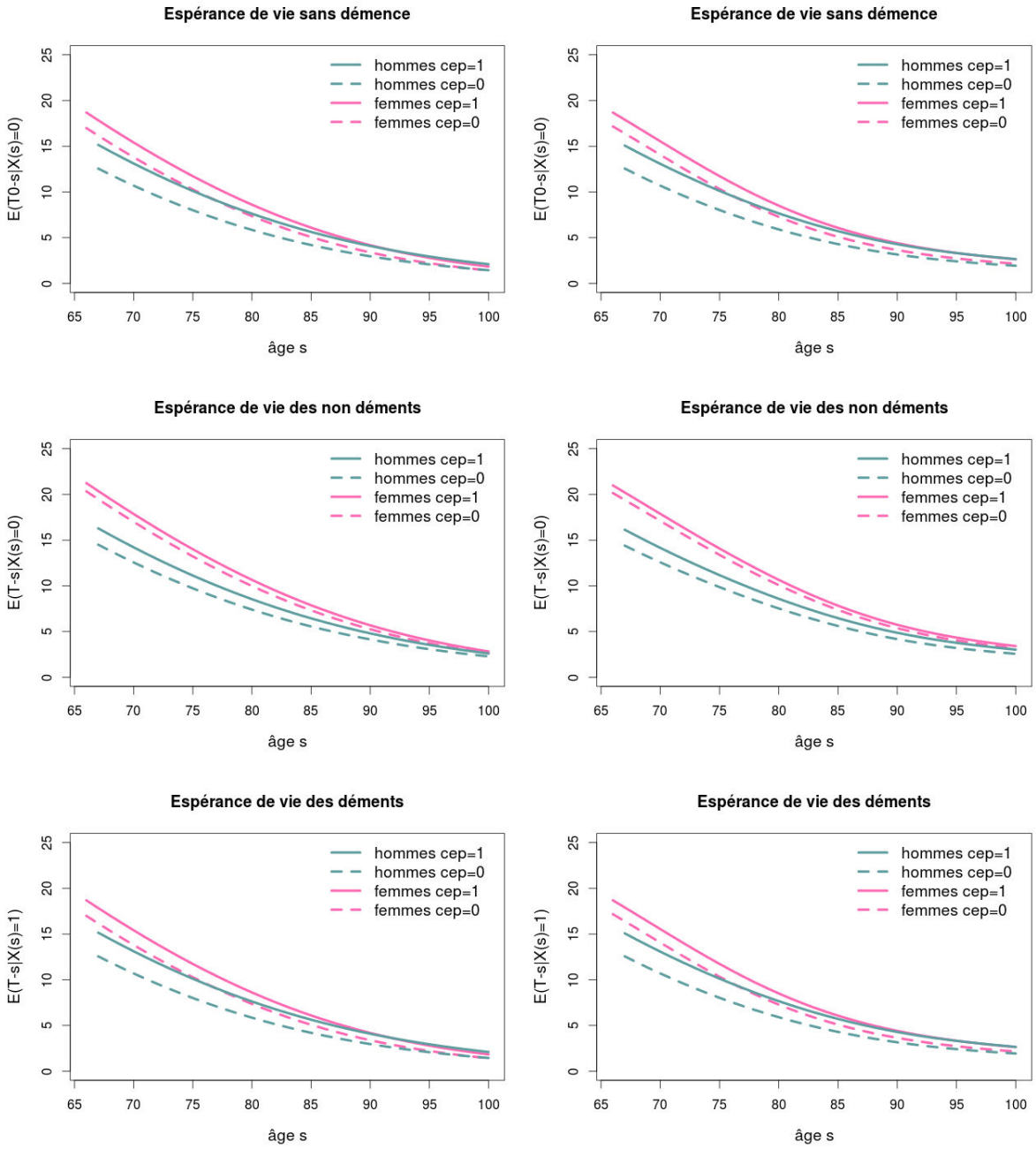


Figure III.11 – Modèle avec le cep : estimations des espérances de vie à l'âge s, $F_{01}(s,\infty)$, $66 \leq s \leq 100$ calculées avec les $\alpha_{kl}(.)$ $i)$ de type Weibull et estimées par MV (à gauche), $ii)$ de type splines et estimées par MVP (à droite). De haut en bas : espérance de vie sans démence (équation III.2), espérance de vie des non déments (équation III.3), espérance de vie des déments (équation III.4).

Tableau III.8 – Modèle avec le cep et **méthode paramétrique** : estimations des espérances de vie $\mathbb{E}\big(T_0 - s | X(s) = 0\big)$ et $\mathbb{E}\big(T_0 - s | X(s) = 0\big)$, $\mathbb{E}\big(T_0 - s | X(s) = 0\big)$ à 70, 80, 90 ans et espérance de vie moyenne d'un dément avec intervalles de confiance.

s	cep	$\hat{\mathbb{E}}(T_0\|_{X(s)=0})$	$\hat{\mathbb{E}}(T\|_{X(s)=0})$	$\hat{\mathbb{E}}(T\|_{X(s)=1})$
hommes				
70	oui	13.11 [12.68 ;13.63]	14.21 [13.78 ;14.84]	6.72 [5.73 ;9.82]
	non	10.69 [10.09 ;11.38]	12.56 [11.99 ;13.43]	8.55 [7.27 ;12.29]
80	oui	7.62 [7.28 ;7.98]	8.54 [8.22 ;9.04]	3.84 [3.39 ;5.29]
	non	5.85 [5.45 ;6.30]	7.40 [7.03 ;8.09]	5.08 [4.42 ;6.97]
90	oui	4.10 [3.76 ;4.46]	4.80 [4.47 ;5.29]	2.19 [1.90 ;3.06]
	non	2.97 [2.67 ;3.31]	4.14 [3.82 ;4.71]	2.97 [2.52 ;4.05]
$\forall s$		espérance de vie moyenne des déments		
	oui		3.61 [3.12 ;4.89]	
	non		5.45 [4.69 ;7.44]	
femmes				
70	oui	15.41 [14.97 ;15.86]	17.86 [17.43 ;18.36]	10.62 [9.40 ;12.68]
	non	13.80 [13.30 ;14.37]	17.00 [16.51 ;17.60]	10.95 [9.76 ;13.04]
80	oui	8.61 [8.30 ;8.92]	10.66 [10.32 ;11.05]	6.06 [5.51 ;7.02]
	non	7.36 [7.03 ;7.73]	9.98 [9.62 ;10.41]	6.29 [5.73 ;7.37]
90	oui	4.18 [3.92 ;4.45]	5.69 [5.37 ;6.06]	3.29 [3.00 ;3.76]
	non	3.39 [3.16 ;3.66]	5.26 [4.98 ;5.64]	3.43 [3.11 ;3.94]
$\forall s$		espérance de vie moyenne des déments		
	oui		4.88 [4.43 ;5.61]	
	non		5.60 [5.04 ;6.56]	

Tableau III.9 – Modèle avec le cep et **méthode semi-paramétrique** : estimations des espérances de vie $\mathbb{E}\big(T_0 - s|X(s) = 0\big)$ et $\mathbb{E}\big(T_0 - s|X(s) = 0\big)$, $\mathbb{E}\big(T_0 - s|X(s) = 0\big)$ à 70, 80, 90 ans et espérance de vie moyenne d'un dément avec intervalles de confiance.

s	cep	$\hat{\mathbb{E}}(T_0\|_{X(s)=0})$	$\hat{\mathbb{E}}(T\|_{X(s)=0})$	$\hat{\mathbb{E}}(T\|_{X(s)=1})$
hommes				
70	oui	13.08 [12.50 ;13.56]	14.15 [13.55 ;14.64]	5.27 [3.08 ;7.82]
	non	10.69 [9.94 ;11.43]	12.58 [11.78 ;13.31]	6.97 [4.38 ;9.87]
80	oui	7.66 [7.27 ;8.06]	8.59 [8.22 ;8.97]	3.86 [3.18 ;4.48]
	non	5.90 [5.46 ;6.38]	7.53 [7.04 ;8.06]	5.08 [4.21 ;6.20]
90	oui	4.29 [3.93 ;4.72]	4.87 [4.48 ;5.33]	2.30 [1.89 ;2.78]
	non	3.16 [2.84 ;3.53]	4.17 [3.80 ;4.64]	3.08 [2.61 ;3.87]
$\forall s$		espérance de vie moyenne des déments		
	oui		3.44 [2.86 ;4.06]	
	non		4.99 [4.09 ;6.15]	
femmes				
70	oui	15.54 [15.04 ;15.96]	17.88 [17.31 ;18.30]	6.74 [4.18 ;9.79]
	non	14.07 [13.45 ;14.59]	17.10 [16.45 ;17.65]	7.28 [4.67 ;10.32]
80	oui	8.50 [8.15 ;8.84]	10.66 [10.30 ;10.98]	5.84 [5.06 ;6.53]
	non	7.28 [6.89 ;7.69]	10.08 [9.67 ;10.51]	6.23 [5.49 ;7.06]
90	oui	4.41 [4.10 ;4.74]	5.75 [5.42 ;6.11]	3.35 [2.98 ;3.74]
	non	3.65 [3.36 ;3.96]	5.37 [5.05 ;5.75]	3.60 [3.24 ;4.09]
$\forall s$		espérance de vie moyenne des déments		
	oui		4.49 [3.97 ;5.07]	
	non		5.12 [4.54 ;5.83]	

Remarque : effet protecteur du cep sur le risque de décès des déments

Il peut sembler étonnant que le cep soit un facteur de risque pour la transition $1 \to 2$ alors qu'il est une facteur protecteur pour les transitions $0 \to 1$ et $0 \to 2$. Cela peut s'expliquer par une notion récente en épidémiologie : la réserve cognitive (Katzman et al., 1989; Stern, 2002, 2003). Les sujets ayant les plus grandes réserves cognitives seraient en capacité d'accomplir une même tâche cognitive malgré l'altération de certaines régions cérébrales en compensant par d'autres réseaux cérébraux (par exemple grâce au nombre de synapses). Les signes cliniques de la démence se manifesteraient donc plus tardivement chez les sujets ayant une réserve cognitive importante. La capacité de réserve cognitive étant fortement corrélée au niveau d'études, un niveau d'études élevé retarderait ainsi le début de l'accélération du déclin cognitif et donc du diagnostic de démence. En revanche, il a été montré que lorsqu'elle a débuté, cette accélération est d'autant plus rapide que les sujets ont une réserve cognitive importante (Hall et al., 2007). Ceci pourrait expliquer la raison pour laquelle les sujets déments d'un niveau d'études élevé (détention du cep) ont un risque de décès plus important que les sujets d'un niveau d'études bas (non détention du cep) du même âge.

Discussion

Nous avons passé en revue les principales quantités d'intérêt épidémiologique dans un modèle *illness-death* et en avons proposé de nouvelles. Leur écriture en fonction des intensités de transition permet de les estimer directement grâce aux estimations de celles-ci, et ce même en présence de censure par intervalle. Il est ainsi possible d'obtenir un large éventail de prévisions tout en faisant moins d'hypothèses qu'avec les méthodes d'estimations proposées pour des modèles multi-états généraux en présence de données observées en temps discret.

Les quantités présentées ont été estimées sur les données de Paquid en considérant un modèle de Markov non homogène à l'aide du paquet R **SmoothHazard**. Ce dernier ne permet pas à ce jour de considérer un modèle de semi-Markov homogène, mais l'écriture de la vraisemblance et des différentes quantités dans un tel modèle a également été donnée. Commenges et al. (2007) ont considéré un modèle prenant en compte à la fois l'âge courant et le temps passé dans l'état 1 à travers un modèle additif : $\alpha_{12}(t, t - T_0) = h(t) + g(t - T_0)$ comme dans Scheike (2001). Ils ont montré que celui-ci ne fournit pas un meilleur critère de validation croisée (qui estime le critère de Kullback-

Leibler) qu'un simple modèle de Markov sur Paquid pour un suivi de 13 ans. Il aurait cependant pu être intéressant de comparer les résultats d'un modèle de Markov et d'un modèle prenant en compte à la fois l'âge courant et le temps passé dans l'état 1 sur Paquid avec un suivi de 20 ans. Par exemple, en considérant un modèle de semi-Markov homogène, nous aurions pu introduire l'âge de démence comme variable explicative sur le risque de décès des déments (*i.e.* sur la transition $1 \to 2$). Dans un schéma d'observations censurées par intervalle, cette variable serait un peu particulière car elle correspondrait à la variable u d'intégration dans les équations de vraisemblance III.9 et III.10. De même dans un modèle de Markov, nous aurions pu introduire la variable correspondant au temps passé en l'état 1. En présence de censure par intervalle, elle correspondrait à $\tilde{T} - u$ dans les équations de la vraisemblance du chapitre I (plus exactement à $C - u$ dans les équations I.9 et I.11 et à $T - u$ dans les équations I.10 et I.12). Une autre solution plus simple aurait été d'imputer l'âge de démence ou le temps passé dans l'état 1 : pour un sujet dément en considérant le milieu de l'intervalle entre la visite de diagnostic et la visite précédente, pour une sujet décédé sans diagnostic de démence en considérant le milieu de l'intervalle entre la dernière visite et le décès.

L'application sur les données de Paquid permet d'illustrer et d'apprécier les différentes prévisions qu'il est possible de faire mais n'est pas très approfondie d'un point de vue épidémiologique. Des prévisions issues d'un modèle avec une variable explicative binaire ont été montrées de façon à comparer les prévisions du groupe exposé par rapport à celles du groupe non exposé. Mais il pourrait être intéressant de prendre en compte un plus grand nombre de variables explicatives, y compris continues et donc d'obtenir des prévisions non plus spécifiques à un groupe de sujet mais à un sujet particulier.

Dans tous les résultats que l'on a présentés, nous avons mis en parallèle les prévisions issues d'intensités de transition estimées avec une méthode paramétrique et celles issues d'intensités de transition estimées avec une méthode semi-paramétrique. Nous avons remarqué que sur les données de Paquid, bien que les allures des courbes n'étaient pas vraiment similaires, les quantités d'intérêt étaient relativement proches, excepté pour celles faisant intervenir les fonctions d'intensités de transition à des âges élevés. De manière générale, la méthode semi-paramétrique est probablement celle qui fournit les meilleures estimations mais elle a certains inconvénients. Tout d'abord, le temps de calcul est beaucoup plus long, et ce d'autant plus qu'il y a de variables explicatives. L'inconvénient majeur est celui des données disponibles lorsqu'on s'intéresse à des quantités pour lesquelles la fenêtre de prévision est celle de la vie entière ($t \to \infty$). En

effet, les nœuds des splines définissant les intensités de chaque transition sont placés sur l' intervalle dans lequel se situent les durées observées. En général, le dernier nœud est placé au dernier temps d'observation ou légèrement plus loin. Il est donc impossible de faire une prévision à un horizon dépassant le dernier nœud. Ainsi, lorsqu'on souhaite estimer des quantités à un horizon infini (espérances de vie, *lifetime risk*), il faut supposer qu'au temps correspondant au dernier nœud, tous les sujets sont décédés (dans l'état absorbant 2). Paquid étant une cohorte de grande envergure avec un temps de suivi très long, la méthode semi-paramétrique a pu être utilisée mais il est important de remarquer qu'elle ne peut l'être avec n'importe quelles données. Avec la méthode paramétrique, il est techniquement possible d'estimer les quantités d'intérêt à n'importe quel horizon. Nous préconisons cependant de ne pas faire de prévisions à des horizons lointains lorsque le temps de suivi est court. En effet, les paramètres des lois régissant les intensités de transition ont été estimés sur la base des durées observées. En dehors des temps d'observation, les intensités de transition sont donc plutôt des extrapolations que des estimations.

Enfin, la technique de calcul d'intervalles de confiance est basée sur l'hypothèse de normalité asymptotique des estimateurs. La normalité asymptotique des estimateurs de MV est bien connue ; celle des estimateurs de MVP l'est moins. Des résultats théoriques ont été montrés, notamment par O'Sullivan (1988b), mais nous aurions pu vérifier par simulations le comportement asymptotique attendu dans notre contexte.

Chapitre IV

Modèle *illness-death* avec effets aléatoires

En analyse de survie, des modèles à fragilité qui sont une extension des modèles à risques proportionnels permettent de tenir compte de l'hétérogénéité d'un échantillon. Des effets aléatoires servent à prendre en compte des facteurs non mesurés, étroitement liés à l'évènement d'intérêt et partagés par les sujets d'un même groupe homogène. Ce chapitre est dédié à l'introduction d'effets aléatoires, spécifiques au groupe d'appartenance, dans le modèle *illness-death* pour données censurées par intervalle. Ce travail a été motivé à la fois par une application à la cohorte Paquid et par une application à la cohorte 3C. En effet, les sujets de Paquid appartenant à la même zone géographique pourraient partager un certain nombre de facteurs qui impacteraient les risques de démence et de décès. De même pour les sujets en couple de la cohorte 3C.

Introduction

Paquid

Comme nous l'avons mentionné précédemment, les sujets de Paquid (n=3675) sont répartis sur 75 communes, et on peut supposer que les sujets d'une même commune partagent des facteurs (par exemple environnementaux) impactant les risques de démence et de décès.

Rondeau et al. (2003) ont précédemment étudié l'« effet commune » sur Paquid en utilisant un modèle à fragilité ajusté sur l'âge (âge choisi comme temps de base) et sur le sexe (modèle stratifié sur le sexe). Ils n'ont cependant pas tenu compte de la censure par intervalle en imputant l'âge de démence entre la visite de diagnostic et la visite précédente (au milieu de l'intervalle). Ils n'ont surtout pas tenu compte du risque de décès en censurant à droite à leur dernière visite les sujets décédés sans diagnostic de démence. Nous avons vu, en particulier dans le chapitre II, qu'il est important de ne

pas ignorer le fait que ces sujets ont pu devenir déments entre leur dernière visite et leur décès lorsqu'on estime l'incidence et les facteurs de risque de la démence. Dans ce travail, nous proposons d'introduire des effets aléatoires à travers des modèles à fragilité sur les trois transitions d'un modèle *illness-death* pour données censurées par intervalle.

3 Cités (3C)

L'étude de cohorte 3C, initiée en 1999, est constituée de 9294 sujets âgés de 65 ans ou plus recrutés dans les villes de Dijon (n=4931), Bordeaux (n=2104) et Montpellier (n=2259), d'où le nom de 3 Cités. Son objectif général est d'analyser de manière approfondie la relation entre les facteurs de risque vasculaire et la démence. Les sujets ont été examinés à l'inclusion puis à 2, 3, 7 et 10 ans de suivi. À 5 ans, ils ont rempli un questionnaire mais le statut dément/non dément n'a pas été évalué. Le diagnostic de démence s'est fait à Dijon de façon similaire à l'étude Paquid : les enquêteurs ont recueilli les résultats de tests cognitifs, et les sujets susceptibles de présenter une démence ont été examinés par un médecin qui a posé ou non un diagnostic de démence. À Montpellier, tous les sujets ont été examinés par un neurologue quelles que soient leurs performances aux tests cognitifs. À Bordeaux, le protocole était analogue à celui de Montpellier à l'inclusion, et à celui de Dijon lors des visites de suivi. Enfin, un panel d'experts extérieur à l'étude a validé ou infirmé le diagnostic.

Une des particularités de l'étude 3C est de contenir 3440 sujets en couple (donc 1720 couples). Bien que l'objectif de 3C ne soit pas d'étudier un éventuel « effet couple » sur le risque de démence ou de décès, il nous a semblé intéressant de l'explorer. On peut en effet supposer que les sujets en couple partagent un certain nombre de facteurs (par exemple liés au style de vie) qui impactent le(s) risque(s) de démence et/ou de décès.

1 Modèle

1.1 Description

Nous considérons toujours le même modèle *illness-death* irréversible (figure III.1) dont X est le processus markovien sous-jacent. $\forall t \geq 0$, $X(t) \in \{0, 1, 2\}$ où 0 correspond à l'état non dément, 1 à l'état dément et 2 à l'état décédé. L'âge est choisi comme temps de base et les sujets sont recrutés à 65 ans ou plus à la condition d'être non déments,

i.e. à la condition $X(a) = 0$ où a est l'âge d'entrée dans la cohorte. Nous sommes donc en présence de données tronquées à gauche puisque les sujets ont survécu en l'état 0 jusqu'à leur âge d'entrée dans la cohorte. Nous supposons que les trois intensités de transition dépendent de l'âge (processus non homogène), de variables explicatives Z_{kl} (facteurs individuels) et d'effets aléatoires U_{kl} (facteurs partagés par les sujets appartenant à un même groupe). Plus précisément, sur chaque transition nous posons un modèle à fragilité partagée :

$$\begin{aligned}
\alpha_{01,ij}(t|U_{01,i}) &= \alpha_{01,0}(t)e^{\beta_{01}^T Z_{01,ij}+U_{01,i}} \\
\alpha_{02,ij}(t|U_{02,i}) &= \alpha_{02,0}(t)e^{\beta_{02}^T Z_{02,ij}+U_{02,i}} \\
\alpha_{12,ij}(t|U_{12,i}) &= \alpha_{12,0}(t)e^{\beta_{12}^T Z_{12,ij}+U_{12,i}}
\end{aligned} \tag{IV.1}$$

où i est l'indice du groupe, j l'indice du sujet. Nous supposons que les trois effets aléatoires sont distribués normalement et, pour plus de généralité, qu'ils ne sont pas indépendants :

$$\begin{pmatrix} U_{01,i} \\ U_{02,i} \\ U_{12,i} \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\Sigma) \qquad \text{avec} \qquad \Sigma = \begin{pmatrix} \sigma_{01}^2 & \sigma_{01,02} & \sigma_{01,12} \\ \sigma_{01,02} & \sigma_{02}^2 & \sigma_{02,12} \\ \sigma_{01,12} & \sigma_{02,12} & \sigma_{12}^2 \end{pmatrix}$$

Les paramètres à estimer sont les paramètres de régression β_{kl} et les 6 paramètres de variance et de covariance de la matrice symétrique Σ.

Plus grande est la variance σ_{kl}^2, plus grande est la variabilité inter-groupes associée à la transition $k \to l$. Plus grande est la covariance $\sigma_{kl,k'l'}$ entre deux effets, plus nombreux sont les facteurs spécifiques au groupe qui agissent à la fois sur la transition $k \to l$ et sur la transition $k' \to l'$. Une covariance positive signifie que les facteurs communs aux deux transitions agissent dans le même sens (d'une augmentation ou d'une diminution du risque de transition). À l'inverse, une covariance négative signifie que lorsque les facteurs communs augmentent le risque associé à l'une des transitions, ils diminuent le risque associé à l'autre transition.

1.2 Significativité des effets aléatoires

Pour tester la significativité de l'effet aléatoire U_{12} par exemple, l'hypothèse de test s'écrit : « $H_0 : \sigma_{12}^2 = \sigma_{01,12} = \sigma_{02,12} = 0$ ». On remarque que la valeur testée 0 se situe à la frontière de l'espace des paramètres puisque le paramètre de variance σ_{12}^2 est toujours positif. Un test classique de rapport de vraisemblance basé sur une

distribution asymptotique du χ^2 n'est pas applicable. En fait, la statistique de test de rapport de vraisemblance :

$$LR = -2ln\left(\frac{L(\theta_0)}{L(\hat{\theta})}\right)$$

(où $\hat{\theta}$ est le vecteur des paramètres estimés et θ_0 le vecteur des paramètres sous H_0) suit un mélange de distributions de χ^2 (Verbeke and Molenberghs, 2009).

Plus exactement, lorsque l'hypothèse nulle consiste à tester la présence de $q+1$ effets par rapport à la présence de q effets dans le modèle, la statistique de test LR suit un mélange de distributions de χ^2 à q et à $q+1$ degrés de liberté de poids égaux 0.5. La p-valeur s'écrit :

$$p = \frac{1}{2}\mathbb{P}(\chi^2_q > LR) + \frac{1}{2}\mathbb{P}(\chi^2_{q+1} > LR)$$

où χ^2_q et χ^2_{q+1} sont des variables distribuées selon une loi de χ^2 à respectivement q et $q+1$ degrés de liberté.

Si l'on souhaite tester la présence de q *vs* $q+k$ effets aléatoires dans le modèle ($k>1$), des méthodes de simulation sont nécessaires pour approcher la distribution de LR sous l'hypothèse nulle.

1.3 Vraisemblance

Soit θ le vecteur des paramètres à estimer *i.e.* le vecteur des paramètres de régression β_{kl} et des paramètres de variance et covariance de la matrice Σ.

La contribution à la vraisemblance du groupe i s'écrit :

$$L_i\Big(\theta|u_{01,i}, u_{02,i}, u_{12,i}, X(a_i) = 0_b ig) = \frac{L_i(\theta|u_{01,i}, u_{02,i}, u_{12,i})}{\mathbb{P}\Big(X(a_i = 0)|u_{01,i}, u_{02,i}, u_{12,i}\Big)}$$

où $u_{01,i}$, $u_{02,i}$, $u_{12,i}$ sont les réalisations des effets aléatoires pour le groupe i et a_i est le vecteur des âges d'entrée dans la cohorte des sujets du groupe i. $\mathbb{P}\Big(X(a_i) = 0|u_{01,i}, u_{02,i}, u_{12,i}\Big)$ est la probabilité que tous les sujets du groupe i aient survécu en l'état 0 jusqu'à leur âge d'entrée dans la cohorte (condition de troncature à gauche) conditionnellement aux effets aléatoires. Utilisons la notation $tr(\theta|u_{01,i}, u_{02,i}, u_{12,i})$ pour désigner cette probabilité de façon à indiquer sa dépendance avec le vecteur des paramètres θ.

La vraisemblance est le produit sur chaque groupe des contributions marginales à

la vraisemblance :

$$L = \prod_{i=1}^{I} \frac{\iiint L_i(\theta|u_{01,i}, u_{02,i}, u_{12,i}) \ \mathrm{d}F(u_{01,i}, u_{02,i}, u_{12,i})}{\iiint tr(\theta|u_{01,i}, u_{02,i}, u_{12,i}) \ \mathrm{d}F(u_{01,i}, u_{02,i}, u_{12,i})}$$

où I est le nombre de groupes et F la fonction de répartition associée au vecteur $(U_{01}, U_{02}, U_{12})^T$.

Comme les sujets au sein de chaque groupe sont supposés indépendants conditionnellement aux effets aléatoires, on obtient :

$$L = \prod_{i=1}^{I} \frac{\iiint \prod_{j=1}^{n_i} L_{ij}(\theta|u_{01,i}, u_{02,i}, u_{12,i}) \ \mathrm{d}F(u_{01,i}, u_{02,i}, u_{12,i})}{\iiint \prod_{j=1}^{n_i} tr(\theta|u_{01,i}, u_{02,i}, u_{12,i}) \ \mathrm{d}F(u_{01,i}, u_{02,i}, u_{12,i})}$$

où n_i est l'effectif du groupe i, L_{ij} la contribution individuelle du sujet j appartenant au groupe i (conditionnellement aux effets aléatoires), a_{ij} l'âge d'entrée dans la cohorte du sujet j appartenant au groupe i.

En développant, on obtient :

$$L = \prod_{i=1}^{I} \frac{\iiint \prod_{j=1}^{n_i} L_{ij}(\theta|u_{01,i}, u_{02,i}, u_{12,i}) f(u_{01,i}, u_{02,i}, u_{12,i}) \ \mathrm{d}u_{01,i}\mathrm{d}u_{02,i}\mathrm{d}u_{12,i}}{\iint \prod_{j=1}^{n_i} tr(\theta|u_{01,i}, u_{02,i}) f(u_{01,i}, u_{02,i}) \ \mathrm{d}u_{01,i}\mathrm{d}u_{02,i}} \tag{IV.2}$$

où la fonction f désigne au numérateur la fonction de densité conjointe du vecteur $(U_{01}, U_{02}, U_{12})^T$ et au dénominateur celle du vecteur $(U_{01}, U_{02})^T$. L'effet aléatoire associé à la transition $1 \to 2$ n'intervient pas dans le dénominateur car la probabilité de survie en l'état 0 ne dépend que des risques associés aux transitions $0 \to 1$ et $0 \to 2$:

$$\begin{aligned}
tr(\theta|u_{01,i}, u_{02,i}, u_{12,i}) &= \mathbb{P}(X(a_{ij} = 0)|u_{01,i}, u_{02,i}, u_{12,i}) \\
&= \exp\left\{-A_{ij,01}(a_{ij}|u_{01,i}, u_{02,i}, u_{12,i}) - A_{ij,02}(a_{ij}|u_{01,i}, u_{02,i}, u_{12,i})\right\} \\
&= \exp\left\{-\int_0^{a_{ij}} \alpha_{ij,01}(v|u_{01,i})\mathrm{d}v - \int_0^{a_{ij}} \alpha_{ij,02}(v|u_{02,i})\mathrm{d}v\right\} \\
&= \mathbb{P}(X(a_{ij} = 0)|u_{01,i}, u_{02,i}) \\
&= tr(\theta|u_{01,i}, u_{02,i})
\end{aligned}$$

Pour connaître le détail de l'écriture de L_{ij}, on se référera à l'annexe C. Une autre façon d'aboutir à l'écriture ci-dessus de la vraisemblance est aussi donnée dans cette annexe.

Jusqu'à maintenant, nous avons analysé séparément les hommes et les femmes. Or, cela n'est pas possible pour l'une des applications de ce chapitre, dont l'enjeu est d'étudier l'effet couple dans la cohorte 3C. De plus, diminuer la taille de l'échantillon peut entraîner une perte de puissance. Enfin, concernant l'application à la cohorte Paquid, nous voulons comparer nos résultats avec ceux de Rondeau et al. (2003) qui ont utilisé un modèle de survie stratifié sur le sexe. Nous avons donc stratifié le modèle *illness-death* sur le sexe, c'est-à-dire que nous avons supposé que les intensités de transition de base étaient différentes chez les hommes et les femmes mais que les paramètres de régression et les paramètres associés aux effets aléatoires étaient identiques pour les deux sexes.

Notons θ' le vecteur des paramètres qui sont supposés égaux chez les hommes et les femmes : $\theta' = (\beta_{01}, \beta_{02}, \beta_{12}, \sigma^2_{01}, \sigma^2_{02}, \sigma^2_{12}, \sigma_{01,02}, \sigma_{01,12}, \sigma_{02,12})^T$. Notons $\alpha_{0,1}$ et $\alpha_{0,2}$ les vecteurs des paramètres associés aux intensités de transition chez les hommes et les femmes respectivement.

La vraisemblance s'écrit :

$$L = \prod_{i=1}^{I} \frac{\iiint \prod_{k=1}^{2} \prod_{j=1}^{n_{ik}} L_{ij}(\theta', \alpha_{0,k} | u_{01,i}, u_{02,i}, u_{12,i}) f(u_{01,i}, u_{02,i}, u_{12,i}) \, \mathrm{d}u_{01,i} \mathrm{d}u_{02,i} \mathrm{d}u_{12,i}}{\iint \prod_{k=1}^{2} \prod_{j=1}^{n_{ik}} tr(\theta', \alpha_{0,k} | u_{01,i}, u_{02,i}) f(u_{01,i}, u_{02,i}) \, \mathrm{d}u_{01,i} \mathrm{d}u_{02,i}}$$

où k est l'indice correspondant à la strate et n_{ik} le nombre de sujets dans le groupe i et la strate k.

1.4 Estimation

La méthode d'estimation choisie est paramétrique : les intensités de transition de base sont spécifiées par des lois de Weibull et la vraisemblance maximisée. Dans le cas le plus général où l'on suppose que l'on a trois effets aléatoires différents et non indépendants, les paramètres à estimer sont : les douze paramètres des lois de Weibull $\alpha_{0,1}$ et $\alpha_{0,2}$ (six pour les hommes, six pour les femmes), les vecteurs des paramètres de régression β_{01}, β_{02}, β_{12} et les six paramètres de variance et de covariance de la matrice Σ.

1.5 Approximation numérique des intégrales

Les intégrales sur les effets aléatoires faisant intervenir des densités à loi normale, nous utilisons la quadrature de Gauss-Hermite pour les approximer. Cette méthode consiste à approximer des intégrales de la forme

$$\int_{-\infty}^{+\infty} \mathrm{e}^{-x^2} f(x)\mathrm{d}x$$

par la somme

$$\sum_{i=1}^{n_{GH}} \omega_i f(x_i)$$

où les points x_i correspondent aux racines d'un polynôme de Hermite d'ordre n_{GH} et où les ω_i sont des poids adaptés.

1.5.1 Cas particulier : effets aléatoires indépendants

Plaçons-nous d'abord dans le cas particulier où l'on suppose que les effets aléatoires du modèle sont indépendants, c'est-à-dire que la matrice Σ est diagonale. De cette façon, la vraisemblance s'écrit :

$$L = \prod_{i=1}^{I} \frac{\iiint \prod_{k=1}^{2}\prod_{j=1}^{n_{ik}} L_{ij}(\theta|u_{01,i},u_{02,i},u_{12,i}) f_{01}(u_{01,i}) f_{02}(u_{02,i}) f_{12}(u_{12,i})\ \mathrm{d}u_{01,i}\mathrm{d}u_{02,i}\mathrm{d}u_{12,i}}{\iint \prod_{k=1}^{2}\prod_{j=1}^{n_{ik}} \mathbb{P}\Big(X(a_{ij})=0|u_{01,i},u_{02,i}\Big) f_{01}(u_{01,i}) f_{02}(u_{02,i})\ \mathrm{d}u_{01,i}\mathrm{d}u_{02,i}}$$

où f_{01}, f_{02} et f_{12} sont les densités de probabilité associées aux effets aléatoires U_{01}, U_{02} et U_{12}.

Dans ce cas particulier, on a trois distributions normales univariées. La méthode de quadrature de Gauss-Hermite peut donc être appliquée directement.

Soit n_{GH} le nombre de points choisis pour la quadrature. La vraisemblance est approchée par :

$$L \approx \prod_{i=1}^{I} \frac{\sum_{z=1}^{n_{GH}}\sum_{y=1}^{n_{GH}}\sum_{x=1}^{n_{GH}} \prod_{k=1}^{2}\prod_{j=1}^{n_{ik}} L_{ij}(\theta|u_{01,i}=x,u_{02,i}=y,u_{12,i}=z) f_{01}(x) f_{02}(y) f_{12}(z)}{\sum_{y=1}^{n_{GH}}\sum_{x=1}^{n_{GH}} \prod_{k=1}^{2}\prod_{j=1}^{n_{ik}} \mathbb{P}\Big(X(a_{ij})=0|u_{01,i}=x,u_{02,i}=y\Big) f_{01}(x) f_{02}(y)} \tag{IV.3}$$

1.5.2 Cas général : effets aléatoires non indépendants

Dans le cas où les effets aléatoires U_{01}, U_{02} et U_{12} ne sont pas indépendants, la méthode de quadrature de Gauss-Hermite ne peut être appliquée directement. Cependant, la factorisation de Cholesky permet de se ramener au cas précédent.

La matrice Σ est symétrique et définie positive (puisque c'est une matrice de variance covariance) donc d'après la factorisation de Cholesky, il existe une matrice triangulaire inférieure B telle que :

$$\Sigma = BB^T$$

Soit $(\tilde{U}_{01}, \tilde{U}_{02}, \tilde{U}_{12})^T$ un vecteur de variables normales centrées réduites :

$$\begin{pmatrix} \tilde{U}_{01} \\ \tilde{U}_{02} \\ \tilde{U}_{12} \end{pmatrix} \sim \mathcal{N}(0, I) \qquad \text{avec } I \text{ la matrice identité.}$$

En appliquant le changement de variable : $U = L\tilde{U}$, la vraisemblance en IV.2 peut se réécrire comme suit :

$$L = \prod_{i=1}^{I} \frac{\iiint \prod_{k=1}^{2} \prod_{j=1}^{n_{ik}} L_{ij}(\theta|\tilde{u}_{01,i}, \tilde{u}_{02,i}, \tilde{u}_{12,i}) f_{01}(\tilde{u}_{01,i}) f_{02}(\tilde{u}_{02,i}) f_{12}(\tilde{u}_{12,i}) \, \mathrm{d}\tilde{u}_{01,i} \mathrm{d}\tilde{u}_{02,i} \mathrm{d}\tilde{u}_{12,i}}{\iint \prod_{k=1}^{2} \prod_{j=1}^{n_{ik}} \mathbb{P}\Big(X(a_{ij}) = 0 | \tilde{u}_{01,i}, \tilde{u}_{02,i}\Big) f_{01}(\tilde{u}_{01,i}) f_{02}(\tilde{u}_{02,i}) \, \mathrm{d}\tilde{u}_{01,i} \mathrm{d}\tilde{u}_{02,i}}$$

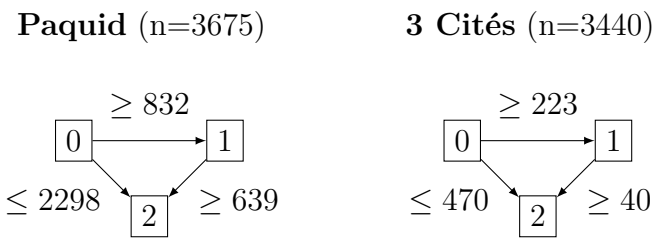
La méthode de quadrature de Gauss-Hermite peut alors être appliquée comme dans la sous-section 1.5.1 et les estimations des six valeurs de la matrice triangulaire B permettent de calculer les estimations des paramètres de covariance de la matrice $\Sigma = BB^T$.

2 Application

Le modèle proposé a été appliqué sur deux jeux de données issus des cohortes Paquid et 3C. Les résultats présentés ont été calculés sur un suivi de 20 ans pour Paquid et sur un suivi de 7 ans pour 3C (les données relatives à un temps de suivi plus long n'étant pas encore disponibles). Les nombres de démences (transitions $0 \to 1$), de décès de non déments (transitions $0 \to 2$) et de décès de déments observés dans chaque cohorte

sont visibles sur la figure IV.1. Les nombres exacts de transitions ne sont pas connus en raison de la censure par intervalle (d'où les signes $\leq$ et $\geq$).

Figure IV.1 – Nombre de transitions pour Paquid avec un suivi de 20 ans et pour 3 Cités avec un suivi de 7 ans.



2.1 Effet commune dans Paquid

La première illustration consiste à intégrer des effets aléatoires dans le modèle *illness-death* pour données censurées par intervalle appliqué aux données de Paquid afin de modéliser la présence supposée de facteurs partagés par les individus appartenant à la même commune. Le nombre de points utilisés pour la quadrature de Gauss est de 20. En effet, nous avons testé successivement 5, 7, 9, 15 et 20 points et il n'y avait quasiment aucune différence dans les estimations selon qu'on utilisait 15 ou 20 points.

Nous avons d'abord considéré un modèle stratifié sur le sexe avec un effet aléatoire sur la transition $0 \to 1$. À la différence de Rondeau et al. (2003), nous avons pris en compte le décès et la censure par intervalle et nous avons utilisé une méthode d'estimation paramétrique. Sans variables explicatives, la variance de l'effet a été estimée par $\hat{\sigma}^2_{01} = 0.0021$ tandis qu'elle était estimée à 0.018 dans Rondeau et al. (2003) et, comme dans Rondeau et al. (2003) l'effet est significatif avec une p-valeur de 0.02 (tandis qu'elle était de 0.045).

Nous avons aussi considéré un modèle avec un seul effet aléatoire sur la transition $0 \to 2$ et un autre avec un seul effet aléatoire sur la transition $1 \to 2$. Ces effets aléatoires sur le risque de décès n'étaient pas significatifs. De même, l'apport d'un effet aléatoire sur le risque de décès ($0 \to 2$ ou $1 \to 2$) en plus d'un effet aléatoire sur le risque de démence ($0 \to 1$) n'était pas significatif, et l'apport d'un troisième non plus. Cependant, dans ce dernier modèle (où l'on modélise 3 effets aléatoires non indépendants), la valeur estimée de la variance de l'effet sur la transition $0 \to 1$, $\hat{\sigma}^2_{01}$,

est beaucoup plus importante que dans les modèles avec un ou deux effets aléatoires. En effet, la matrice de variance-covariance des trois effets aléatoires a été estimée par :

$$\hat{\Sigma} = \begin{pmatrix} 0.0112 & -0.0055 & -0.0081 \\ -0.0055 & 0.0035 & 0.0037 \\ -0.0081 & 0.0037 & 0.0060 \end{pmatrix}$$

Nous avons ensuite ajouté au modèle stratifié avec un seul effet aléatoire sur la transition $0 \to 1$ le cep en variable explicative sur les trois transitions. La variance a été estimée par $\hat{\sigma}^2_{01} = 0.0009$ et n'était pas significativement non nulle ($p = 0.17$). Dans Rondeau et al. (2003), elle avait été estimée par 0.0001 et n'était pas non plus significativement non nulle ($p = 0.35$).

Nous avons également considéré deux modèles séparés pour les hommes et les femmes au lieu d'un modèle stratifié. Sans variables explicatives et avec un effet aléatoire sur $0 \to 1$, on obtient une variance de $\hat{\sigma}^2_{01} = 0.0215$ pour les hommes et de $\hat{\sigma}^2_{01} = 0.0011$ pour les femmes. De plus, l'effet est significatif chez les hommes ($p = 0.003$) mais pas chez les femmes ($p = 0.26$). Lorsqu'on introduit le cep sur les trois transitions, on obtient une variance de $\hat{\sigma}^2_{01} = 0.0102$ pour les hommes et l'effet aléatoire n'est plus significatif ($p = 0.056$).

Conclusion

Sur les données de Paquid, l'introduction sur la transition $0 \to 1$ d'un effet aléatoire partagé par les sujets d'une même commune nous conduit à penser que les sujets appartenant à la même commune partagent des facteurs ayant un effet significatif sur le risque de démence. En analysant séparément l'échantillon des hommes et l'échantillon des femmes au lieu d'utiliser un modèle stratifié sur le sexe, nous nous sommes rendu compte que cet effet commune n'est en fait significatif que pour les hommes. La prise en compte du niveau d'études dans le modèle (cep en variable explicative) divise par deux l'estimation de la variance de l'effet aléatoire qui n'est plus significativement non nulle. Cependant, la p-valeur est encore très proche de 5 % sur l'échantillon des hommes.

2.2 Effet couple dans 3 Cités

La deuxième illustration consiste à intégrer des effets aléatoires dans le modèle appliqué aux données des 3440 sujets en couple de 3C. Nous avons utilisé 9 points de

quadrature de Gauss-Hermite après avoir usé du même procédé qu'avec la précédente application pour choisir le nombre de points.

Présumant que les facteurs partagés au sein d'un couple auraient surtout un impact sur le risque de décès, nous avons d'abord considéré un modèle *illness-death* pour données censurées par intervalle stratifié sur le sexe avec un effet aléatoire sur la transition $0 \to 2$ (décès des non déments). Nous obtenons une estimation de la variance de l'effet de $\sigma^2_{02} = 0.4627$ qui est significativement non nulle ($p = 0.032$).

En ajoutant au modèle un effet aléatoire sur la transition $0 \to 1$ (démence), nous avons obtenu comme estimation de la variance-covariance des effets aléatoires :

$$\hat{\Sigma} = \begin{pmatrix} 0.0708 & 0.1772 \\ 0.1772 & 0.4435 \end{pmatrix}$$

L'ajout dans le modèle d'un effet aléatoire sur $0 \to 1$ n'est cependant pas significatif ($p = 0.13$).

Nous n'avons pas pu modélisé d'effet aléatoire sur le risque de décès des déments car sur les 40 transitions $1 \to 2$ observées, il y a seulement 1 couple.

3 Simulations

Des simulations de 3000 sujets ont été réalisées avec un seul effet aléatoire sur la transition $0 \to 1$, U_{01}. Nous avons fait varier la variance de l'effet aléatoire σ^2_{01} ainsi que le nombre de sujets par groupe. Nous avons considéré 9 scénarios correspondant à :

— des valeurs de 0.5, 0.1 et 0.05 pour σ^2_{01} ;

— 300 groupes de 10 sujets, 600 groupes de 5 sujets, 1500 groupes de 2 sujets.

Le but était de comparer les estimations de σ^2_{01} ainsi que les p-valeurs produites par le test de significativité de l'effet aléatoire.

Pour chaque scénario, 1000 jeux de données ont été générés. Les simulations ont été faites dans le même esprit que celles du chapitre II, de façon à se conformer autant que possible aux données de Paquid. Les âges correspondant aux transitions d'un état à l'autre ont été générés selon trois lois de Weibull. Les paramètres choisis pour ces lois étaient ceux estimés sur Paquid. Une variable distribuée uniformément a été utilisée pour les âges de censure à droite. Les âges associés à la transition $0 \to 1$ étaient censurés par intervalle. Une variable binaire d'effets -0.4, -0.2 et 0.2 sur respectivement

Tableau IV.1 – Résultats obtenus sur 1000 simulations. $moy(\hat{\sigma}^2_{01})$: moyenne des estimations de σ^2_{01} ; $var(\hat{\sigma}^2_{01})$: variance des estimations de σ^2_{01} ; % $p < 0.05$: pourcentage de p-valeurs <0.05 (pourcentage de simulations pour lesquelles le test conduit à un effet significatif de U_{01}).

		300 groupes de 10	600 groupes de 5	1500 groupes de 2
$\sigma^2_{01} = 0.5$	$moy(\hat{\sigma}^2_{01})$	0.547	0.529	0.512
	$var(\hat{\sigma}^2_{01})$	0.0078	0.0065	0.0140
	% $p < 0.05$	100%	100%	100%
$\sigma^2_{01} = 0.1$	$moy(\hat{\sigma}^2_{01})$	0.1004	0.1013	0.1000
	$var(\hat{\sigma}^2_{01})$	0.00108	0.00168	0.00438
	% $p < 0.05$	98.5%	86%	45.6%
$\sigma^2_{01} = 0.05$	$moy(\hat{\sigma}^2_{01})$	0.0496	0.0494	0.0530
	$var(\hat{\sigma}^2_{01})$	0.00073	0.00113	0.00286
	% $p < 0.05$	63.9%	40.2%	18.9%

les transitions $0 \to 1$, $0 \to 2$ et $1 \to 2$ a également été générée afin de reproduire l'effet du cep.

Le temps de calcul étant relativement long, ces simulations ont été réalisées en parallèle avec le logiciel R, c'est-à-dire qu'au lieu de s'exécuter de façon séquentielle sur un seul processeur, elles ont été réalisées en partie simultanément sur plusieurs processeurs (souvent 250 ou 500) à l'aide du mésocentre de calcul intensif aquitain (voir http://cran.r-project.org/web/views/HighPerformanceComputing.html).

Le tableau IV.1 résume les résultats des simulations. Nous avons calculé pour chaque scénario la moyenne des estimations de l'effet aléatoire $moy(\hat{\sigma}^2_{01})$, la variance des estimations de l'effet aléatoire $var(\hat{\sigma}^2_{01})$ et le pourcentage de simulations pour lesquelles la statistique de test de rapport de vraisemblance nous conduirait à conclure à la significativité de l'effet aléatoire (% $p < 0.05$).

Si l'on lit le tableau de gauche à droite, on remarque que pour un même nombre de sujets, plus on augmente le nombre de groupes (et donc le nombre de réalisations de l'effet aléatoire), plus l'estimation de la variance de l'effet aléatoire est proche de la vraie valeur. Cependant, cela ne paraît être vrai qu'à la condition que la variance soit suffisamment importante (voir les deux premières lignes du tableau qui correspondent à des variances $\sigma^2_{01} = 0.5$ et $\sigma^2_{01} = 0.1$). On remarque également que plus on augmente le nombre de groupes (et donc plus on diminue le nombre de sujets au sein de chaque groupe), plus on augmente la variance de l'estimation. On remarque enfin que la capacité du test à détecter un effet aléatoire significatif se dégrade, du moins lorsque la variance de l'effet est faible (voir les deux dernières lignes du tableau qui correspondent

à des variances $\sigma^2_{01} = 0.1$ et $\sigma^2_{01} = 0.05$).

Si l'on lit maintenant le tableau de haut en bas, on remarque que le fait de baisser la variance de l'effet aléatoire entraîne une diminution du pourcentage de fois où celui-ci est détecté comme étant significatif.

4 Conclusion et discussion

Les modèles à fragilité sont largement utilisés dans des modèles de survie classiques. Nous avons étendu leur utilisation à un modèle *illness-death* qui prend en compte les données censurées par intervalle.

Leur intérêt a été illustré par deux applications.

Nous avons d'abord prolongé le travail de Rondeau et al. (2003) qui ont étudié la présence de facteurs de risque de démence spécifiques à la commune sur les données de Paquid, en prenant en compte la censure par intervalle et le risque de décès des déments et des non déments. À l'inverse de Rondeau et al. (2003), nous avons utilisé une méthode d'estimation paramétrique. En effet, nous avons observé dans les chapitre II et III que la méthode paramétrique donne des résultats similaires à la méthode semi-paramétrique sur les données de Paquid, notamment en ce qui concerne l'estimation des paramètres de régression (*i.e.* des effets des facteurs de risque associés aux trois transitions). De plus, la différence de temps de calcul entre les deux méthodes d'estimation est plus importante dans un modèle *illness-death* que dans un modèle de survie, et, très probablement bien plus importante dans un modèle *illness-death* avec effets aléatoires (car accentuée par des calculs d'intégrales).

Cette application a mis en évidence (en particulier chez les hommes) la présence de facteurs de risque de démence partagés par les sujets d'une même commune. La prise en compte du niveau d'études dans l'analyse a diminué de moitié la variance de l'effet aléatoire. Il serait intéressant dans une prochaine étape d'introduire dans le modèle l'exposition au silice et à l'aluminium dans l'eau du robinet de chaque commune. Si comme dans Rondeau et al. (2003) la variance de l'effet aléatoire diminue largement nous pourrions alors considérer que la prise en compte du niveau d'études et de l'exposition au silice et à l'aluminium dans l'eau du robinet suffisent à « capter » la dépendance intra-communes et, donc, qu'il est inutile de rechercher d'autres facteurs spécifiques à la commune qui auraient un effet sur le risque de démence.

Cependant, dans cette application, des effets aléatoires spatiaux auraient peut-être

été plus adaptés que des effets aléatoires non spatiaux afin de prendre en compte la disposition géographique des communes (Banerjee et al., 2003; Cressie, 1993; Gamerman, 1991).

Sur les données de 3C, nous avons mis en évidence le fait que les risques de décès des sujets non déments d'un même couple sont corrélés. Le modèle proposé suppose que cette corrélation est expliquée par le fait que les sujets d'un même couple partagent un certain nombre de facteurs qui ont un effet sur le risque de décès. Il est toutefois légitime de se demander si la corrélation inter-couples ne serait pas plutôt d'ordre temporel, c'est-à-dire si le décès d'un sujet en couple ne précipiterait pas le décès de son conjoint. De même, lorsqu'on ajoute un effet aléatoire sur le risque de démence en plus de l'effet aléatoire sur le risque de décès des non déments, on peut se demander si le fait que l'estimation de la covariance entre les deux effets soit plus importante que celle de la variance de l'effet sur le risque de démence ne serait pas la conséquence d'une corrélation d'ordre temporel. Peut-être le décès d'un sujet en couple précipite-t-il la démence de son conjoint (et/ou la survenue de la démence d'un sujet en couple précipite-t-elle le décès de son conjoint) ? De plus, il a été montré que le fait d'être marié minore le risque de démence (Dartigues et al., 2012).

Bien que l'effet aléatoire associé au risque de décès soit significatif, il paraît ainsi difficile de soutenir que le modèle proposé capte bien les effets de facteurs partagés par les sujets d'un même couple sur le risque de décès, et non la modification du risque de décès d'un sujet à partir de la date de décès de son conjoint.

La possibilité d'introduire des effets aléatoires dans le modèle *illness-death* n'a pas été ajoutée au paquet R. La raison principale est que nous avons rencontré des problèmes numériques avec les données utilisées. En particulier, il est arrivé que pour certains groupes i, la valeur des L_{ij} soit très faible et donc que le produit sur j des L_{ij} soit nul. Sans effets aléatoires, ce problème ne se pose car on calcule la log-vraisemblance et le logarithme d'un produit de contributions à la vraisemblance peut être transformé en somme des logarithmes des contributions. La présence d'effets aléatoires induit des sommes sur les points de quadrature qui empêchent le passage au logarithme pour calculer le produit $\prod_{j=1}^{n_i} L_{ij}$ (voir équation IV.3). Selon le jeu de données, le calcul de la vraisemblance a donc dû être adapté. Ce problème s'est également posé pour certains scénarios de simulations non présentés ici.

Enfin, les simulations présentées ont permis en particulier de remarquer que lorsqu'on simule un effet aléatoire dont la variance est faible, celle-ci est correctement

estimée mais la significativité de l'effet ne peut être mise en évidence avec un test de rapport de vraisemblance. Il est donc possible que dans l'application sur Paquid, il existe des facteurs spécifiques à la commune dont l'effet soit trop faible pour être détecté.

Chapitre V

Paquet **R SmoothHazard**

Il nous a semblé important que le modèle *illness-death* et les méthodes d'estimations présentés dans cette thèse puissent être utilisés par d'autres. C'est pourquoi une partie importante de cette thèse a été dédiée à la construction du paquet R **SmoothHazard**. Un article visant à expliquer son utilisation plus amplement que dans le présent chapitre est disponible à l'annexe A.

Introduction

Des paquets R pour l'analyse de modèles multi-états, et donc pour modèles *illness-death*, existent. Le paquet **mstate** (Beyersmann et al., 2011; De Wreede et al., 2010, 2011) peut être utilisé lorsque les transitions sont observées en temps continu. Il fournit des estimations pour les intensités de transition cumulées et pour les probabilités de transition. Le paquet R **TPmsm** (Meira-Machado and Roca-Pardiñas, 2011) permet quant à lui d'estimer de façon non paramétrique les probabilités de transition pour des modèles *illness-death* et des modèles progressifs à 3 états non markoviens. Mais lorsque l'on étudie la démence à partir de données de cohorte, les temps d'apparition de la démence sont censurés par intervalle. Or, nous avons vu dans le chapitre II que le fait d'ignorer la présence de censure par intervalle dans un contexte *illness-death* peut induire des biais dans l'estimation de l'incidence de la démence et des effets des facteurs de risque.

Le paquet **msm** (Jackson, 2011) peut être utilisé pour des schémas d'observation en temps discret. Il fournit des estimations des intensités de transition et des probabilités de transition. Les intensités de transition y sont supposées constantes ou constantes par morceaux (entre deux temps d'observation consécutifs) et sont estimées par résolution des équations différentielles de Kolmogorov. Un ensemble de fonctions R permet d'es-

timer des espérances de vie dans un modèle *illness-death* à partir d'un modèle estimé avec **msm** (van den Hout, 2013).

Cependant, le modèle *illness-death* pour données censurées par intervalle est un modèle multi-état simple dont l'écriture de la vraisemblance est explicite et il est plus intéressant de recourir à des méthodes d'estimation basées sur la vraisemblance et fournissant des estimations plus flexibles des intensités de transition. Dans ce contexte, le paquet **SmoothHazard** permet d'estimer les intensités de transition de façon paramétrique par maximum de vraisemblance en spécifiant les intensités de transition de base par des lois de Weibull ou de façon semi-paramétrique, par maximum de vraisemblance pénalisée, en approchant les estimateurs des intensités de transition de base par des M-splines. Ces méthodes sont décrites en détail dans le chapitre I. Des modèles à intensités de transition proportionnelles permettent d'introduire des variables explicatives sur chaque transition. Ce paquet fournit également des estimations pour les quantités d'intérêt présentées dans le chapitre III : probabilités de transition, probabilités cumulées, espérances de vie.

Pour une vue d'ensemble des paquets R pour l'analyse de survie en général et les modèles multi-états et *illness-death* en particulier, on pourra se référer à la page `http://cran.r-project.org/web/views/Survival.html`.

Dans ce chapitre, nous commençons par expliquer l'architecture d'un paquet R en général, puis celle du paquet **SmoothHazard**. Enfin, nous présentons et illustrons ses principales fonctionnalités.

1 Architecture...

1.1 ... d'un paquet R

Un paquet R est un ensemble de programmes permettant d'augmenter les fonctionnalités du logiciel R. Il est constitué de fichiers et de répertoires, tous réunis dans un répertoire racine portant le nom du paquet. Plus précisément, il contient :

- un fichier nommé `DESCRIPTION` : sert à décrire le paquet (titre, auteurs, version, dépendances (si le paquet fait appel à d'autres paquets), etc.) ;
- un fichier optionnel nommé `NAMESPACE` : sert à définir la visibilité des fonctions du paquet et des autres paquets (public/privé, visible/invisible) ;
- le répertoire `/R/` : contient les codes R du paquet (fichiers d'extension `.R`) ;
- le répertoire optionnel `/data/` : contient les jeux de données ;

— le répertoire `/man/` : contient les fichiers d'aide (d'extension `.Rd`) des fonctions qui peuvent être appelées et des jeux de données qui peuvent être chargés par l'utilisateur – cela lui permet de faire afficher dans sa session R les différents fichiers d'aide grâce à la commande `help` – ;
— le répertoire optionnel `/src/` : contient les codes écrits dans des langages de programmation autre que du R (`C`, `FORTRAN`, etc.).

La figure V.1 montre la structure du paquet **SmoothHazard**.

1.2 ... de SmoothHazard

Le paquet **SmoothHazard** a été créé dans l'optique de rendre accessible à la communauté scientifique l'utilisation de modèles *illness-death* pour données censurées par intervalle. Il a été développé avec l'aide d'Amadou Diakité et en collaboration avec Thomas Gerds de l'université de Copenhague. Il est basé sur des programmes existants écrit en `FORTRAN 77`. Le travail a consisté à :

— « nettoyer » ces programmes puis les convertir en `FORTRAN 90` ;
— construire le paquet : insérer les fichiers `FORTRAN` dans le répertoire `src`, remplir les fichiers `DESCRIPTION`, `NAMESPACE` et les dossiers `/R/`, `/data/`, `/man/`.

En particulier, le répertoire `/R/` contient les deux fonctions principales du paquet :

— `idm` (pour *illness-death model*) ;
— `shr` (pour *smooth hazard regression*).

Elles sont appelées par l'utilisateur pour estimer respectivement les paramètres d'un modèle *illness-death* et d'un modèle de survie à partir de données possiblement censurées par intervalle. Ces fonctions servent d'interface entre l'utilisateur et les fichiers `FORTRAN`.

Le répertoire `/R/` contient également des fonctions qui permettent de faire des prévisions dans un modèle *illness-death* (voir le chapitre III pour une illustration). En effet, l'objet retourné par la fonction `idm` peut être utilisé comme argument dans les fonctions :

— `predict` qui estime les probabilités de transition et les probabilités cumulées entre deux temps donnés ;
— `lifexpect` qui estime les espérances de vie à un temps donné.

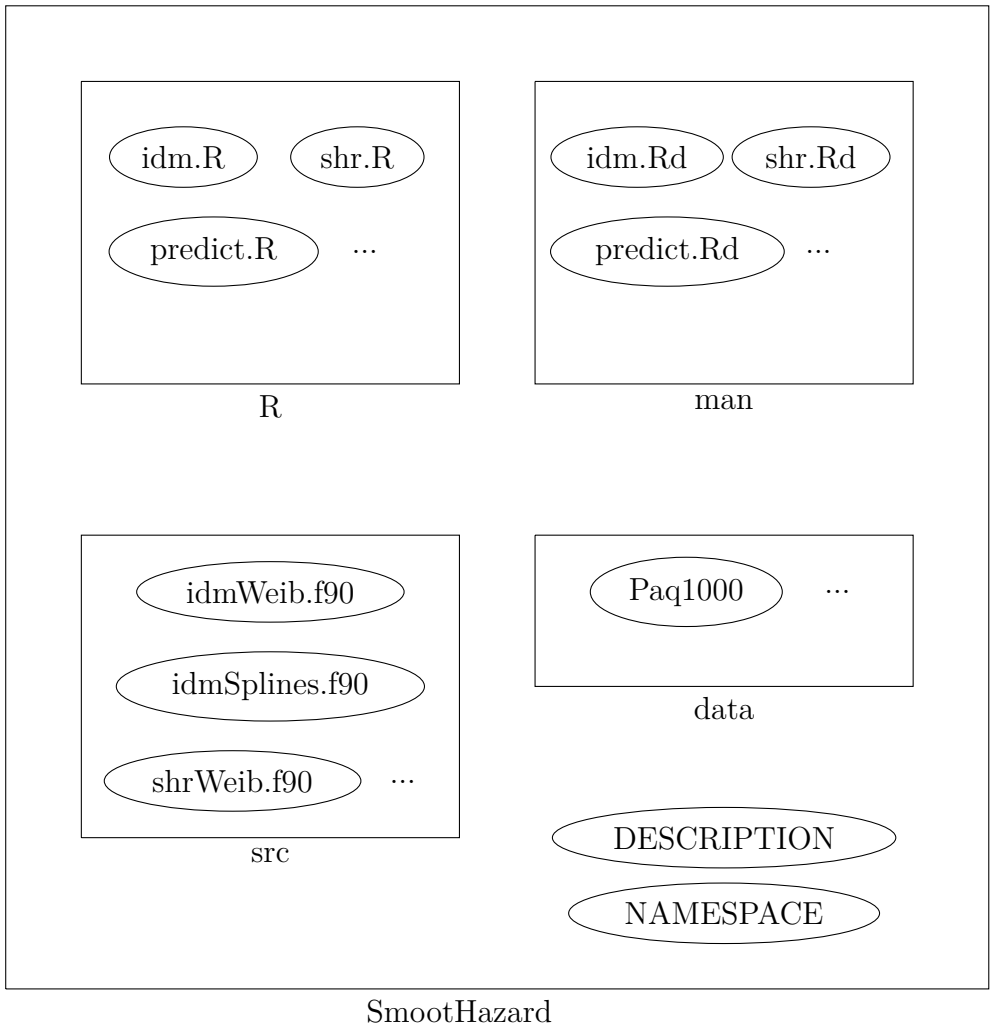


Figure V.1 – Structure du paquet **SmoothHazard**. Les rectangles représentent les répertoires dont le nom est en-dessous ; les ellipses représentent les fichiers dont le nom est à l'intérieur. « Paq1000 » est un jeu de données constitué de 1000 sujets tirés aléatoirement dans la cohorte Paquid. « idm.R », « shr.R » et « predict.R » sont des fonctions que l'utilisateur peut appeler et « idm.Rd », « shr.Rd » et « predict.Rd » sont les fichiers d'aide correspondant. « idm.R » (resp. « shr.R ») est la fonction à utiliser lorsqu'on veut considérer un modèle *illness-death* (resp. un modèle de survie) ; si la méthode d'estimation choisie est paramétrique, il fait appel à la fonction FORTRAN « idmWeib.f90 » (resp. « shrWeib.f90 ») tandis que si la méthode d'estimation choisie est semi-paramétrique, il fait appel à la fonction FORTRAN « idmSplines.f90 » (resp. « shrSplines.f90 »).

2 Utilisation

Le paquet **SmoothHazard** est disponible sur le site du CRAN (http://cran.r-project.org/web/packages/SmoothHazard/index.html). Il est aussi possible d'installer la version la plus récente à partir du site R-forge, via la commande R :

```
> install.packages("SmoothHazard",repos="http://R-Forge.R-project.org")
```

Après avoir chargé le paquet à l'aide de la commande `library(SmoothHazard)`, on peut utiliser la fonction `idm` pour estimer les paramètres d'un modèle *illness-death* sur des données possiblement censurées par intervalle. Les fichiers d'aide contiennent des exemples d'appel utilisant le jeu de données `Paq1000`. Les 1000 lignes de `Paq1000` correspondent à une sélection aléatoire de 1000 sujets de la cohorte Paquid. La table peut être chargée grâce à la commande `data(Paq1000)`. Les colonnes correspondent à l'indicateur de démence (`dementia`), l'indicateur de décès (`death`), l'âge d'entrée dans la cohorte (`e`), les bornes gauche et droite des âges de censure par intervalle (`l` et `r`), les variables sexe (`gender`) et le certificat d'études primaires (`certif`).

```
> head(Paq1000)
  dementia death        e         l         r         t  certif gender
1        1     1  72.3333  82.34014  84.73303  87.93155       0      0
2        0     1  77.9167  78.93240  78.93240  79.60048       0      1
3        0     1  79.9167  79.91670  79.91670  80.92423       0      0
4        0     1  74.6667  78.64750  78.64750  82.93501       1      1
5        0     1  76.6667  76.66670  76.66670  79.16636       0      1
6        0     0  66.2500  71.38070  71.38070  84.16975       1      0
```

Ci-après un exemple d'appel :

```
> fit.splines <- idm(formula02=Hist(time=t,event=death,entry=t0)~certif,
                     formula01=Hist(time=list(l,r),event=dementia)~certif,
                     formula12=~1,
                     intensities="Splines",
                     data=Paq1000)
```

Les arguments `formula02`, `formula01` et `formula12` doivent contenir les formules de régression associées respectivement aux risques de décès et de démence des non déments et au risque de décès des déments. La fonction `Hist` crée un objet utilisé

comme variable réponse dans les formules. Elle prend en entrée la variable des temps de transition (sous forme de liste si les temps sont censurés par intervalle), et la variable d'indicateur de l'évènement associé à la transition. La partie gauche de la formule associée à la transition $1 \to 2$ ne nécessite pas d'être remplie. La partie droite qui sert à indiquer les variables explicatives nécessite d'être remplie si l'on souhaite que les variables associées à la transition $1 \to 2$ ne soient pas les mêmes que celles associées à la transition $0 \to 2$ (affectation par défaut).

L'option `intensities` indique par le biais de la forme spécifiant les intensités de transition le type de méthode d'estimation :

- `"Weib"` (méthode semi-paramétrique) : maximum de vraisemblance où les intensités de transition de base sont paramétrées selon des lois de Weibull (valeur par défaut) ;
- `"Splines"` (méthode paramétrique) : maximum de vraisemblance pénalisée où les intensités de transition de base sont approchées par des M-splines.

Les mêmes méthodes sont disponibles avec l'option `hazard` de la fonction `shr` qui estime les paramètres d'un modèle de survie (dans le cas particulier d'un modèle de survie, l'intensité de transition est plutôt appelée fonction de risque).

Les autres options de `idm` et `shr` (par exemple concernant les choix des nœuds des splines ou du paramètre de lissage avec la méthode semi-paramétrique) sont plus amplement détaillées dans l'annexe A.

Un objet retourné par la fonction `idm` peut être utilisé pour faire des prévisions, comme dans le chapitre III. La commande :

```
> pred <- predict(fit.splines,s=70,t=80,Z01=c(1),Z02=c(1))
```

sert à calculer les quantités $p_{01}(s,t)$, $p_{12}(s,t)$, $p^0_{02}(s,t)$, $p^1_{02}(s,t)$, $F_{01}(s,t)$, $F_{0\bullet}(s,t)$ entre les âges $s = 70$ ans et $t = 80$ ans pour un sujet qui a le cep.

Aux arguments `Z01`, `Z02` et `Z12` doivent être affectés des vecteurs qui contiennent les valeurs des variables explicatives sur chaque transition, rangées dans le même ordre que dans l'appel de `idm` qui a produit l'objet utilisé (ici `fit.splines`). Par défaut, la valeur 0 est attribuée à toutes les variables explicatives.

La commande :

```
LE <- lifexpect(fit,s=85,Z01=c(1),Z02=c(1),CI=FALSE)
```

sert à calculer l'espérance de vie sans maladie et l'espérance de vie d'un sujet non malade de 75 ans qui a le cep, ainsi que l'espérance de vie d'un sujet malade de 75 ans qui a le cep.

2.1 Discussion

Le modèle *illness-death* est l'un des modèles multi-états les plus utilisés. La création du paquet **SmoothHazard** et sa diffusion sur le CRAN a rendu possible l'analyse de données censurées par intervalle dans un modèle *illness-death* en utilisant une méthode d'estimation semi-paramétrique initialement proposée dans Joly et al. (2002) ou une méthode paramétrique, toutes deux basées sur la vraisemblance. Il permet également d'estimer une large palette de quantités d'intérêt (voir chapitre III). La carte de la figure V.2 donne une idée du nombre de téléchargements de **SmoothHazard** dans le monde.

Dans l'avenir, des améliorations du paquet sont envisagées. En particulier, il devrait être possible de :

— considérer un modèle de semi-Markov homogène ;
— considérer le cas particulier d'un modèle à deux risques concurrents ;
— estimer les intensités de transition de façon non paramétrique lorsque les temps de transition sont observés de façon exacte (pas de censure par intervalle).

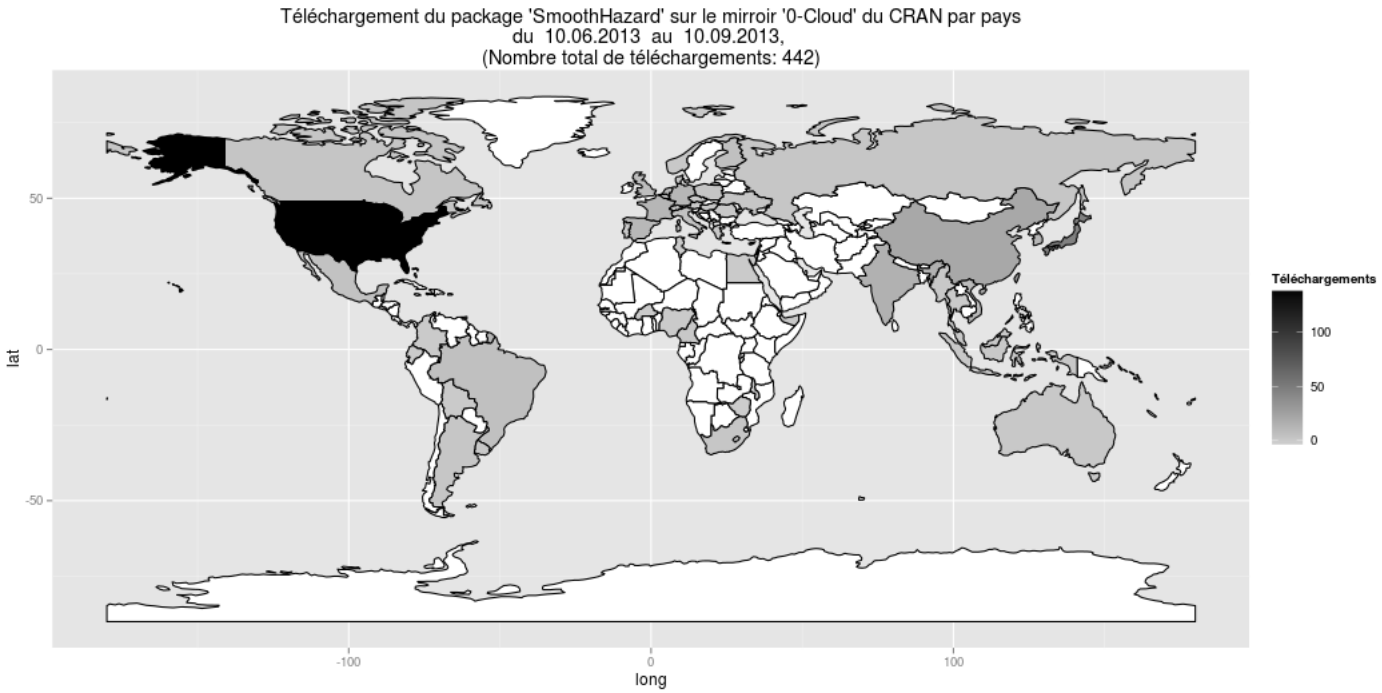


Figure V.2 – Carte représentant le nombre de téléchargement de **SmoothHazard** à partir du miroir '0-Cloud' du CRAN sur une période de temps allant du 10/06/2013 au 10/09/2013.

Conclusion générale et perspectives

Récapitulatif du travail de thèse

Cette thèse a été motivée par l'étude de la démence sénile à partir de données de cohorte. Le fait que les sujets soient suivis par intermittence donne lieu à des données censurées par intervalle : lorsqu'un sujet est observé dément, la date d'apparition de la maladie n'est pas connue exactement mais comprise entre la visite de diagnostic et la visite précédente. À cela vient s'ajouter un risque important de décès. La principale conséquence est que lorsqu'un sujet décède alors qu'il a été observé non dément à sa dernière visite, on ne connaît pas sa trajectoire : il a pu décéder directement ou devenir dément puis décéder.

Dans ce contexte, les techniques classiques d'analyse de survie ne s'appliquent plus. Il en est d'ailleurs de même des techniques pour modèles multi-états avec des données observées en temps continu : la vraisemblance ne se factorisant plus, les différentes transitions ne peuvent être traitées séparément avec les techniques habituelles d'analyse de survie. Une pratique courante lorsqu'on modélise le risque de démence consiste cependant à censurer à droite les sujets « gênants » (les décédés sans diagnostic de démence) à la date de dernière visite sans démence, de façon à pouvoir malgré tout les appliquer. Il a déjà été montré qu'une telle pratique induit une sous-estimation de l'incidence de la démence qui est d'autant plus importante que la différence entre le risque de décès des déments et des non déments est grande. Nous avons montré dans le chapitre II et dans Leffondré et al. (2013) que les effets des facteurs de risque de démence peuvent aussi être biaisés.
Le modèle approprié dans ce contexte est un modèle *illness-death* qui prend en compte des temps exacts d'observation pour le décès et des temps censurés par intervalle pour la

démence, dans lequel les paramètres associés aux trois transitions sont estimés conjointement. Ce modèle permet de plus d'estimer d'autres quantités qui ont un intérêt certain en épidémiologie. Nous en avons fait un inventaire dans le chapitre III et dans Touraine et al. (2013b) et les avons largement expliquées et illustrées. Par exemple, nous pouvons estimer l'espérance de vie d'un sujet dément ou le risque de démence au cours de la vie entière (*lifetime risk*) à un âge donné et pour un ensemble de facteurs individuels donnés. La vocation de ce travail de thèse n'étant pas de faire de la recherche en épidémiologie, les applications n'ont pas été approfondies. Nous pensons qu'elles méritent de l'être, que cela soit sur les données des deux cohortes présentées ou sur d'autres.
Nous avons bon espoir à ce sujet car les méthodes d'estimation des intensités de transition, des effets des facteurs de risque et des autres quantités d'intérêt d'un modèle *illness-death* pour données censurées sont maintenant implémentées dans le paquet R **SmoothHazard** (présenté dans le chapitre V), développé durant la thèse. Il rend accessible l'utilisation du modèle *illness-death* sur des données censurées par intervalle, d'autant plus qu'un article a été écrit pour faciliter la compréhension du paquet et sa mise en pratique (Touraine et al., 2013a).
Enfin, le modèle *illness-death* pour données censurées par intervalle a été adapté aux données de cohorte dont nous disposons, c'est-à-dire généralisé à des données groupées. Des effets aléatoires ont été introduits afin de prendre en compte les effets de facteurs partagés par les individus d'un même groupe (commune, couple) sur les risques de démence et de décès.

Notons que ce travail a été motivé par l'étude de la démence à partir de données de cohorte, mais bien évidemment il reste valable pour d'autres applications. Il s'inscrit plus généralement dans un contexte *illness-death* où les temps de transition vers l'état intermédiaire sont censurés par intervalle. À notre connaissance, le paquet **SmoothHazard** a été utilisé entre autres pour l'étude de la dépendance lourde des sujets déments (les états 0, 1 et 2 représentant respectivement la démence, la dépendance lourde et le décès). Il peut également être adapté à des modèles multi-états plus complexes (à plus de trois états) tant que le nombre de chemins possibles entre deux temps d'observation est fini. Cependant, dans des modèles plus complexes, il arrive que le nombre d'évènements associés à chacune des transitions ne soit pas suffisamment grand et que l'on ne parvienne donc pas à estimer les paramètres. Dans ce cas, des hypothèses peuvent être faites ; on peut par exemple supposer que deux intensités de transition sont égales.

Discussion

Nous avons mis en avant les avantages de ce travail de thèse mais nous ne nous sommes pas encore attardé sur ses limites.

Limites et perspectives

Commençons par discuter de celles qui nous paraissent les plus importantes. Ce sont aussi celles qui ont été soulevées dans la pratique par les utilisateurs de **SmoothHazard**.

Variables dépendantes du temps

Un inconvénient majeur dans l'utilisation du paquet **SmoothHazard** est qu'il ne permet pas de prendre en compte des variables dépendantes du temps. Cela réduit le champ des applications possibles. Concernant l'étude de la démence par exemple, on ne peut pas prendre en compte des variables comme l'indice de masse corporelle ou l'hypertension artérielle, à moins de faire l'hypothèse (forte) que leur valeur reste inchangée tout au long du suivi. De plus, cela est contraignant lorsqu'on cherche des marqueurs pronostiques de la maladie afin de la détecter le plus précocement possible. Par exemple, lorsqu'on s'intéresse à la vitesse de marche comme facteur pronostique de la démence, il est seulement possible d'utiliser la valeur mesurée au début du suivi. Pourtant, la modification de la vitesse de marche pourrait aussi être importante dans l'apparition de la démence, de la même façon que l'accélération du déclin cognitif l'est. Mais pour étudier cela, des modèles conjoints pourraient être plus adaptés (Jacqmin-Gadda et al., 2006; Yu and Ghosh, 2010).

Adéquation au modèle

Un autre inconvénient majeur est celui de l'adéquation au modèle, en particulier de l'hypothèse d'intensités de transition proportionnelles. Comme nous venons de le voir, elle ne peut être vérifiée en considérant une variable dépendante du temps. Les méthodes basées sur la distribution de la fonction de survie (validation graphique, résidus de Cox Snell) ne sont pas non plus applicables car dans un modèle *illness-death*, la fonction de survie (ou fonction de survie globale) dépend à la fois du risque de transition vers l'état malade et du risque de transition vers l'état décédé. De plus, en présence de censure par intervalle, le fait de ne pas connaître la trajectoire des sujets décédés sans diagnostic de démence rend difficile l'utilisation de méthodes comme celle

des résidus de Shœnfeld (en particulier pour la transition $0 \to 2$). La seule méthode que nous avons utilisée est une méthode de validation graphique qui consiste à estimer les intensités de transition sur autant de sous-échantillons qu'il y a de modalités de la variable explicative testée, et ainsi vérifier la proportionnalité des fonctions. Il paraît cependant essentiel de s'intéresser par la suite à la question de l'adéquation au modèle.

Choix de la méthode d'estimation

Une question qui peut se poser lorsqu'on utilise le modèle *illness-death* présenté dans cette thèse est le choix de la méthode d'estimation. Jusqu'à présent, nous avons conseillé aux utilisateurs de **SmoothHazard** de tester dans un premier temps les deux méthodes (paramétrique et semi-paramétrique) sans variables explicatives et de comparer les courbes des intensités de transition. Lorsque les courbes des intensités paramétrées selon une loi de Weibull étaient proches de celles des intensités de type splines, nous avons alors recommandé la méthode paramétrique qui nécessite un temps de calcul moindre et qui est plus facile d'utilisation. Cependant, dans le futur, il serait intéressant de réfléchir à l'élaboration d'un critère qui permettrait de choisir entre le modèle paramétrique et le modèle semi-paramétrique. Il pourrait aussi être intéressant de proposer une nouvelle méthode d'estimation dans le paquet qui soit aussi flexible que la méthode semi-paramétrique actuelle mais moins gourmande en temps de calcul. On pourrait par exemple toujours utiliser des splines pour spécifier la forme des intensités de transition mais maximiser la vraisemblance non pénalisée.

Autres limites et perspectives

Censure informative

Une autre limite de ce travail de thèse est que nous ne nous sommes pas intéressé à la présence d'une éventuelle censure informative chez les sujets censurés à droite pour la démence. En effet, si le déclin cognitif augmente le risque d'être perdu de vue, les sujets censurés à droite car perdus de vue seraient à plus haut risque de démence que les autres. Une façon d'étudier cette censure informative serait d'ajouter les deux états « perdu de vue » et « dément et perdu de vue » au modèle *illness-death* comme dans Barrett et al. (2011).

Simulation de risques concurrents

Tout au long de ce manuscrit, nous avons mis en garde le lecteur, à travers plusieurs remarques, contre l'approche qui considère des temps d'évènement latents (*latent failure times*) dans un modèle à risques concurrents ou dans un modèle *illness-death*. Dans les simulations effectuées dans les chapitres II et IV, nous avons pourtant généré nos données selon cette approche. En effet, pour chaque sujet nous avons simulé deux temps : l'un associé à la maladie (transition $0 \rightarrow 1$) et l'autre associé au décès des non malades (transition $0 \rightarrow 2$). Le minimum de ces deux temps a été utilisé pour déterminer le temps d'évènement et l'évènement associé (maladie ou décès). Une autre méthode de simulation, appropriée à la manière dont les données sont ensuite analysées, a été proposée par Beyersmann et al. (2011, 2009) et devrait être privilégiée par la suite.

Annexe A : The SmoothHazard package for R : Fitting regression models to interval-censored observations of illness-death models

Abstract The irreversible illness-death model describes the pathway from an initial state to an absorbing state either directly or through an intermediate state. This model is frequently used in medical applications where the intermediate state represents illness and the absorbing state represents death. In many studies, disease onset times are not known exactly. This happens for example if the disease status of a patient can only be assessed at follow-up visits. In this situation the disease onset times are interval censored. This article presents the **SmoothHazard** package for R. It implements algorithms for simultaneously fitting regression models to the three transition intensities of an illness-death model where the transition times to the intermediate state may be interval censored and the all event times can be right censored. The program parses the individual data structure of the subjects in a data set to find the individual contributions to the likelihood. The three baseline hazard functions are modelled by Weibull distributions, alternatively in a semi-parametric approach by an M-spline approach. For a given set of covariates, the estimated transition intensities can be combined into predictions of cumulative event probabilities and life expectancies.

1 Introduction

The irreversible illness-death model is a multi-state model which has many applications in various areas of research, for example in the medical field. The model describes the transitions from an initial state (e.g., alive and disease-free) to an absorbing state (e.g., death) either directly or via an intermediate state (e.g., disease) (Figure A.1). The transition intensities α_{01}, α_{02}, and α_{02} are positive functions of time which can also depend on covariates.

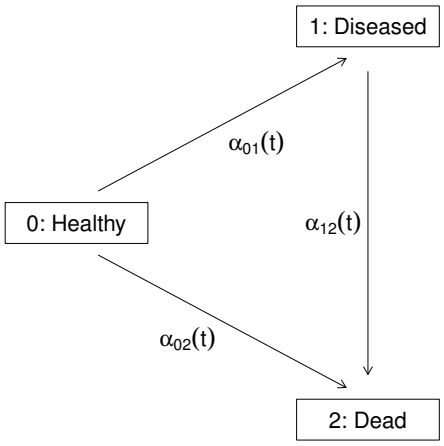


Figure A.1 – The irreversible illness-death model

In some applications it happens for some or all subjects that the transition times from the initial state to the intermediate state are interval censored. This occurs for example when the status of the intermediate state can only be determined at a sequence of visit times. In this case, if a subject is diagnosed as diseased at one of the visit times, say R, then it is only known that the subject was last seen disease-free at the previous visit time, say L, and hence the time of the onset of the disease is interval censored between L and R for this subject. Furthermore, both the process of visit times and the observation of the time of the transition into the absorbing state are usually right censored, i.e., limited to the individual follow-up period of the subjects. This yields a rather complex general observational pattern, because for a subject who died without being diagnosed as diseased at earlier visit times, it may or it may not be possible to determine retrospectively if and when the subject became diseased between the last visit time and the time of death.

The **SmoothHazard** package provides estimates of the baseline transition intensities and of covariate effects when the data fall into one of the 6 cases that are displayed

in Figure A.2. Thus, the case of left-truncated event times (delayed entry) is covered,

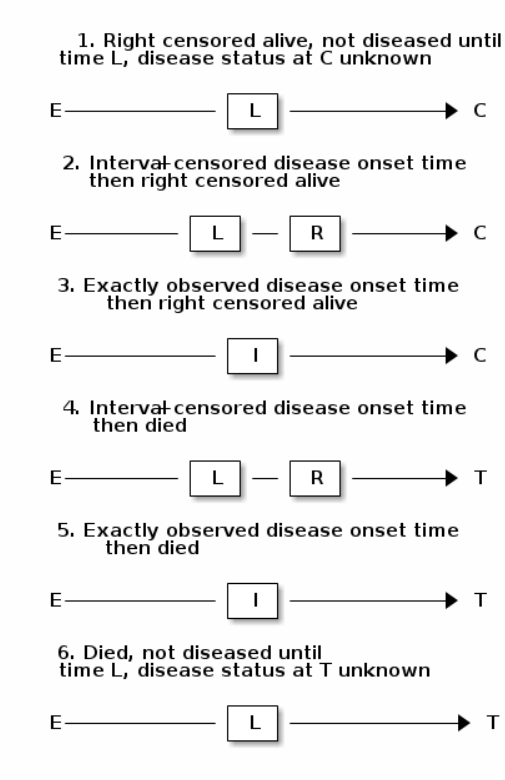


Figure A.2 – Observational patterns that are recognized by **SmoothHazard**. The letters I and T denote the transition times into the intermediate and absorbing state, respectively. The letters E and C denote the start and end of follow-up, respectively, and the letters L and R the visit times between which the transition into the intermediate happened.

as well as the case where for some subjects the transition time into the intermediate state is observed exactly and for others it is interval censored. Finally, the special case is covered where for some or all subjects no intermediate information is available about the disease status such that it is only known whether or not the subjects became diseased between the start and the end of follow-up. The latter occurs in Figure A.2 when $E = L$ and $R = \min(T, C)$ in cases 2. or 4.

To estimate covariate effects on the three transition intensities, implemented are regression models which assume proportional transition intensities and a non-homogeneous Markov process. The user chooses between a fully parametric model where each of the baseline intensities is described by the parameters of a Weibull distribution and a semi-parametric model where the baseline intensities are left unspecified and approximated by M-splines. For the parametric model, the regression coefficients and Weibull parameters are estimated by maximising the likelihood, for the semi-parametric model, the

coefficients of the M-splines and the regression coefficients are estimated by maximising a penalized likelihood.

The package **SmoothHazard** then allows to do predictions of transition probabilities, cumulative probabilities of event and life expectancies for a given set of covariates, based on estimated baseline transition intensities and on estimated covariates effects.

If the exact transition times are observed, standard procedures like those implemented, e.g. in the packages **survival**, **rms**, **etm**, **mstate** can be used to estimate transition intensities, regression coefficients and functionals thereof (see Beyersmann et al., 2011; De Wreede et al., 2010). In particular, the regression coefficients can be estimated using Cox partial likelihood (Cox, 1975) without the need to model the baseline intensities. However, when transition times to the intermediate event are interval censored, it is generally not possible to arrive at consistent estimates with the software provided by the packages listed above. Indeed, the approach to handle subjects who died with unknown disease status, consists in artificially ending their follow-up at the last time they were seen without disease and subsequently treat them as right censored. However, this approach can lead to a systematic bias in the estimates of transition intensities and of regression coefficients (Joly et al., 2002; Leffondré et al., 2013). The bias will be especially pronounced if the risk of death is higher for diseased subjects than the risk of death for disease-free subjects.

The **msm** package (Jackson, 2011) allows to fit Markov multi-state models to panel data where the status of the subjects is known at a finite series of inspection times. As a special case the setting includes the illness-death model and it can be used with interval-censored disease times and exact death times. However, in this package the likelihood is calculated using the Kolmogorov differential equations that relate the transition probabilities and the transition intensities and to make this work a time-homogeneity assumption is made where all transition intensities are constant or piecewise-constant between two successive observation times.

1.0.1 Outline

The main functions of **SmoothHazard** is

— `idm` : for fitting illness-death regression models based on possibly interval-censored disease times and right censored times.

A fitted illness-death model as produced by `idm` can be used in the following functions to calculate predictions :

— `predict.idm` : for estimating transition probabilities and cumulative probabilities of event for a given set of covariates ;

— `lifexpect` : for estimating life expectancies for a given set of covariates.

The `R` function `idm` is essentially an interface between the user and FORTRAN programs which constitute the heart of the package **SmoothHazard**.

Section 2 presents the model and the likelihood. Section 3 presents the estimation methods. Section 4 briefly presents predictions that can be made in an illness-death model. Section 5 provides some examples illustrating **SmoothHazard**.

2 Model and likelihood

We consider an illness-death process $X = (X(t), t \geq 0)$ which takes values in $\{0, 1, 2\}$ (Figure A.1). Subjects are initially disease-free ($X(0) = 0$) and may become diseased (transition $0 \to 1$) and die (transition $1 \to 2$), or die directly without disease (transition $0 \to 2$.) X is assumed to be a non-homogeneous Markov process which means that the future evolution of the process $\{X(t), t > s\}$ depends on the current time s and only on the current state $X(s)$. Thus, the distribution of X is fully characterized by the set of transition probabilities :

$$p_{hl}(s,t) = \mathbb{P}(X(t) = l | X(s) = h) \qquad hl \in \{01, 02, 12\}.$$

The transition probabilities are related to the instantaneous transition intensities α_{hl} shown in Figure A.1 by the relation :

$$\alpha_{hl}(t) = \frac{p_{hl}(t, t + \Delta t)}{\Delta t}.$$

We introduce covariate effects separately for each transition through proportional transition intensities regression models which are a natural extension of the Cox proportional hazard model :

$$\alpha_{hl}(t|Z_{hli}) = \alpha_{0,hl}(t) \exp\{\beta_{hl}^T Z_{hli}\}; \qquad hl \in \{01, 02, 12\}. \tag{1}$$

Here $\alpha_{0,hl}$ are baseline transition intensities, Z_{hli} are covariate vectors for subject i and β_{hl} are vectors of regression parameters for transition $h \to l$.

In the situation where the time to disease and the time to death are not interval

censored but either observed exactly or right censored, the regression coefficients could be estimated by the partial likelihood method without the need to specify and estimate the baseline hazard functions $\alpha_{0,hl}(t)$. For interval-censored transition times to the intermediate state, the situation is more complex. It turns out that we have to estimate all parameters simultaneously and that we need a model for the baseline transition intensity functions. This can be seen by inspecting the likelihood function.

For subject i, denote the conditional disease-free survival function by

$$S(t|Z_{01i}, Z_{02i}) = e^{-A_{01}(t|Z_{01i})-A_{02}(t|Z_{02i})}$$

where $A_{hl}(.|Z_{hli})$ is the conditional cumulative intensity function of transition $h \to l$:

$$A_{hl}(t|Z_{hli}) = \int_0^t \alpha_{hl}(u|Z_{hli})du.$$

Note that if subject i has entered the intermediate state, the conditional survival function in the intermediate state between times s and t is given by :

$$\frac{e^{-A_{12}(t|Z_{12i})}}{e^{-A_{12}(s|Z_{12i})}}.$$

We allow that the event times are left truncated, i.e., that subjects enter the study at the delayed entry time $E > 0$. The left truncation condition $X(E_i) = 0$ implies that subject i has survived in state 0 until time E_i.

In addition to the covariate vectors $Z_{01i}, Z_{02i}, Z_{12i}$ we observe the vector $(E_i, L_i, R_i, \delta_{1i}, \tilde{T}_i, \delta_{2i})$ where $\tilde{T}_i = \min(T_i, C_i)$ is the minimum between the transition time into the absorbing state T_i and the right censoring time C_i and $\delta_{2i} = \mathbb{1}\{T_i \leq C_i\}$. Also, $\delta_{1i} = 1$ if we know for sure that subject i was diseased between E_i and $\tilde{T}_i$ and $\delta_{1i} = 0$ otherwise. The visit times L_i and R_i are defined by $E_i \leq L_i \leq R_i \leq \tilde{T}_i$ if $\delta_{1i} = 1$ and by $E_i \leq L_i \leq \tilde{T}_i, R_i = \infty$ if $\delta_{1i} = 0$. When the transition time into the intermediate state is observed exactly, we have $\delta_{1i} = 1$ and $L_i = R_i$. In the latter case we also denote I_i for the transition time into the intermediate state.

We now detail the likelihood contributions according to the different observational patterns shown in Figure A.2 in the special case where there is no left truncation i.e. $E_i = 0$. Left-truncated event times are taken into account by simply dividing the above

likelihood contributions by the term $S(E_i|Z_{01i}, Z_{02i})$.

$$
\begin{aligned}
\text{case 1}: \quad & \mathcal{L}_i = S(C_i|Z_{01i}, Z_{02i}) + \int_{L_i}^{C_i} S(u|Z_{01i}, Z_{02i})\alpha_{01}(u|Z_{01i})\frac{e^{-A_{12}(C_i|Z_{12i})}}{e^{-A_{12}(u|Z_{12i})}}du \\
\text{case 2}: \quad & \mathcal{L}_i = \int_{L_i}^{R_i} S(u|Z_{01i}, Z_{02i})\alpha_{01}(u|Z_{01i})\frac{e^{-A_{12}(C_i|Z_{12i})}}{e^{-A_{12}(u|Z_{12i})}}du \\
\text{case 3}: \quad & \mathcal{L}_i = S(I_i|Z_{01i}, Z_{02i})\alpha_{01}(I_i|Z_{01i})\frac{e^{-A_{12}(C_i|Z_{12i})}}{e^{-A_{12}(I_i|Z_{12i})}} \\
\text{case 4}: \quad & \mathcal{L}_i = \int_{L_i}^{R_i} S(u|Z_{01i}, Z_{02i})\alpha_{01}(u|Z_{01i})\frac{e^{-A_{12}(T_i|Z_{12i})}}{e^{-A_{12}(u|Z_{12i})}}\alpha_{12}(T_i|Z_{12i})du \\
\text{case 5}: \quad & \mathcal{L}_i = S(I_i|Z_{01i}, Z_{02i})\alpha_{01}(I_i|Z_{01i})\frac{e^{-A_{12}(T_i|Z_{12i})}}{e^{-A_{12}(I_i|Z_{12i})}}\alpha_{12}(T_i|Z_{12i}) \\
\text{case 6}: \quad & \mathcal{L}_i = S(T_i|Z_{01i}, Z_{02i})\alpha_{02}(T_i|Z_{02i}) \\
& \quad + \int_{L_i}^{T_i} S(u|Z_{01i}, Z_{02i})\alpha_{01}(u|Z_{01i})\frac{e^{-A_{12}(T_i|Z_{12i})}}{e^{-A_{12}(u|Z_{12i})}}\alpha_{12}(T_i|Z_{12i})du
\end{aligned}
\tag{2}
$$

3 Estimation

The `idm` function computes estimates for the three baseline transition intensities and for the regression parameters using the Levenberg-Marquardt's algorithm (Levenberg, 1944; Marquardt, 1963) to maximize the (penalized) likelihood. The algorithm is a combination of a Newton-Raphson algorithm and a gradient descent algorithm (also known as the steepest descent algorithm). It has the avantage of being more robust than the Newton-Raphson algorithm while preserving its fast convergence property.

3.1 Parametric estimation

In the default estimation method of function `idm`, a Weibull parametrization for the baseline transition intensities is assumed :

$$\alpha_{0,hl}(t) = a_{hl}\ b_{hl}^{a_{hl}}\ t^{a_{hl}-1}; \quad hl \in \{01, 02, 12\}.$$

where a_{hl} and $\frac{1}{b_{hl}}$ are shape and scale parameters. The Weibull parameters estimates $\hat{a}_{hl}$ and $\hat{b}_{hl}$ and the vectors of regression parameters estimates $\hat{\beta}_{hl}$ are obtained simultaneously by maximizing the likelihood which is the product over the subjects' contributions according to equation 2 :

$$\mathcal{L}(\beta_{01}, \beta_{02}, \beta_{12}, a_{01}, a_{02}, a_{12}, b_{01}, b_{02}, b_{12}) = \prod_{i=1}^{n} \mathcal{L}_i(\beta_{01}, \beta_{02}, \beta_{12}, a_{01}, a_{02}, a_{12}, b_{01}, b_{02}, b_{12}).$$

Confidence intervals for the regression parameters are obtained using standard errors estimated by inverting the Hessian matrix of the log-likelihood, that is the matrix of the second partial derivatives of $\log \mathcal{L}$ given in the previous display. Pointwise confidence bands for the baseline transition intensities are obtained using a simulation-based approach explained below (section 4.1).

3.2 Semi-parametric estimation

In situations where it is suspected that the Weibull distribution does not fit the data very well one can think of extending the model and to leave the baseline intensity functions completely unspecified, as in the Cox regression model. Unfortunately, in interval-censored data there is no direct analogue to the partial likelihood and the Breslow estimator of the Cox model in right-censored data. The function `idm` implements a semi-parametric model where the three baseline transition intensities are approximated by linear combinations of M-splines. In this section we explain the basic steps of the approach.

3.2.1 The penalized likelihood

To control the smoothness of the estimated intensity functions, we penalize the log-likelihood by a term which specificies the curvature of the intensity functions. It is given by the square of the second derivates. The penalized log-likelihood (pl) is defined as :

$$pl = l - \kappa_{01} \int \alpha_{01}^{''\,2}(u|Z_{01})du - \kappa_{02} \int \alpha_{02}^{''\,2}(u|Z_{02})du - \kappa_{12} \int \alpha_{12}^{''\,2}(u|Z_{12})du \qquad (3)$$

where l is the log-likelihood and κ_{01}, κ_{02} and κ_{12} are three positive parameters which control the trade-off between the data fit and the smoothness of the functions. It is proposed that the penalization parameters are chosen by maximizing a cross-validated likelihood score. Here, leave-one-out is appealing as the result does not depend on the random seed as it would, e.g., for 10-fold cross-validation. However, since leave-one-out requires as many maximizations of the likelihood as there are subjects in the data set,

this can be computationally very expensive. To avoid extremely long run times we have implemented the following algorithm :

Step 1. We ignore the covariates and use a grid search method to find the values for $(\kappa_{01}, \kappa_{02}, \kappa_{12})$ based on an approximiation of the leave-one-out log-likelihood score. The approximation is equivalent to one step of the Newton-Raphson algorithm and reduces the number of calculations considerably. This approach was proposed by O'Sullivan (1988a) for survival models and studied by Joly et al. (2002) in an illness-death model with interval censored data.

Step 2. We use the results of Step 1, i.e. the optimized value of $(\kappa_{01}, \kappa_{02}, \kappa_{12})$ to maximize the penalized likelihood (3) with covariates. The parameters being maximized are the regression coefficients and the coefficients of the linear combination of the M-splines defined below.

3.2.2 M-splines

A family of M-spline functions of order k, $M_1, \ldots, M_n$ is defined by a set of m knots where $n = m + k - 2$ (Ramsay, 1988). We consider only cubic M-splines of order $k = 4$. Denote by $t_{01} = (t_{01,1}, \ldots, t_{01,m_{01}})$ a sequence of m_{01} knots used for approximating α_{01} and by $t_{02} = (t_{02,1}, \ldots, t_{02,m_{02}})$ and $t_{12} = (t_{12,1}, \ldots, t_{12,m_{12}})$ similar sequences of knots for approximating α_{02} and α_{12} respectively. We denote by $M_{hl}^T = M_{hl,1}, \ldots, M_{hl,n_{hl}}$ the families of n_{hl} cubic M-splines, with $n_{hl} = m_{hl}+2$ and for $hl \in \{01, 02, 12\}$. The baseline transition intensity $\alpha_{0,hl}$ is approximated using the following linear combination :

$$\tilde{\alpha}_{0,hl}(t) = \sum_{i=1}^{n_{hl}} (a_{hl,i})^2 M_{hl,i}(t)$$

where $a_{hl,i}$ are unkown parameters. The n_{hl} M-splines are integrated in order to produce a family of monotone splines, these are called I-splines. Thus, with each M-spline $M_{hl,i}$ we associate an I-spline $I_{hl,i}$:

$$I_{hl,i}(t) = \int_{t_{hl,1}}^{t} M_{hl,i}(u)du.$$

For given values of the parameters $a_{hl,i}$, we can approximate the cumulative baseline transition intensities A_{hl} by a linear combination of I-splines :

$$\tilde{A}_{0,hl}(t) = \sum_{i=1}^{n_{hl}} (a_{hl,i})^2 I_{hl,i}(t).$$

Because M-splines are non-negative, the positivity constraint on $(a_{hl,i})^2$ ensures that $\tilde{A}_{0,hl}$ is monotone increasing.

Confidence intervals of the regression parameters are obtained using estimated standard errors which are obtained by inverting the Hessian matrix of the penalized log-likelihood.

Confidence intervals for the transition intensities $\alpha_{hl}(t)$ are obtained using the Bayesian approach proposed in O'Sullivan (1988a) for survival analysis where the standard errors are estimated by $M_{hl}(t)^T H^{-1} M_{hl}(t)$ where H denotes the Hessian matrix of the penalized log-likelihood.

4 Predictions

Often in illness-death models the functions of interest are the transition intensities. However, other quantities (transition probabilities, cumulative probabilities and life expectancies) which can be expressed in terms of the transition intensities (Touraine et al., 2013b) may provide additional information and have a more natural interpretation.

For example, given a set of covariates $Z_{01,i}, Z_{02,i}, Z_{12,i}$ for a subject i who is diseased at time s, one could be interested in probability to be still alive at some time $t > s$, or in life expectancy. Given a set of covariates $Z_{01,j}, Z_{02,j}, Z_{12,j}$ for a subject j who is diseased-free at time s, one could be interested in lifetime risk of disease or in healthy life expectancy (expected remaining sojourn time in the healthy state). Since these quantities can be written in terms of the transition intensities, **SmoothHazard** provides estimates of them using estimates of the transition intensities. Confidence intervals are calculated using the simulation-based method immediately following.

4.1 Confidence regions

A simulation based approach (Mandel, 2013) is used to calculate confidence intervals for the transition intensities $\alpha_{hl}(t)$ in the parametric approach and for the other quantities of interest in both parametric and semi-parametric approaches. To briefly outline how it works, we generically denote by θ the vector of all the parameters that characterize the likelihood and by $\hat{\theta}$ the maximum (penalized) likelihood estimator. θ contains the Weibull parameters in the parametric model, the spline parameters in the semi-parametric model and the regression parameters in both models.

Case	Description	δ_1	δ_2	L	R	T	Remark
1	No illness observed, right-censored death time	0	0	L_i	L_i	C_i	$L_i \leq C_i$
2	Interval-censored ill time, right-censored death time	1	0	L_i	R_i	C_i	$L_i < R_i \leq C_i$
3	Exact ill time, right-censored death time	1	0	L_i	L_i	C_i	$L_i \leq C_i$
4	Interval-censored ill time, death time observed	1	1	L_i	R_i	T_i	$L_i < R_i \leq T_i$
5	Exact ill time, death time observed	1	1	L_i	L_i	T_i	$L_i \leq T_i$
6	No illness observed, death time observed	0	1	L_i	L_i	T_i	$L_i \leq T_i$

Tableau A.1 – Description of how the data set must be built to be understood by the `idm` function

We assume the asymptotic normality for the estimator $\hat{\theta}$ and denote by $\hat{V}_{\hat{\theta}}$ the estimated covariance matrix of $\hat{\theta}$. We consider a multivariate normal distribution with the parameters estimates as expectation and $\hat{V}_{\hat{\theta}}$ as covariance matrix. We generate n vectors ($n = 2000$ in practice) from this distribution : $\theta^{(1)}, \ldots, \theta^{(n)}$. Based on them, we can calculate n values for the transition intensities : $\alpha_{hl}^{(1)}(t)$,...,$\alpha_{hl}^{(n)}(t)$, and therefore n values for any quantity of interest written in terms of the transition intensities. The n values reflecting the sample variation (Aalen et al., 1997), we order them and the 2.5^{th} and the 97.5^{th} empirical percentiles are then used as lower and upper confidence bounds for 95% confidence intervals. This procedure can be repeated for any t, so we can obtain pointwise confidence bands for $\alpha_{hl}(.)$.

5 Using SmoothHazard

5.1 How to prepare the data

Table A.1 shows how the program interpretes the structure of the data set. In all cases, L_i may be equal to the entry time. Some more details are necessary to distinguish the case where the ill status is known at the last follow-up time for death from the case where this is not possible.

— In case 1, if $L_i < C_i$ then it is assumed that the subject may become ill between L_i and C_i. If $L_i = C_i$ it is assumed that the subject is disease-free at time C_i.

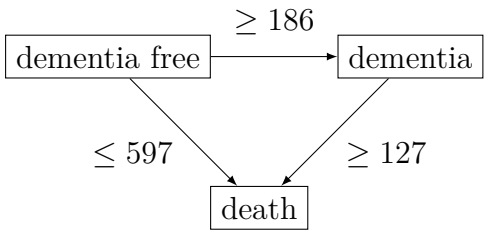


Figure A.3 – The exact number of transitions in the illness-death model with interval-censored time to disease is unknown : example of the `Paq1000` data set.

In the latter case the integral of the likelihood equals zero.

— In case 6, if $L_i < T_i$ then it is assumed that the subject may become ill between L_i and T_i. If $L_i = T_i$ it is assumed that the subject is disease-free at time T_i. In the latter case the integral of the likelihood equals zero.

5.2 Paquid study

In order to illustrate the functionality of the package we provide a random subset containing data from 1000 subjects that were enrolled in the Paquid study (Letenneur et al., 1999), a large cohort study on mental and physical aging.

```
1 library(SmoothHazard)
2 data(Paq1000)
```

The population consists of subjects aged 65 years and older living in Southwestern France. The event of interest is dementia and death without dementia is a competing risk. Furthermore, the time to dementia onset is interval censored between the diagnostic visit and the previous one and demented subjects are at risk of death. Thus, subjects who died without being diagnosed as demented at their last visit may have become demented between last visit and death.

In this subset `186` subjects are diagnosed as demented and `724` died from whom `597` without being diagnosed as demented before. Because of interval censoring more than `186` should have been demented, more than `127` should have been dead with dementia and less than `597` should have been dead without dementia (see Figure A.3).

Age is chosen as the basic time scale and subjects are dementia-free (and alive) at entry into study. Consequently, we need to deal with left-truncated event times.

```
head(round(Paq1000,1))
```

```
  dementia death   e    l    r    t certif gender
1        1     1 72.3 82.3 84.7 87.9      0      0
2        0     1 77.9 78.9 78.9 79.6      0      1
3        0     1 79.9 79.9 79.9 80.9      0      0
4        0     1 74.7 78.6 78.6 82.9      1      1
5        0     1 76.7 76.7 76.7 79.2      0      1
6        0     0 66.2 71.4 71.4 84.2      1      0
```

Each row in the data corresponds to one subject. The variables `dementia` and `death` are δ_1 and δ_2, the status variables for dementia and death. The variable `e` contains ages of subjects at entry into study. The variables `l` and `r` contain the left and right endpoints of the censoring intervals. For demented subjects, `r` is the age at the diagnostic visit and `l` is the age at the previous one. For non demented subjects, `l` and `r` are the age at the latest visit without dementia (`l`=`r`). The variable `t` is the age at death or at latest news on vital status. There are two binary covariates : `certif` for primary school diploma (762 with diploma and 238 without diploma) and `gender` (578 women and 422 men).

The function `idm` computes estimates for the three transition intensities $\alpha_{01}(.)$, $\alpha_{02}(.)$, $\alpha_{12}(.)$ which represents age-specific incidence rate of dementia, age-specific mortality rate of dementia-free subjects and age-specific mortality rate of demented subjects, respectively. Proportional transition intensities regression models allow for covariates on each transition. Covariates are specified independently for the regression models of the three transition intensities by the right hand side of the respective formula `formula01`, `formula02` and `formula12`.

Interval censoring and left truncation must be specified at the left side of the formula arguments using the `Hist` function. For left-truncated data, the `entry` argument of `Hist` must contain the vector of delayed entry times. For interval-censored data, the `time` argument of `Hist` must contain a list of the left and right endpoints of the intervals. The `data` argument contains the data frame in which to interpret the variables of `formula01`, `formula02` and `formula12`. The left side of `formula12` argument does not need to be filled because all the data informations are already contained in `formula01` and `formula02`. The left side of `formula12` argument is required only if we want the covariates impacting transition $1 \to 2$ different from those impacting transition $0 \to 2$.

5.3 Fitting the illness-death model based on interval-censored data

The main function `idm` computes estimates for the three baseline transition intensities and for the regression parameters of an illness-death model. The `intensities` argument by specifying the form of the transition intensities allows to select either the parametric or a semi-parametric estimation method :

— With the default value `"Weib"`, a Weibull distribution is assumed for the baseline transition intensities and the parameters are estimated by maximizing the log-likelihood ;

— With the `"Splines"` value, the baseline transition intensities are approximated by linear combinations of M-splines and the parameters are estimated by maximizing the penalized log-likelihood.

We stop the iterations of the maximization algorithm when the differences between two consecutive parameters values, log-likelihood values, and gradient values is small enough. The default convergence criteria are 10^{-5}, 10^{-5} and 10^{-3} and can be changed by means of the `eps` argument.

We now illustrate how to fit the illness-death model to the `Paq1000` data set, based on interval-censored dementia times and exact death times.

In the following call, a Weibull parametrization is used for the three baseline transition intensities and we include two covariates on the transition to dementia, one covariate on the transition from no dementia to death and no covariates on the transition from dementia to death. Note that in case of missing `formula12` argument the covariates on the $1 \to 2$ transition are the same as the ones specified in the `formula02` argument.

```
1 fit.weib <- idm(formula01=Hist(time=list(l,r),event=dementia,entry=e)~
      certif+gender,
2        formula02=Hist(time=t,event=death,entry=e)~gender,
3        formula12= ~ 1,
4        data=Paq1000)
5 fit.weib
```

```
Call:
idm(formula01 = Hist(time = list(l, r), event = dementia, entry = e) ~
```

```
    certif + gender, formula02 = Hist(time = t, event = death,
    entry = e) ~ gender, formula12 = ~1, data = Paq1000)

Illness-death model: Results of Weibull regression for the intensity functions.

number of subjects:  1000
number of events '0-->1':  186
number of events '0-->2' or '0-->1-->2':  724
number of covariates:  2 1 0

                coef SE.coef     HR          CI      Wald p.value
certif_01_01 -0.4117  0.1827 0.6625 [0.46;0.95]  5.077106 0.02424
gender_01_01 -0.2621  0.1561 0.7694 [0.57;1.04]  2.818364 0.09319
gender_02_02  0.6712  0.1143 1.9565 [1.56;2.45] 34.449583 < 1e-04

               Without cov  With cov
Log likelihood   -3075.308 -3053.648

Parameters of the Weibull distribution: 'S(t) = exp(-(b*t)^a)'
      alpha01    alpha02    alpha12
a 11.12344625 8.82268159 6.44006486
b  0.01102198 0.01074539 0.01381268

----
Model converged.
number of iterations:  6
convergence criteria: parameters= 7.3e-10
                    : likelihood= 2.3e-08
                    : second derivatives= 2.8e-12
```

The hazard ratios HR (e^{coef}) have the usual interpretation, as in a parametric Cox regression model.

The three baseline transition intensity functions can be displayed as functions of time, functions of age in our illustrative example (Figure A.4).

```
1 par(mgp=c(4,1,0),mar=c(5,5,5,5))
```

```
plot(fit.weib,conf.int=TRUE,lwd=3,citype="shadow",xlim=c(65,100), axis2.
    las=2,axis1.at=seq(65,100,5),xlab="Age (years)")
```

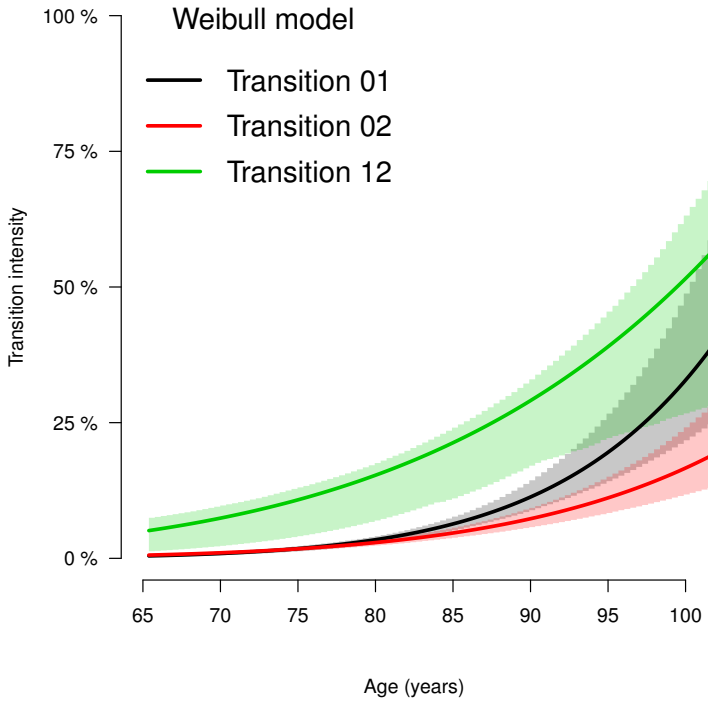


Figure A.4 – Baseline transition intensities estimated using the Weibull parametrization of the parametric approach

The other estimation option in the function `idm` permits to relax the strict parametric assumptions of the Weibull regression models. With the option `intensities="Splines"`, linear combinations of M-splines are used to approximate the three baseline transition intensities. Although this option implies a considerable amount of extra computations (see Section 3.2), the call and the printed output are very similar to the Weibull model :

```
fit.splines <- idm(formula01=Hist(time=list(l,r),event=dementia,entry=e)~
    certif+gender,
         formula02=Hist(time=t,event=death,entry=e)~gender,
         formula12= ~ 1,
         intensities="Splines",data=Paq1000)
fit.splines
```

```
Call:
idm(formula01 = Hist(time = list(l, r), event = dementia, entry = e) ~
    certif + gender, formula02 = Hist(time = t, event = death,
    entry = e) ~ gender, formula12 = ~1, data = Paq1000, intensities = "Splines")

Illness-death regression model using M-spline approximations
 of the baseline transition intensities.

number of subjects:  1000
number of events '0-->1':  186
number of events '0-->2' or '0-->1-->2':  724
number of subjects:  1000
number of covariates:  2 1 0

Smoothing parameters:
      transition01 transition02 transition12
knots        7e+00        7e+00            7
kappa        8e+05        2e+05        50000

                coef SE.coef     HR          CI      Wald p.value
certif_01_01 -0.3762  0.1853 0.6865 [0.48;0.99]  4.122728 0.04231
gender_01_01 -0.2297  0.1580 0.7948 [0.58;1.08]  2.113669 0.14599
gender_02_02  0.6529  0.1119 1.9211 [1.54;2.39] 34.039816 < 1e-04

                         Without cov  With cov
Penalized log likelihood   -3072.464 -3052.046

----
Model converged.
number of iterations:  8
convergence criteria: parameters= 4e-09
                    : likelihood= 9.5e-08
                    : second derivatives= 2.2e-11
```

Again, the estimated baseline transition intensities can conveniently be visualized in a joint graph (Figure A.5).

```
par(mgp=c(4,1,0),mar=c(5,5,5,5))
plot(fit.splines,conf.int=TRUE,lwd=3,citype="shadow",xlim=c(65,100),
    axis2.las=2,axis1.at=seq(65,100,5),xlab="Age (years)")
```

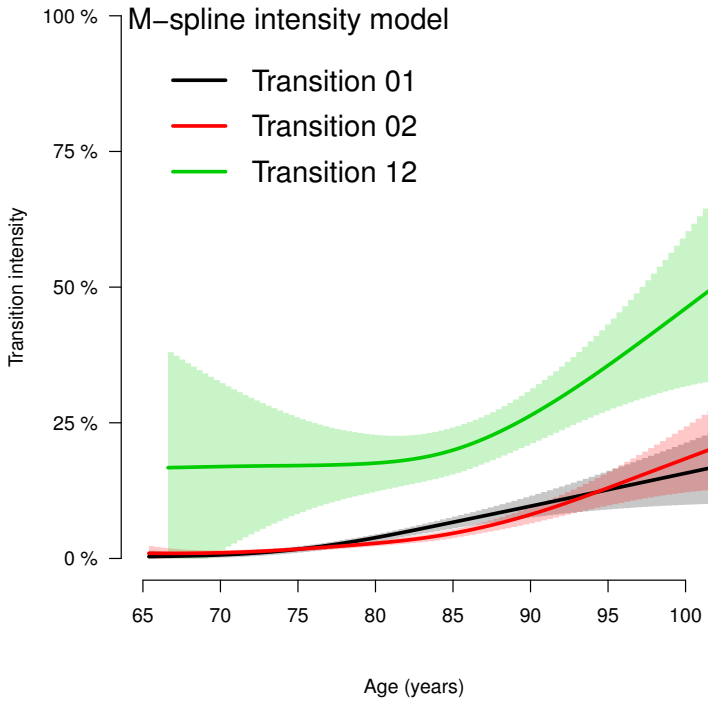


Figure A.5 – Baseline transition intensities estimated using the splines approximation of the semi-parametric approach

5.3.1 Semi-parametric estimation method : choice of smoothing parameters

Some optional arguments are specific to the semi-parametric approach (when using the option `intensities="Splines"`) :

— `n.knots` contains a vector (by default `c(7,7,7)`) specifying the number of knots on the $0 \to 1$, $0 \to 2$ and $1 \to 2$ transitions, respectively ;

— `knots` contains the choice of the knots placement (equidistant by default or quantile-based placement) or a list of sequences of knots for transitions $0 \to 1$, $0 \to 2$ and $1 \to 2$ respectively, to be specified by the user ;
— `CV` (FALSE by default) is set to TRUE for using approximate leave-one-out cross-validation score to choose the smoothing parameters κ_{01}, κ_{02}, κ_{12} ;
— `kappa` contains the smoothing parameters if `CV=FALSE` (arbitrary choice of the smoothing parameters κ_{01}, κ_{02}, κ_{12}) ; the initial smoothing parameters for the grid search method which maximize the approximate leave-one-out cross-validation score if `CV=TRUE`.

By default the function `idm` selects equidistant sequences of 7 knots between the minimal and maximal event times (`e`, `l` and `r` for `Paq1000`). There must be a knot before or at the first time from which there are subjects at risk and after or at the last time of transition. The current implementation of our program requires a minimum of 4 knots for each transition intensitiy.

Consequently, the semi-parametric approach requires much more information than the parametric one to achieve convergence. The number of parameters to be estimated is larger, and enough observation times on each transition are required to fit the splines. In particular, in data sets where few $1 \to 2$ transitions times are observed, we does not recommended this approach. Increasing the number of knots does not deteriorate the estimates of the transition intensities : this is because the degree of smoothing in the penalized likelihood method is tuned by the smoothing parameters κ_{01}, κ_{12} and κ_{02}. On the other hand, once a sufficient number of knots is established, there is no advantage in adding more. Moreover, the more knots, the longer the running time. Some numerical problem can arise, particularly for a large number of knots. So it is recommended to start with a small number of knots (e.g. 5 or 7) and increase the number of knots until the graph of the transition intensities function remains unchanged (from our own experience rarely more than 12 knots).

The default values for the smoothing parameters κ_{01}, κ_{02}, κ_{01}, are suitable for the `Paq1000` data set. However, these values can be expected to be very different depending on time scale, number of subjects and number of knots. The cross-validation option can be used to find appropriate smoothing parameters. However, the running time with cross-validation is very long and an empirical technique can be preferred. It consists in repeating the `idm` running trying different smoothing parameters. After each estimation, the transition intensities are plotted. If the curves seem too smooth, it may be useful to reduce the smoothing parameter. Similarly, if the curves are too

wiggly, the smoothing parameter may be increased.

5.4 Making predictions

A object as returned by the `idm` function can be used as argument of the `predict` function in order to obtain transition probabilities, cumulative probabilities of event and life expectancies with confidence intervals. For example, the following call give predictions regarding a 70 years-old male subject who have primary school diploma, over a 10 years horizon :

```
pred <- predict(fit.weib,s=70,t=80,Z01=c(1,1),Z02=1)
x<-round(do.call("rbind",pred),2)
colnames(x) <- c("Probability","Lower","Upper")
x
```

```
      Probability Lower Upper
p00          0.64  0.59  0.68
p01          0.05  0.03  0.07
p11          0.33  0.27  0.67
p12          0.67  0.33  0.73
p02_0        0.29  0.24  0.33
p02_1        0.03  0.01  0.05
p02          0.32  0.27  0.36
F01          0.08  0.05  0.12
F0.          0.36  0.32  0.41
```

The covariates values must be specified in the `Z01`, `Z02` and `Z12` arguments in the same order as they were entered in the preceding `idm` call.

The ouput attributes are :

— for a dementia-free 70 years-old subject :
  — the probability of being still alive and dementia-free 10 years later $p_{00}(70, 80)$,
  — the probability of being still alive but demented 10 years later $p_{01}(70, 80)$,
  — the probability of dying in the next 10 years $p_{02}(70, 80)$ having been demented before ($p^1_{02}(70, 80)$) or not ($p^0_{02}(70, 80)$),
  — the absolute risk of dementia in the 10 years (10 years later, the subject may be dead or not) $F_{01}(s, t)$,

— the absolute risk of exit from the no dementia state in the 10 years $F_{0\bullet}(s,t)$ (due to either dementia or death) ;

— for a demented 70 years-old subject : the probability of dying in the next 10 years $p_{12}(s,t)$ or not $p_{11}(s,t)$.

The following calls give life expectancies regarding a 80 years-old female subject who have primary school diploma based on the transition intensities estimates from respectively the parametric approach and the semi-parametric approach :

```
LE.weib <- lifexpect(fit.weib,s=80,Z01=c(1,0),Z02=0)
x<-round(do.call("rbind",LE.weib),2)
colnames(x) <- c("LE","Lower","Upper")
x
```

```
                           LE Lower Upper
life.in.0.expectancy     8.87  7.89  9.78
life.expectancy.nondis  10.45  9.79 11.61
life.expectancy.dis      4.89  4.40  7.87
```

```
LE.splines <- lifexpect(fit.splines,s=80,Z01=c(1,0),Z02=0,CI=FALSE)
x<-round(do.call("rbind",LE.splines),2)
colnames(x) <- c("LE")
x
```

```
                           LE
life.in.0.expectancy     8.82
life.expectancy.nondis  10.42
life.expectancy.dis      4.91
```

The confidence intervals calculation may take time, especially using the splines estimates of the transition intensities. To suppress this calculation, the `CI` argument must be set to `FALSE` (see above). To reduce the computation time of the confidence intervals, the number of simulations can also be modified using the `nsim` argument (by default 2000 for the `predict` function and 1000 for the `lifexpect` function).

The output attributes of the `lifexpect` function are :

— for a dementia-free 80 years-old subject :

— the life expectancy in state 0 (healthy life expectancy),

— the life expectancy ;
— for a demented 80 years-old subject : the life expectancy.

5.4.1 Warnings regarding predictions

Predictions using the splines estimates of the transition intensities are not possible if involving times prior to the first knot or times beyond the last knot. Moreover, the life expectancies are calculated using integration until infinity using the Weibull estimates and until the last knot using the splines estimates. Consequently, to calculate life expectancies using the splines estimates, we implicitly assume that the last knot time is the maximal time of death. The above life expectancies calculating from the Weibull estimates or the splines estimates of the transition intensities are very close because the follow-up period of the `Paq1000` data set is long. However, in other data sets this assumption may not hold anymore. For data sets with short follow-up period, it is possible to calculate quantities involving any time, even infinity like life expectancies. However, beyond the follow-up time, they are not based anymore on estimations of the transition itensity functions but rather on extrapolations on them. Consequently, we do not recommend to do predictions involving times beyond the follow-up period. Finally, to avoid numerical problem in the predictions calculations, the first and last knots for all transitions must be the same or very close.

Annexe B : Chapitre III (sous-section 4.2) : tableaux et figures

Ci-après figurent les tableaux et figures qui ont été retirés du chapitre III pour en faciliter la lecture. Elles concernent les prévisions issues du modèle contenant la variable explicative cep.

Tableau B.1 – Modèle avec la variable cep sur les trois transitions : estimations des probabilités de transition $\hat{p}_{ij}(s,t)$ **chez les hommes** basée sur la **méthode paramétrique**.

(s,t)	cep	$\hat{p}_{00}$	$\hat{p}_{01}$	$\hat{p}^0_{02}$	$\hat{p}^1_{02}$	$\hat{p}_{11}$	$\hat{p}_{12}$
(70,80)	oui	0.63 [0.60 ;0.66]	0.06 [0.05 ;0.07]	0.27 [0.24 ;0.29]	0.05 [0.03 ;0.06]	0.24 [0.17 ;0.46]	0.76 [0.54 ;0.83]
	non	0.51 [0.47 ;0.56]	0.10 [0.08 ;0.13]	0.32 [0.28 ;0.37]	0.06 [0.03 ;0.09]	0.36 [0.28 ;0.60]	0.64 [0.40 ;0.72]
(70,90)	oui	0.19 [0.17 ;0.22]	0.04 [0.04 ;0.07]	0.55 [0.51 ;0.59]	0.21 [0.17 ;0.27]	0.01 [0.01 ;0.06]	0.99 [0.94 ;0.99]
	non	0.09[0.07 ;0.11]	0.07[0.06 ;0.11]	0.57 [0.51 ;0.64]	0.27 [0.19 ;0.32]	0.05 [0.03 ;0.15]	0.95 [0.85 ;0.97]
(80,90)	oui	0.30 [0.28 ;0.33]	0.06 [0.05 ;0.09]	0.46 [0.41 ;0.50]	0.17 [0.14 ;0.21]	0.05 [0.03 ;0.13]	0.95 [0.87 ;0.97]
	non	0.17[0.14 ;0.21]	0.11 [0.09 ;0.15]	0.49 [0.44 ;0.55]	0.22 [0.16 ;0.27]	0.13 [0.09 ;0.26]	0.87[0.74 ;0.91]
(80,100)	oui	0.02 [0.01 ;0.03]	0.01 [0.01 ;0.02]	0.63 [0.56 ;0.69]	0.35 [0.28 ;0.41]	0.00 [0.00 ;0.00]	1.00 [1.00 ;1.00]
	non	0.00 [0.00 ;0.01]	0.01[0.00 ;0.02]	0.58 [0.51 ;0.65]	0.40 [0.33 ;0.47]	0.00[0.00 ;0.02]	1.00 [0.98 ;1.00]

Tableau B.2 – Modèle avec la variable cep sur les trois transitions : estimations des probabilités de transition $\hat{p}_{ij}(s,t)$ **chez les femmes** basée sur la **méthode paramétrique**.

(s,t)	cep	$\hat{p}_{00}$	$\hat{p}_{01}$	$\hat{p}^0_{02}$	$\hat{p}^1_{02}$	$\hat{p}_{11}$	$\hat{p}_{12}$
(70,80)	oui	0.75 [0.73 ;0.77]	0.09 [0.07 ;0.10]	0.13 [0.11 ;0.15]	0.03 [0.02 ;0.04]	0.50 [0.42 ;0.62]	0.50 [0.38 ;0.58]
	non	0.69 [0.67 ;0.72]	0.13 [0.11 ;0.15]	0.13 [0.11 ;0.16]	0.05 [0.03 ;0.06]	0.52 [0.44 ;0.64]	0.48 [0.36 ;0.56]
(70,90)	oui	0.29 [0.26 ;0.31]	0.13 [0.11 ;0.15]	0.35 [0.31 ;0.38]	0.24 [0.20 ;0.27]	0.10 [0.07 ;0.16]	0.90 [0.84 ;0.93]
	non	0.20 [0.17 ;0.22]	0.17[0.15 ;0.20]	0.32 [0.28 ;0.37]	0.31 [0.26 ;0.36]	0.11 [0.08 ;0.18]	0.89 [0.82 ;0.92]
(80,90)	oui	0.38 [0.36 ;0.40]	0.15 [0.13 ;0.17]	0.29 [0.26 ;0.32]	0.18 [0.15 ;0.20]	0.19 [0.15 ;0.26]	0.81 [0.74 ;0.85]
	non	0.28[0.26 ;0.31]	0.20 [0.18 ;0.23]	0.27 [0.23 ;0.32]	0.24 [0.20 ;0.28]	0.21 [0.17 ;0.28]	0.79[0.72 ;0.83]
(80,100)	oui	0.02 [0.02 ;0.03]	0.03 [0.03 ;0.05]	0.44 [0.39 ;0.49]	0.51 [0.45 ;0.56]	0.01 [0.00 ;0.01]	0.99 [0.99 ;1.00]
	non	0.01 [0.00 ;0.01]	0.03[0.02 ;0.04]	0.37 [0.31 ;0.43]	0.60 [0.54 ;0.65]	0.01[0.00 ;0.01]	0.99 [0.99 ;1.00]

Tableau B.3 – Modèle avec la variable cep sur les trois transitions : estimations des probabilités de transition $\hat{p}_{ij}(s,t)$ **chez les hommes** basées sur la **méthode semi-paramétrique**.

(s,t)	cep	$\hat{p}_{00}$	$\hat{p}_{01}$	$\hat{p}^0_{02}$	$\hat{p}^1_{02}$	$\hat{p}_{11}$	$\hat{p}_{12}$
(70,80)	oui	0.63 [0.59 ;0.65]	0.06 [0.04 ;0.07]	0.26 [0.23 ;0.30]	0.06 [0.04 ;0.08]	0.15 [0.05 ;0.30]	0.85 [0.70 ;0.95]
	non	0.51 [0.46 ;0.55]	0.10 [0.08 ;0.13]	0.31 [0.26 ;0.37]	0.08 [0.05 ;0.12]	0.26 [0.12 ;0.44]	0.74 [0.56 ;0.88]
(70,90)	oui	0.19 [0.17 ;0.21]	0.04 [0.03 ;0.05]	0.55 [0.50 ;0.59]	0.22 [0.18 ;0.26]	0.01 [0.00 ;0.02]	0.99 [0.98 ;1.00]
	non	0.09 [0.07 ;0.11]	0.07 [0.05 ;0.09]	0.56 [0.48 ;0.62]	0.28 [0.22 ;0.35]	0.03 [0.01 ;0.08]	0.97 [0.92 ;0.99]
(80,90)	oui	0.30 [0.27 ;0.33]	0.06 [0.05 ;0.08]	0.45 [0.41 ;0.50]	0.18 [0.14 ;0.22]	0.06 [0.03 ;0.09]	0.94 [0.91 ;0.97]
	non	0.18 [0.14 ;0.21]	0.12 [0.09 ;0.14]	0.48 [0.42 ;0.54]	0.23 [0.18 ;0.29]	0.13 [0.08 ;0.21]	0.87 [0.79 ;0.92]
(80,100)	oui	0.02 [0.02 ;0.03]	0.01 [0.00 ;0.01]	0.65 [0.59 ;0.70]	0.32 [0.27 ;0.38]	0.00 [0.00 ;0.00]	1.00 [1.00 ;1.00]
	non	0.00 [0.00 ;0.01]	0.01 [0.00 ;0.02]	0.58 [0.51 ;0.65]	0.40 [0.33 ;0.47]	0.00 [0.00 ;0.01]	1.00 [0.99 ;1.00]

Tableau B.4 – Modèle avec la variable cep sur les trois transitions : estimations des probabilités de transition $\hat{p}_{ij}(s,t)$ **chez les femmes** basées sur la **méthode semi-paramétrique**.

(s,t)	cep	$\hat{p}_{00}$	$\hat{p}_{01}$	$\hat{p}^0_{02}$	$\hat{p}^1_{02}$	$\hat{p}_{11}$	$\hat{p}_{12}$
(70,80)	oui	0.76 [0.74 ;0.79]	0.07 [0.06 ;0.09]	0.13 [0.11 ;0.15]	0.04 [0.03 ;0.05]	0.25 [0.10 ;0.43]	0.75 [0.57 ;0.90]
	non	0.72 [0.68 ;0.74]	0.10 [0.09 ;0.12]	0.13 [0.11 ;0.16]	0.05 [0.03 ;0.07]	0.28 [0.13 ;0.45]	0.72 [0.55 ;0.87]
(70,90)	oui	0.28 [0.25 ;0.30]	0.14 [0.12 ;0.15]	0.34 [0.30 ;0.38]	0.25 [0.22 ;0.29]	0.05 [0.02 ;0.09]	0.95 [0.91 ;0.98]
	non	0.19 [0.17 ;0.22]	0.18 [0.15 ;0.20]	0.32 [0.27 ;0.37]	0.31 [0.26 ;0.36]	0.06 [0.03 ;0.11]	0.94 [0.89 ;0.97]
(80,90)	oui	0.36 [0.34 ;0.39]	0.16 [0.15 ;0.18]	0.27 [0.24 ;0.31]	0.20 [0.17 ;0.23]	0.19 [0.14 ;0.23]	0.81 [0.77 ;0.86]
	non	0.27 [0.24 ;0.30]	0.22 [0.19 ;0.24]	0.26 [0.22 ;0.30]	0.25 [0.21 ;0.30]	0.21 [0.16 ;0.27]	0.79 [0.73 ;0.84]
(80,100)	oui	0.03 [0.02 ;0.04]	0.03 [0.02 ;0.04]	0.44[0.39 ;0.48]	0.51 [0.46 ;0.56]	0.01 [0.00 ;0.01]	0.99 [0.99 ;1.00]
	non	0.01 [0.01 ;0.02]	0.03 [0.02 ;0.04]	0.36 [0.31 ;0.42]	0.60 [0.54 ;0.65]	0.01 [0.00 ;0.02]	0.99 [0.98 ;1.00]

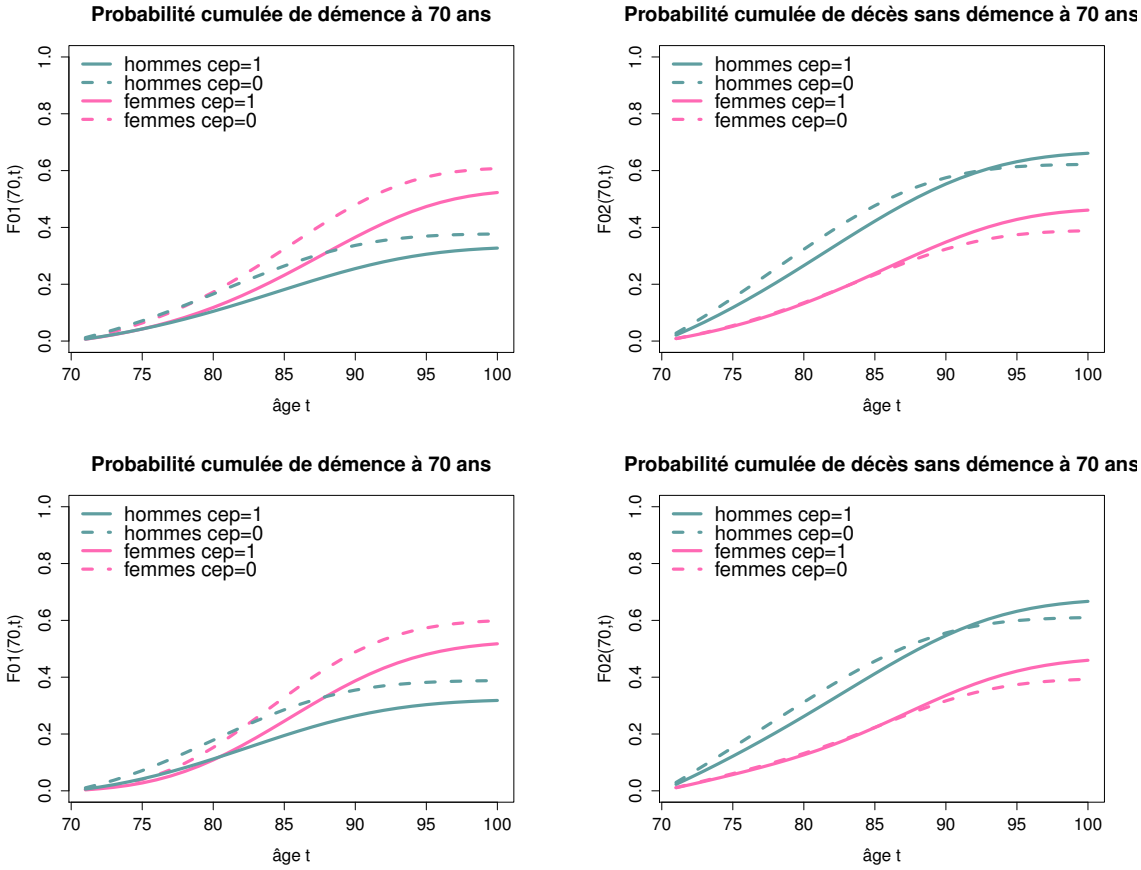


Figure B.1 – Modèle avec le cep : estimations des probabilités cumulées entre 70 ans et t, $71 \leq t \leq 100$ ($F_{01}(70,t)$ à gauche, $F_{02}(70,t)$ à droite) calculées avec les $\alpha_{kl}(.)$ i) de type Weibull et estimées par MV (partie haute), ii) de type splines et estimées par MVP (partie basse).

Tableau B.5 – Modèle avec le cep : estimations des probabilités cumulées $\hat{F}_{01}(s,t)$ et $\hat{F}_{02}(s,t)$ avec intervalles de confiance basées sur la méthode paramétrique (à gauche) semi-paramétrique (à droite) chez les hommes (♂) et les femmes (♀).

		méthode paramétrique		méthode semi-paramétrique	
(s,t)	cep	$\hat{F}_{01}$	$\hat{F}_{02}$	$\hat{F}_{01}$	$\hat{F}_{02}$
hommes					
(70,80)	oui	0.09 [0.08 ;0.13]	0.27 [0.24 ;0.29]	0.11 [0.09 ;0.14]	0.26 [0.23 ;0.30]
	non	0.17 [0.12 ;0.21]	0.32 [0.28 ;0.37]	0.18 [0.13 ;0.24]	0.31 [0.26 ;0.37]
(70,90)	oui	0.26 [0.22 ;0.30]	0.55 [0.51 ;0.59]	0.26 [0.23 ;0.31]	0.55 [0.50 ;0.59]
	non	0.34 [0.27 ;0.40]	0.57 [0.51 ;0.64]	0.35 [0.29 ;0.43]	0.56 [0.48 ;0.62]
(80,90)	oui	0.24 [0.20 ;0.28]	0.46 [0.41 ;0.50]	0.24 [0.20 ;0.29]	0.45 [0.41 ;0.50]
	non	0.33 [0.28 ;0.39]	0.49 [0.44 ;0.55]	0.35 [0.28 ;0.41]	0.48 [0.42 ;0.54]
(80,100)	oui	0.35 [0.29 ;0.42]	0.63 [0.56 ;0.69]	0.33 [0.27 ;0.39]	0.65 [0.59 ;0.70]
	non	0.41 [0.35 ;0.48]	0.58 [0.51 ;0.65]	0.41 [0.34 ;0.48]	0.58 [0.51 ;0.65]
femmes					
(70,80)	oui	0.12 [0.10 ;0.14]	0.13 [0.11 ;0.15]	0.11 [0.09 ;0.13]	0.13 [0.11 ;0.15]
	non	0.17 [0.14 ;0.20]	0.13 [0.11 ;0.16]	0.15 [0.13 ;0.18]	0.13 [0.11 ;0.16]
(70,90)	oui	0.37 [0.33 ;0.40]	0.35 [0.31 ;0.38]	0.39 [0.35 ;0.43]	0.34 [0.30 ;0.38]
	non	0.48 [0.43 ;0.53]	0.32 [0.28 ;0.37]	0.49 [0.44 ;0.54]	0.32 [0.27 ;0.37]
(80,90)	oui	0.33 [0.30 ;0.36]	0.29 [0.26 ;0.32]	0.36 [0.33 ;0.40]	0.27 [0.24 ;0.31]
	non	0.44 [0.40 ;0.49]	0.27 [0.23 ;0.32]	0.47 [0.42 ;0.52]	0.26 [0.22 ;0.30]
(80,100)	oui	0.54 [0.49 ;0.59]	0.44 [0.39 ;0.49]	0.53 [0.49 ;0.59]	0.44 [0.39 ;0.48]
	non	0.63 [0.57 ;0.68]	0.37 [0.31 ;0.43]	0.62 [0.57 ;0.68]	0.36 [0.31 ;0.42]

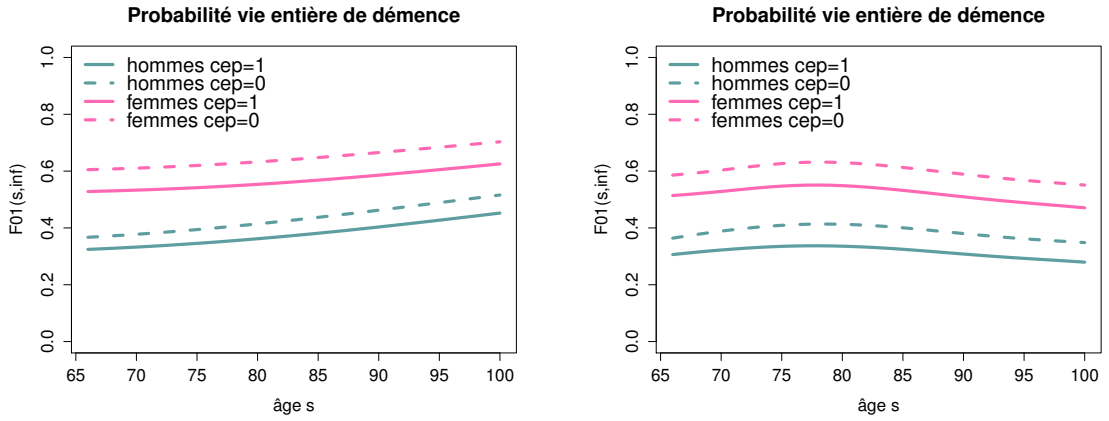


Figure B.2 – Modèle avec le cep : estimations des probabilités de démence au cours de la vie entière à l'âge s, $F_{01}(s,\infty)$, $66 \leq s \leq 100$, avec leurs bandes de confiance, calculées avec les $\alpha_{kl}(.)$ *i*) de type Weibull et estimées par MV (à gauche), *ii*) de type splines et estimées par MVP (à droite).

Tableau B.6 – Modèle avec le cep : estimations de $\hat{F}_{01}(s,\infty)$ la probabilité de démence au cours de la vie entière (*lifetime risk of dementia*) avec intervalles de confiance basées sur la méthode paramétrique (à gauche) semi-paramétrique (à droite).

		méthode paramétrique	méthode semi-paramétrique
s	cep	$\hat{F}_{01}$	$\hat{F}_{01}$
70	oui	0.33 [0.28 ;0.39]	0.32 [0.27 ;0.37]
	non	0.38 [0.31 ;0.44]	0.39 [0.32 ;0.46]
80	oui	0.36 [0.30 ;0.43]	0.34 [0.28 ;0.40]
	non	0.41 [0.35 ;0.48]	0.41 [0.34 ;0.48]
90	oui	0.40 [0.31 ;0.50]	0.31 [0.23 ;0.40]
	non	0.46 [0.37 ;0.56]	0.38 [0.30 ;0.46]

Annexe C : Chapitre IV (sous-section 1.3) : écriture de la vraisemblance

Cette annexe présente une autre façon d'aboutir à l'écriture de la vraisemblance d'un modèle *illness-death* avec trois effets aléatoires U_{01}, U_{02}, U_{12} non indépendants sur chaque transition et troncature à gauche.

Nous rappelons que I est le nombre de groupes, n_i le nombre de sujets dans le groupe i ; a_{ij} l'âge d'entrée dans la cohorte du sujet j du groupe i, a_i le vecteur des âges d'entrée dans la cohorte des sujets du groupe i, $X(a) = 0$ la condition d'entrée dans la cohorte avec a l'âge d'entrée (troncature à gauche).

La vraisemblance s'écrit comme le produit sur les groupes des contributions marginales à la vraisemblance de chaque groupe conditionnellement aux effets aléatoires :

$$\begin{aligned} L &= \prod_{i=1}^{I} \iiint L_i(\theta|u_{01,i}, u_{02,i}, u_{12,i}, X(a_i) = 0)\ \mathrm{d}F(u_{01,i}, u_{02,i}, u_{12,i}|X(a_i) = 0) \\ &= \prod_{i=1}^{I} \iiint \underbrace{L_i(\theta|u_{01,i}, u_{02,i}, u_{12,i}, X(a_i) = 0)}_{(\star)} \underbrace{f(u_{01,i}, u_{02,i}, u_{12,i}|X(a_i) = 0)}_{(\star\star)} \\ &\quad \mathrm{d}u_{01_i}\mathrm{d}u_{02_i}\mathrm{d}u_{12_i} \end{aligned}$$

où $u_{01,i}$, $u_{02,i}$, $u_{12,i}$ sont les réalisations des effets aléatoires pour le groupe i.

Ne connaissant pas la distribution conditionnelle des effets aléatoires, nous nous débarrassons du conditionnement dû à la troncature à gauche en utilisant la formule

de Bayes :

$$\begin{aligned}(\star\star) &= \frac{\mathbb{P}(X(a_i)=0|u_{01,i},u_{02,i},u_{12,i})f(u_{01,i},u_{02,i},u_{12,i})}{\mathbb{P}(X(a_i)=0)} \\ &= \frac{\prod_{j=1}^{n_i}\mathbb{P}(X(a_{ij})=0|u_{01,i},u_{02,i},u_{12,i})f(u_{01,i},u_{02,i},u_{12,i})}{\iint\prod_{j=1}^{n_i}\mathbb{P}(X(a_{ij})=0)f(u_{01,i},u_{02,i})\mathrm{d}u_{01_i}\mathrm{d}u_{02_i}}\end{aligned}$$

De plus :

$$\begin{aligned}(\star) &= \prod_{j=1}^{n_i}L_{ij}(\theta|u_{01,i},u_{02,i},u_{12,i},X(a_i)=0) \\ &= \frac{\prod_{j=1}^{n_i}L_{ij}(\theta|u_{01,i},u_{02,i},u_{12,i})}{\prod_{j=1}^{n_i}\mathbb{P}(X(a_{ij})=0|u_{01,i},u_{02,i},u_{12,i})}\end{aligned}$$

On obtient :

$$L=\prod_{i=1}^{I}\frac{\iiint\prod_{j=1}^{n_i}L_{ij}(\theta|u_{01,i},u_{02,i},u_{12,i})f(u_{01,i},u_{02,i},u_{12,i})\ \mathrm{d}u_{01,i}\mathrm{d}u_{02,i}\mathrm{d}u_{12,i}}{\iint\prod_{j=1}^{n_i}\mathbb{P}(X(a_{ij})=0|u_{01,i},u_{02,i})f(u_{01,i},u_{02,i})\ \mathrm{d}u_{01,i}\mathrm{d}u_{02,i}}$$

qui correspond bien à la formule en IV.2 si l'on note $tr(\theta|u_{01,i},u_{02,i})=\mathbb{P}(X(a_{ij}=0)|u_{01,i},u_{02,i})$.

Détaillons L_{ij} en distinguant les différents cas. Dans un souci de simplification des notations nous omettons dans ce qui suit le conditionnement sur les effets aléatoires $(|u_{01,i},u_{02,i},u_{12,i})$. Notons l (pour *left*) et r (pour *right*) les bornes gauche et droite de l'intervalle de censure d'un sujet dément (r = âge à la visite de diagnostic ; l = âge à la visite précédente). Pour un sujet non observé dément, l est l'âge à la dernière visite. Enfin, notons t l'âge de dernières nouvelles des sujets vivants ou l'âge de décès des sujets décédés.

— Pour un sujet non diagnostiqué dément et non décédé, âgé de l à sa dernière

visite et de t à sa date de dernières nouvelles :

$$L_{ij} = \mathrm{e}^{-A_{01}(t)-A_{02}(t)} + \int_l^t \mathrm{e}^{-A_{01}(u)-A_{02}(u)} \alpha_{01}(u) \mathrm{e}^{-A_{12}(u,t)} \, \mathrm{d}u$$

On rappelle que $A_{kl}(t) = \int_0^t \alpha_{kl}(u)\mathrm{d}u$ et que $A_{kl}(s,t) = \int_s^t \alpha_{kl}(u)\mathrm{d}u = \dfrac{A_{kl}(t)}{A_{kl}(s)}$. Dans le cas particulier où $l = t$, c'est-à-dire lorsque la date de dernière visite correspond aussi à la date de dernière nouvelle :

$$L_{ij} = \mathrm{e}^{-A_{01}(l)-A_{02}(l)}$$

— Pour un sujet non diagnostiqué dément à sa dernière visite et décédé à l'âge t :

$$L_{ij} = \mathrm{e}^{-A_{01}(t)-A_{02}(t)} \alpha_{02}(t) + \int_l^t \mathrm{e}^{-A_{01}(u)-A_{02}(u)} \alpha_{01}(u) \mathrm{e}^{-A_{12}(u,t)} \alpha_{12}(t) \, \mathrm{d}u$$

— Pour un sujet dément, âgé de r à la visite de diagnostic et de l à la visite précédente, qui est toujours vivant à l'âge t :

$$L_{ij} = \int_l^r \mathrm{e}^{-A_{01}(u)-A_{02}(u)} \alpha_{01}(u) \mathrm{e}^{-A_{12}(u,t)} \, \mathrm{d}u$$

— Pour un sujet dément, âgé de r à la visite de diagnostic et de l à la visite précédente, qui est décédé à l'âge t :

$$L_{ij} = \int_l^r \mathrm{e}^{-A_{01}(u)-A_{02}(u)} \alpha_{01}(u) \mathrm{e}^{-A_{12}(u,t)} \alpha_{12}(t) \, \mathrm{d}u$$

Annexe C : Chapitre IV (sous-section 1.3) : écriture de la vraisemblance

Références bibliographiques

Aalen, O. (1980). A model for nonparametric regression analysis of counting processes. In *Mathematical statistics and probability theory*, pages 1–25. Springer. 19

Aalen, O. O. (1975). *Statistical inference for a family of counting processes.* PhD thesis, Department of Statistics, University of California, Berkeley. 26

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6(4) :701–726. 15, 26

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in medicine*, 8(8) :907–925. 19

Aalen, O. O. (1994). Effects of frailty in survival analysis. *Statistical Methods in Medical Research*, 3(3) :227–243. 21

Aalen, O. O., Farewell, V. T., de Angelis, D., Day, N. E., and Nöel Gill, O. (1997). A markov model for hiv disease progression including the effect of hiv diagnosis and treatment : application to AIDS prediction in England and Wales. *Statistics in medicine*, 16(19) :2191–2210. 76, 143

Aalen, O. O. and Johansen, S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5 :141–150. 26, 63

Al Hazzouri, A. Z., Haan, M. N., Kalbfleisch, J. D., Galea, S., Lisabeth, L. D., and Aiello, A. E. (2011). Life-course socioeconomic position and incidence of dementia and cognitive impairment without dementia in older Mexican Americans : results from the Sacramento area Latino study on aging. *American journal of epidemiology*, 173(10) :1148–1158. 40

Andersen, P. K., Abildstrom, S. Z., and Rosthøj, S. (2002). Competing risks as a multi-state model. *Statistical Methods in Medical Research*, 11(2) :203–215. 29

Andersen, P. K. and Borgan, O. (1985). Counting process models for life history data : a review. *Scandinavian Journal of Statistics*, 12 :97–158. 26

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer Verlag. 22

Andersen, P. K., Geskus, R. B., de Witte, T., and Putter, H. (2012). Competing risks in epidemiology : possibilities and pitfalls. *International journal of epidemiology*, 41(3) :861–870. 27

Andersen, P. K. and Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in medicine*, 31(11-12) :1074–1088. 28, 70

Andersen, P. K. and Klein, J. P. (2007). 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. 64

Andersen, P. K. and Perme, M. P. (2010). Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, 19(1) :71–99. 64

Banerjee, S., Wall, M. M., and Carlin, B. P. (2003). Frailty modeling for spatially correlated survival data, with application to infant mortality in minnesota. *Biostatistics*, 4(1) :123–142. 116

Barrett, J. K., Siannis, F., and Farewell, V. T. (2011). A semi-competing risks model for data with interval-censoring and informative observation : An application to the MRC cognitive function and ageing study. *Statistics in Medicine*, 30(1) :1–10. 130

Benichou, J. (2005). Absolute risk. *Encyclopedia of biostatistics*. 64

Berr, C., Vercambre, M.-N., and Akbaraly, T. N. (2009). Épidémiologie de la maladie d'Alzheimer : aspects méthodologiques et nouvelles perspectives. *Psychologie et NeuroPsychiatrie du vieillissement*, 7(1) :7–14. 4

Betensky, R. A., Lindsey, J. C., Ryan, L. M., and Wand, M. (2002). A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine*, 21(2) :263–275. 6

Beyersmann, J., Allignol, A., and Schumacher, M. (2011). *Competing Risks and Multistate Models with R*. Use R ! Springer. 119, 131, 136

Beyersmann, J., Latouche, A., Buchholz, A., and Schumacher, M. (2009). Simulating competing risks data in survival analysis. *Statistics in Medicine*, 28(6) :956–971. 131

Bretagnolle, J. and Huber-Carol, C. (1988). Effects of omitting covariates in Cox's model for survival data. *Scandinavian Journal of Statistics*, pages 125–138. 21

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1) :141–151. 21

Commenges, D., Joly, P., Gégout-Petit, A., and Liquet, B. (2007). Choice between semiparametric estimators of Markov and non-Markov multi-state models from coarsened observations. *Scandinavian Journal of Statistics*, 34(1) :33–52. 36, 100

Commenges, D., Letenneur, L., Joly, P., Alioum, A., and Dartigues, J. F. (1998). Modelling age-specific risk : application to dementia. *Statistics in Medicine*, 17(17) :1973–1988. 58

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220. 19

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62 :269–276. 136

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. Chapman an Hall. 24

Cressie, N. A. (1993). Statistics for spatial data, revised edition. 116

Dartigues, J.-F., Helmer, C., Letenneur, L., Péres, K., Amieva, H., Auriacombe, S., Orgogozo, J.-M., Commenges, D., Jacqmin-Gadda, H., Richard-Harston, S., et al. (2012). Paquid 2012 : illustration et bilan. *Gériatrie et Psychologie Neuropsychiatrie du Vieillissement*, 10(3) :325–331. 4, 116

De Wreede, L. C., Fiocco, M., and Putter, H. (2010). The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. *Computer methods and programs in biomedicine*, 99(3) :261–274. 27, 119, 136

De Wreede, L. C., Fiocco, M., and Putter, H. (2011). mstate : An R package for the analysis of competing risks and multi-state models. *Journal of Statistical Software*, 38(7) :1–30. 27, 119

Fine, J. P. (2001). Regression modeling of competing crude failure probabilities. *Biostatistics*, 2(1) :85–97. 64

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446) :496–509. 64

Finkelstein, D. M. and Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*, pages 933–945. 6

Fratiglioni, L., Viitanen, M., Von Strauss, E., Tontodonati, V., Herlitz, A., and Winblad, B. (1997). Very old women at highest risk of dementia and Alzheimer's disease : incidence data from the Kungsholmen Project, Stockholm. *Neurology*, 48(1) :132–138. 58

Freitag, M. H., Peila, R., Masaki, K., Petrovitch, H., Ross, G. W., White, L. R., and Launer, L. J. (2006). Midlife pulse pressure and incidence of dementia the Honolulu-Asia aging study. *Stroke*, 37(1) :33–37. 42

Frydman, H. (1995). Nonparametric estimation of a Markov illness-death process from interval-censored observations, with application to diabetes survival data. *Biometrika*, 82(4) :773–789. 7, 31, 64

Frydman, H. and Szarek, M. (2009). Nonparametric estimation in a Markov illness-death process from interval censored observations with missing intermediate transition status. *Biometrics*, 65(1) :143–151. 31

Frydman, H. and Szarek, M. (2010). Estimation of overall survival in an illness-death model with application to the vertical transmission of HIV-1. *Statistics in Medicine*, 29(19) :2045–2054. 31, 64

Gamerman, D. (1991). Dynamic bayesian models for survival data. *Applied Statistics*, pages 63–79. 19, 116

Gerds, T. A., Scheike, T. H., and Andersen, P. K. (2012). Absolute risk regression for competing risks : interpretation, link functions, and prediction. *Statistics in Medicine*, 31(29) :3921–3930. 64

Gill, R. D. and Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *The annals of statistics*, pages 1501–1555. 26

Gooley, T. A., Leisenring, W., Crowley, J., Storer, B. E., et al. (1999). Estimation of failure probabilities in the presence of competing risks : new representations of old estimators. *Statistics in medicine*, 18(6) :695–706. 64

Hall, C., Derby, C., LeValley, A., Katz, M., Verghese, J., and Lipton, R. (2007). Education delays accelerated decline on a memory test in persons who develop dementia. *Neurology*, 69(17) :1657–1664. 100

Hougaard, P. (1995). Frailty models for survival data. *Lifetime data analysis*, 1(3) :255–273. 21

Hougaard, P. (1999). Multi-state models : a review. *Lifetime data analysis*, 5(3) :239–264. 22

Hougaard, P. (2000). *Analysis of multivariate survival data*. Springer. 21

Hubbard, R. A., Inoue, L., and Fann, J. (2008). Modeling nonhomogeneous markov processes via time transformation. *Biometrics*, 64(3) :843–850. 30

Izmirlian, G., Brock, D., Ferrucci, L., and Phillips, C. (2000). Active life expectancy from annual follow-up data with missing responses. *Biometrics*, 56(1) :244–248. 65

Jackson, C. H. (2011). Multi-state models for panel data : the msm package for R. *Journal of Statistical Software*, 38(8) :1–29. 30, 119, 136

Jacqmin-Gadda, H., Commenges, D., and Dartigues, J.-F. (2006). Random changepoint model for joint modeling of cognitive decline and dementia. *Biometrics*, 62(1) :254–260. 129

Janssen, J. and Limnios, N. (1999). *Semi-Markov models and applications*. Kluwer Academic Publishers Dotrecht. 24

Joly, P., Commenges, D., Helmer, C., and Letenneur, L. (2002). A penalized likelihood approach for an illness-death model with interval-censored data : application to age-specific incidence of dementia. *Biostatistics*, 3(3) :433–443. 7, 62, 125, 136, 141

Joly, P., Touraine, C., Georget, A., Dartigues, J.-F., Commenges, D., and Jacqmin-Gadda, H. (2013). Prevalence projections of chronic diseases and impact of public health intervention. *Biometrics*. 4

Kalbfleisch, J. D. and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80(392) :863–871. 29, 63

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons. 64

Katzman, R., Aronson, M., Fuld, P., Kawas, C., Brown, T., Morgenstern, H., Frishman, W., Gidez, L., Eder, H., and Ooi, W. L. (1989). Development of dementing illnesses in an 80-year-old volunteer cohort. *Annals of neurology*, 25(4) :317–324. 100

Kay, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, 42(4) :855–865. 30

Klein, J. P. and Andersen, P. K. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics*, 61(1) :223–229. 64

Law, C. G. and Brookmeyer, R. (1992). Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in medicine*, 11(12) :1569–1578. 62

Leffondré, K., Touraine, C., Helmer, C., and Joly, P. (2013). Interval-censored time-to-event and competing risk with death : is the illness-death model more accurate than the Cox model ? *International journal of epidemiology*, 42(4) :1177–1186. 39, 62, 127, 136

Letenneur, L., Gilleron, V., Commenges, D., Helmer, C., Orgogozo, J., and Dartigues, J. (1999). Are sex and educational level independent predictors of dementia and Alzheimer's disease ? Incidence data from the paquid project. *Journal of Neurology, Neurosurgery and Psychiatry*, 66(2) :177–183. 144

Levenberg, K. (1944). A method for the solution of certain problems in least squares. *Quarterly of applied mathematics*, 2 :164–168. 37, 139

Lièvre, A., Brouard, N., and C, H. (2003). The estimation of life expectancies from cross-longitudinal surveys. *Mathematical Population Studies*, 10(4) :211–248. 65

Lobo, A., Launer, L., Fratiglioni, L., Andersen, K., Di Carlo, A., Breteler, M., Copeland, J., Dartigues, J.-F., Jagger, C., Martinez-Lage, J., et al. (2000). Prevalence of dementia and major subtypes in Europe : a collaborative study of population-based cohorts. *Neurology*, 54(11) :S4–S9. 4

Mandel, M. (2013). Simulation-based confidence intervals for functions with complicated derivatives. *The American Statistician*, 67(2) :76–81. 75, 142

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(3) :431–441. 37, 139

Meira-Machado, L., de Uña-Álvarez, J., and Cadarso-Suarez, C. (2006). Nonparametric estimation of transition probabilities in a non-Markov illness–death model. *Lifetime Data Analysis*, 12(3) :325–344. 63

Meira-Machado, L. and Roca-Pardiñas, J. (2011). p3state.msm : Analyzing survival data from an illness-death model. *Journal of Statistical Software*, 38(3) :1–18. 119

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4) :945–966. 15

Omar, R. Z., Stallard, N., and Whitehead, J. (1995). A parametric multistate model for the analysis of carcinogenicity experiments. *Lifetime data analysis*, 1(4) :327–346. 30

O'Sullivan, F. (1988a). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing*, 9(2) :363–379. 36, 141, 142

O'Sullivan, F. (1988b). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal on Scientific and Statistical Computing*, 9(3) :531–542. 102

Pan, W. and Chappell, R. (2002). Estimation in the Cox proportional hazards model with left-truncated and interval-censored data. *Biometrics*, 58(1) :64–70. 6

Petersen, J. H., Andersen, P. K., and Gill, R. D. (1996). Variance components models for survival data. *Statistica Neerlandica*, 50(1) :193–211. 21

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics : Competing risks and multi-state models. *Statistics in Medicine*, 26 :2389–2430. 26, 27, 67

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Annals of Statistics*, 11 :453–466. 18, 26

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, 3(4) :425–441. 35, 141

Rondeau, V., Commenges, D., and Joly, P. (2003). Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime data analysis*, 9(2) :139–153. 103, 108, 111, 112, 115

Rosenberg, P. S. (1995). Hazard function estimation using B-splines. *Biometrics*, pages 874–887. 18

Scheike, T. (2001). A generalized additive regression model for survival times. *Annals of Statistics*, 29(5) :1344–1360. 100

Stern, Y. (2002). What is cognitive reserve ? Theory and research application of the reserve concept. *Journal of the International Neuropsychological Society*, 8(03) :448–460. 100

Stern, Y. (2003). The concept of cognitive reserve : a catalyst for research. *Journal of Clinical and Experimental Neuropsychology*, 25(5) :589–593. 100

Titman, A. C. (2011). Flexible nonhomogeneous Markov models for panel observed data. *Biometrics*, 67(3) :780–787. 30

Touraine, C., Gerds, T. A., and Joly, P. (2013a). The Smoothhazard package for R : Fitting regression models to interval-censored observations of illness-death models. *Research report 13/12. Department of Biostatistics, University of Copenhagen.* 128

Touraine, C., Helmer, C., and Joly, P. (2013b). Predictions in an illness-death model. *Statistical methods in medical research.* 65, 128, 142

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1) :20–22. 28

van den Hout, A. (2013). Elect : Estimation of life expectancies using continuous-time multi-state survival models. 120

van den Hout, A. and Matthews, F. E. (2008). Multi-state analysis of cognitive ability data : a piecewise-constant model and a Weibull model. *Statistics in Medicine*, 27(26) :5440–5455. 65, 75

van den Hout, A. and Matthews, F. E. (2009). A piecewise-constant Markov model and the effects of study design on the estimation of life expectancies in health and ill health. *Statistical Methods in Medical Research*, 18(2) :145. 65

van den Hout, A. and Matthews, F. E. (2010). Estimating stroke-free and total life expectancy in the presence of non-ignorable missing values. *Journal of the Royal Statistical Society Series A Statistics in Society*, 173(2) :331–349. 75

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3) :439–454. 21

Vaupel, J. W. and Yashin, A. I. (1985). Heterogeneity's ruses : some surprising effects of selection on population dynamics. *The American Statistician*, 39(3) :176–185. 21

Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer. 22, 106

Xu, J., Kalbfleisch, J. D., and Tai, B. (2010). Statistical analysis of illness–death processes and semicompeting risks data. *Biometrics*, 66(3) :716–725. 29

Yu, B. and Ghosh, P. (2010). Joint modeling for cognitive trajectory and risk of dementia in the presence of death. *Biometrics*, 66(1) :294–300. 129

Zheng, M. and Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1) :127–138. 28

Résumé : Lorsqu'on étudie la démence à partir de données de cohorte, les sujets sont suivis par intermittence ce qui donne lieu à des temps d'apparition de la démence censurés par intervalle et ont un risque important de décès, d'où un nombre non négligeable de sujets qui décèdent sans avoir été diagnostiqués déments. Le modèle adapté à l'étude de la démence dans ce contexte est un modèle *illness-death* dans lequel les sujets initialement non malades peuvent transiter vers l'état décédé directement ou en passant par l'état malade. La vraisemblance du modèle permet en particulier de tenir compte du fait que les sujets décédés sans diagnostic de démence ont pu passer par deux chemins différents entre leur dernière visite et leur décès. Elle ne se factorise pas comme dans le cas où les différents temps de transition sont connus exactement ; tous les paramètres sont donc estimés conjointement. Or, une pratique courante lorsqu'on s'intéresse aux facteurs de risque de démence consiste à considérer uniquement la transition de l'état non malade à l'état malade. Afin de pouvoir appliquer les techniques d'analyse de survie classiques, les sujets décédés sans diagnostic de démence sont artificiellement censurés à droite à leur dernière visite. La première partie de cette thèse permet de montrer que cette approche, contrairement à l'approche illness-death, peut induire des biais dans l'estimation des effets des facteurs de risque. Le fait de modéliser le décès en plus de la démence permet aussi d'exprimer des quantités directement liées au décès comme des espérances de vie ou le risque absolu de démence au cours de la vie entière. Dans la deuxième partie de cette thèse, nous nous efforçons de dégager toutes les quantités pertinentes d'un point de vue épidémiologique qui peuvent être exprimées dans un contexte *illness-death*. Elles peuvent être estimées en plus des différentes intensités de transition et des effets des facteurs de risque à l'aide du paquet R **SmoothHazard**, développé au cours de cette thèse. Enfin, la dernière partie de cette thèse consiste à prendre en compte l'hétérogénéité de nos données. Nous introduisons des effets aléatoires sur les trois transitions du modèle *illness-death* afin de prendre en compte des facteurs de risque partagés par les sujets appartenant à un même groupe.
Mots clés : analyse de survie, modèle *illness-death*, données censurées par intervalle, démence, modèles à fragilité

Abstract: In dementia research, difficulties arise when studying cohort data. Time-to-disease onset is interval censored because the diagnosis is made at intermittent follow-up visits. As a result, disease status at death is unknown for subjects who are disease-free at the last visit before death. The illness-death model allows initially disease-free subjects to first become ill and then die, or die directly. Those two possible trajectories of the subjects who died without dementia diagnosis can be taken into account into the likelihood. Unlike the case where transition times are exactly observed, the latter do not factorizes and parameters of the three transitions have to be estimated jointly. However, when studying risk factors of dementia, a common approach consists in artificially ending follow-up of subjects who died without dementia diagnosis by considering them as right censored at the last time they were seen without disease. The first part of the present work shows that this approach (unlike the illness-death modeling approach) can lead to biases when estimating risk factor effects of dementia. Modeling death in addition to disease also allows to consider quantities which are closely related with risk of death, like lifetime risk of disease or life expectancies. In the second part of this work, we detail all the quantities which are of epidemiological interest in an illness-death model. They can be estimated, in addition to the transition intensities and the effects or risk factors, using the R package **SmoothHazard** which has been implemented during this thesis. Finally, in the last part of this work, we consider shared frailty regression models for the three transitions of the illness-death model.
Keywords: survival analysis, illness-death model, interval-censored data, dementia, frailty models