Université de Montréal

Validation des modèles statistiques tenant compte des variables dépendantes du temps en prévention primaire des maladies cérébrovasculaires

par

Loredana Kis

Département de mathématiques et de statistique

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures

en vue de l'obtention du grade de

Maître ès sciences (M.Sc.)
en Statistique

juillet 2012

© Loredana Kis, 2012

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Validation des modèles statistiques tenant compte des variables dépendantes du temps en prévention primaire des maladies cérébrovasculaires

présenté par

Loredana Kis

a été évalué par un jury composé des personnes suivantes :

Pierre Duchesne
(président-rapporteur)

Jean-François Angers
(directeur de recherche)

Sylvie Perreault
(co-directeur)

Alejandro Murua
(membre du jury)

Mémoire accepté le:
Date d'acceptation

SOMMAIRE

L'intérêt principal de cette recherche porte sur la validation d'une méthode statistique en pharmaco-épidémiologie. Plus précisément, nous allons comparer les résultats d'une étude précédente réalisée avec un devis cas-témoins niché dans la cohorte utilisé pour tenir compte de l'exposition moyenne au traitement :

- aux résultats obtenus dans un devis cohorte, en utilisant la variable exposition variant dans le temps, sans faire d'ajustement pour le temps passé depuis l'exposition ;
- aux résultats obtenus en utilisant l'exposition cumulative pondérée par le passé récent ;
- aux résultats obtenus selon la méthode bayésienne.

Les covariables seront estimées par l'approche classique ainsi qu'en utilisant l'approche non paramétrique bayésienne. Pour la deuxième le moyennage bayésien des modèles sera utilisé pour modéliser l'incertitude face au choix des modèles. La technique utilisée dans l'approche bayésienne a été proposée en 1997 mais selon notre connaissance elle n'a pas été utilisée avec une variable dépendante du temps. Afin de modéliser l'effet cumulatif de l'exposition variant dans le temps, dans l'approche classique la fonction assignant les poids selon le passé récent sera estimée en utilisant des splines de régression.

Afin de pouvoir comparer les résultats avec une étude précédemment réalisée, une cohorte de personnes ayant un diagnostique d'hypertension sera construite en utilisant les bases des données de la RAMQ et de Med-Echo.

Le modèle de Cox incluant deux variables qui varient dans le temps sera utilisé. Les variables qui varient dans le temps considérées dans ce mémoire sont

la variable dépendante (premier évènement cérébrovasculaire) et une des variables indépendantes, notamment l'exposition.

Mots clefs : modèle de Cox, B-spline, moyennage bayésien des modèles, analyse de survie.

SUMMARY

The main interest of this research is the validation of a statistical method in pharmacoepidemiology. Specifically, we will compare the results of a previous study performed with a nested case-control which took into account the average exposure to treatment to :

– results obtained in a cohort study, using the time-dependent exposure, with no adjustment for time since exposure ;
– results obtained using the cumulative exposure weighted by the recent past ;
– results obtained using the Bayesian model averaging.

Covariates are estimated by the classical approach and by using a nonparametric Bayesian approach. In the later, the Bayesian model averaging will be used to model the uncertainty in the choice of models. To model the cumulative effect of exposure which varies over time, in the classical approach the function assigning weights according to recency will be estimated using regression splines.

In order to compare the results with previous studies, a cohort of people diagnosed with hypertension will be constructed using the databases of the RAMQ and Med-Echo.

The Cox model including two variables which vary in time will be used. The time-dependent variables considered in this paper are the dependent variable (first stroke event) and one of the independent variables, namely the exposure.

Keywords : Cox model, B-spline, Bayesian model averaging, survival analysis.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

REMERCIEMENTS

J'aimerais tout d'abord remercier Pr. Jean-François Angers pour l'aide et le support apporté tout au long de la maîtrise au Département de mathématiques et statistique de l'Université de Montréal.

J'aimerais aussi remercier Pr. Sylvie Perreault de la Faculté de pharmacie de l'Université de Montréal pour l'opportunité de travailler sur ce projet en pharmaco-épidémiologie. Ce travail m'a permis de m'initier à plusieurs notions spécifiques au domaine de l'épidémiologie.

J'aimerais aussi remercier les membres et collègues du département de mathématiques et statistique pour l'ambiance scolaire et humaine qu'ils ont créée et qui a rendu les conditions de travail très agréables. Un gros merci à Miquele Nassoni et Guillaume Provencher qui ont eu la patience de répondre à toutes mes questions en matière d'« informatique ».

Finalement j'aimerais remercier mon mari Dan et ma fille Maria qui m'ont soutenu et encouragé durant mes années d'études.

INTRODUCTION

Les maladies cérébrovasculaires sont une des causes les plus importantes de mortalité et d'invalidité dans le monde (Thom *et al.*, 2006). L'hypertension artérielle représente le facteur le plus important qui peut modifier le risque des maladies cérébrovasculaires (Goldstein *et al.*,2010).

Des recherches cliniques ont montré que les antihypertenseurs peuvent être associées à une baisse de 30% à 40% de l'incidence des maladies cérébrovasculaires en quelques années seulement(Collins et MacMahon, 1994). La prise de médicaments antihypertenseurs a été prouvée efficace pour le contrôle de la pression sanguine de même que pour la réduction de la morbidité et la mortalité cardiovasculaire. Toutefois cette baisse est fortement reliée à l'adhérence à la thérapie.

Diverses études pharmaco-épidémiologiques estiment l'effet de l'exposition dépendante du temps, où l'exposition et son intensité peuvent varier dans le temps. Dans ce cas, l'exposition est une variable complexe (variable dans le temps, d'intensité et des périodes d'exposition variables) et l'estimation de l'association entre le traitement et l'issu étudié est plus susceptible de biais. Il faut donc tenir compte de l'incertitude sur la façon dont l'effet s'accumule dans le temps, de l'importance des périodes d'exposition, ou des doses utilisées durant le suivi. Le concept de dose cumulative pondérée par le passé récent est présent dans la littérature spécifique à ce domaine depuis plus de 20 ans (Thomas, 1988). Malgré cela, son utilisation est restreinte à quelques études. Une de ces études est celle sur les blessures causées par des chutes associées à l'utilisation de la benzodiazépine chez les personnes âgées réalisée par Abrahamowicz *et al.* (2006). Ils ont constaté que des combinaisons des covariables qui varient dans le temps (dose cumulative, dose courante, durée de l'utilisation) pondérées par le passé récent ont amélioré les estimations

par rapport à des modèles plus simples. Plus récemment, Sylvestre et Abrahamowicz (2009) ont considéré une méthode flexible pour la modélisation des effets cumulatifs des expositions variables dans le temps. Ils ont utilisé la régression spline cubique pour estimer la fonction qui assigne les poids à des doses prises dans le passé.

Un modèle ne prenant pas en considération la variation dans le temps de l'exposition peut conduire à des conclusions biaisées. Donc, le but de ce projet est d'évaluer l'impact de considérer la présence de l'exposition, de son intensité et des périodes d'exposition qui varient dans le temps de même que les risques compétitifs sur les estimés évaluant l'impact de l'adhésion aux agents antihypertenseurs sur l'apparition des événements cérébrovasculaires.

Chapitre 1

MOTIVATION

1.1. Contexte

L'accident vasculaire cérébral représente l'une des principales causes de mortalité et d'invalidité dans le monde entier (Thom *et al.*, 2006). La morbidité qui y associée est très grande. Sur 100 personnes ayant subi un accident vasculaire cérébral, 25 se rétablissent avec une déficience mineure, 40 souffrent d'une incapacité modérée ou grave et 10 ont un handicap si important qu'elles nécessitent des soins de longue durée, parfois à vie. Les répercussions économiques de l'accident vasculaire cérébral sont tout aussi considérables puisque l'accident vasculaire cérébral engendre des dépenses annuelles de 2,7 milliards de dollars pour l'économie canadienne.

Dans ce contexte, la prévention d'un premier épisode s'avère très importante, surtout lorsque nous pensons que plus de 70% des accidents vasculaires cérébraux sont des premiers événements (Goldstein *et al.*, 2006). Plusieurs facteurs contribuent au développement de l'accident vasculaire cérébral (Goldstein *et al.*, 2011), mais l'hypertension artérielle demeure sans contredit le facteur de risque modifiable le plus important.

L'hypertension artérielle est fortement prévalente au sein de la population canadienne, affectant 22% des adultes. En plus de devoir modifier les habitudes de vie, une pharmacothérapie est souvent nécessaire. Plusieurs traitements efficaces et sûrs sont disponibles pour réduire la pression artérielle, tels que les diurétiques, les bêta-bloqueurs, les bloqueurs des canaux calciques, les inhibiteurs de l'enzyme

de conversion de l'angiotensine et les antagonistes des récepteurs à l'angiotensine. Ces médicaments réduisent la morbidité et la mortalité cardiovasculaire, fait démontré dans le cadre de nombreux essais cliniques randomisés, tant en prévention primaire qu'en prévention secondaire (Collins *et al.*, 1994 ; Blood Pressure Lowering Treatment Trialists' Collaboration, 2003).

Le contrôle de la pression artérielle au sein de la population générale est faible, surtout en raison d'une utilisation non optimale du traitement antihypertenseur. En effet, des études observationnelles récentes ont montré que la moitié des patients arrête leur traitement deux ans seulement après l'avoir initié (Perreault *et al.*, 2005). Ce phénomène est très inquiétant d'autant plus que l'effet cardioprotecteur et l'effet cérébroprotecteur de ces thérapies n'apparaissent qu'à la suite d'une exposition minimale d'une année au traitement. L'adhésion est également non optimale et elle varie entre 73% et 90% lors de la première année d'utilisation (Elliott *et al.*, 2007).

Les conséquences de l'utilisation non optimale des thérapies antihypertensives sur les événements cliniques ont été éxaminées dans plusieurs études. Une de ces études est celle réalisée par Kettani *et al.*(2009) qui avait pour objectif principal d'évaluer la relation existant entre le niveau d'adhésion au traitement antihypertenseur et la survenue de l'accident vasculaire cérébral non fatal chez des patients sans antécédents cardiovasculaires dans un contexte réel d'utilisation. L'étude a montré qu'une plus grande adhésion aux agents antihypertenseurs est associée à une baisse de 22% du risque d'occurrence des événements cérébrovasculaires.

Dans la recherche en pharmaco-épidémiologie, nous cherchons à établir s'il existe une relation entre une maladie/décès et l'exposition à un médicament. Pour ce faire, des modèles multivariés peuvent être utilisés, dont le modèle de Cox. Celui-ci permet d'étudier le délai d'apparition d'un événement et d'exprimer le risque instantané de l'événement. Pour le modèle de Cox traditionnel, un facteur de risque est mesuré habituellement au début de l'étude. Mais souvent, soit l'effet de ce facteur varie dans le temps, soit le facteur lui même change de valeur pendant l'étude et le modèle de Cox permet de tenir compte des variables

qui varient dans le temps. Dans les études épidémiologiques où l'exposition varie avec le temps, nous avons besoin d'un niveau supplémentaire de complexité méthodologique pour rendre compte de la dépendance temporelle de l'exposition.

Afin de déterminer l'effet de l'exposition sur le risque d'un événement (maladie ou décès), nous avons besoin d'un seul proxy de l'exposition. Dans le but d'inclure toute l'information disponible dans un proxy qui prend en considération le moment de l'exposition (ou l'intensité), nous pouvons utiliser la notion de dose cumulative (non pondérée) (Stranges, 2006) représentée par :

$$\sum_t X(t),$$

où $X(t)$ reprèsente la valeur de la variable d'intérêt au moment t. Le problème avec ce genre de modèle est que la dose cumulative (non pondérée) ne prend pas en considération le moment de l'exposition. Ce modèle peut être amélioré pour prendre cet aspect en considération, en utilisant une combinaison linéaire :

$$\sum_t w(t) \times X(t),$$

où $w(t)$ est une fonction de poids.

Une première façon d'aborder ce type de modèle est en spécifiant *a priori* une certaine forme pour la fonction de poids. Ce type de modèle a été utilisé par Vacek (1997), qui a étudié la relation entre l'exposition à l'amiante et le cancer des poumons de même que par Abrahamowicz *et al.* (2006) qui ont regardé les effets adverses des benzodiazépines.

Une modélisation plus flexible inclut l'estimation non paramétrique de la fonction de poids. Hauptmann *et al.*(2000) ont utilisé les B-splines dans le cadre d'un modèle linéaire généralisé pour estimer la fonction de poids. Sylvestre et Abrahamowicz (2009) ont estimé la fonction de poids en utilisant les splines cubiques de régression dans le modèle de Cox. Ils ont constaté que des combinaisons des covariables qui varient dans le temps (dose cumulative, dose courante, durée de l'utilisation) pondérées par la passé récent ont amélioré les estimations par rapport à des modèles plus simples.

Selon nos connaissances, l'association entre l'adhésion au traitement antihypertenseur et le risque d'une maladie cérébrovasculaire qui prend en considération

l'exposition et les périodes d'exposition qui varient dans le temps n'a pas encore été évaluée dans le contexte de la pratique réelle.

1.2. Prévalence

Un des indicateurs classiques pour déterminer le niveau de présence d'une maladie dans la population est la prévalence.

Définition 1.2.1. *La prévalence représente le nombre de fois qu'un événement a été observé dans un milieu déterminé, à un moment donné ou pendant une période écoulée, et sans distinction entre les nouvelles manifestations de cet événement et les anciennes.*

Chaque année, entre 40 000 et 50 000 accidents vasculaires cérébraux sont rapportés au Canada, parmi lesquels 16 000 décès, ce qui en fait la quatrième cause de mortalité dans le pays (Statistique Canada, 2012). Parmi les facteurs de risque pour les maladies cardiovasculaires dans la population générale, nous pouvons mentionner le tabagisme, la sédentarité, l'excès de poids, l'hypertension artérielle, le diabète, la dyslipidémie, les antécédents cardio-vasculaires etc. Tous ces facteurs contribuent au développement de l'accident vasculaire cérébral. L'hypertension artérielle, en particulier, est très répandue et constitue en soi un problème majeur de santé dans le monde entier. Au Canada, nous estimons à environ 22% la proportion des adultes âgés de 18 à 70 ans qui en souffrent et à près de 50% celle de personnes âgées de plus de 65 ans qui en sont affectées (Joffres *et al.*, 2001).

1.3. Efficacité clinique des antihypertenseurs (prévention primaire)

Cinq classes majeures d'antihypertenseurs sont souvent utilisées aujourd'hui dans le traitement de l'hypertension artérielle. Il s'agit des diurétiques, des bêta-bloqueurs (BB), des bloqueurs des canaux calciques (BCC), des inhibiteurs de l'enzyme de conversion de l'angiotensine (IECA) et des antagonistes des récepteurs à l'angiotensine (ARA). À part leurs propriétés antihypertensives établies,

ces agents suscitent un grand intérêt par rapport à leur capacité à réduire la morbidité et la mortalité cardiovasculaire. Ils ont ainsi fait l'objet de nombreuses études cliniques effectuées aussi bien chez des patients sans antécédents cardiovasculaires (prévention primaire) que chez des patients ayant déjà eu un événement cardiovasculaire (prévention secondaire). De nombreuses méta-analyses ont également été effectuées pour quantifier leur effet cardioprotecteur et cérébroprotecteur.

Les premiers essais cliniques contrôlés par placebo conduits en prévention primaire chez des personnes hypertendues d'âge moyen ou avancé avec une hypertension diastolique prédominante, ont pu démontré l'efficacité des diurétiques (principalement hydrochlorothiazide et chlorthalidone) et des BBs dans la réduction du risque d'accident vasculaire cérébral et de maladie coronarienne (Collins *et al.*, 1994). D'autres études ont montré que chez les personnes âgées souffrant d'hypertension artérielle systolique isolée, la thérapie antihypertensive maintenue durant 4 années a diminué la pression artérielle diastolique d'environ 10 mm Hg et réduit le risque d'accident vasculaire cérébral d'environ 30% (Staessen *et al.*, 1999). Durant les deux dernières décennies, les BCC (voir le tableau d'abréviations utilisées à l'annexe A) et des IECA, «nouvelles» classes d'antihypertenseurs, ont été investiguées à leur tour. Des essais contrôlés par placebo ont montré de manière évidente leurs effets cérébrovasculaires bénéfiques (HOPE, 2000 ; Pitt *et al.*, 2000).

1.4. Adhésion et persistence

Comme mentionné auparavant, des études antérieures ont montré que l'utilisation du traitement antihypertenseur est souvent sous-optimale. L'étude de Kettani *et al.* (2009) a démontré qu'une plus grande adhésion est associée à une baisse importante du risque d'événements.

Définition 1.4.1. *L'adhésion est le degré auquel les comportements d'un patient à l'égard de la prise d'un médicament coïncident avec les recommandations médicales.*

Une notion très utilisée dans les études pharmaco-épidemiologiques pour calculer l'adhésion est celle du ratio de possession du médicament (MPR). Dans ce

cas, l'adhésion se calcule par la proportion de journées exposées à un médicament sur une période définie. La définition des catégories d'adhésion varie selon le choix des chercheurs. Le seuil de 80%, qui est celui le plus souvent rencontré dans la littérature, sépare les adhérents des non-adhérents (DiMatteo *et al.*, 2002). Par contre, tous les auteurs s'entendent pour dire que l'adhésion au traitement antihypertenseur est essentielle pour qu'un bénéfice similaire à celui obtenu dans les essais cliniques soit perçu dans la population générale.

Les niveaux d'adhésion observés dans les essais cliniques dépassent souvent 80%. Ceci s'explique par le fait que les essais cliniques représentent un milieu d'expérimentation hautement contrôlé. En effet, les patients sont suivis de façon systématique et ils sont encouragés à prendre correctement leur médicament.

Dans l'étude ALLHAT(2002) par exemple, les patients sous chlorthalidone étaient adhérents à 87,1% à un an de suivi et à 80,5% après 5 ans. Dans le groupe sous amlodipine les patients étaient adhérents à 87,6% au bout d'un an et à 80,4% à 5 ans, et le groupe sous lisinopril avait un taux d'adhérence de 82,4% à un an et de 72,6% à 5 ans. Ces données doivent être interprétées avec prudence car les essais cliniques tentent à maximiser la bonne utilisation des médicaments. De plus, les patients sont suivis intensément et une attention particulière leur est accordée. Enfin, les participants sont souvent plus motivés et plus informés que la population générale.

Définition 1.4.2. *La persistance représente la durée pendant laquelle un patient continue de prendre le médicament qui lui a été prescrit.*

Des études observationnelles faites à partir de bases de données administratives donnent une approximation plus exacte de la persistance au traitement en situation réelle. Celle-ci est généralement plus faible que dans les études cliniques. Caro *et al.* (1999) ont examiné la persistance de 79 591 patients hypertendus au traitement antihypertenseur à partir des bases de données du ministère de la santé de la Saskatchewan. Ils ont ainsi constaté que la persistance au traitement a diminué au cours des six premiers mois qui ont suivi le début du traitement et qu'il a continué à diminuer au cours des quatre années suivantes. Parmi les nouveaux diagnostiqués avec une hypertension artérielle, seulement 78% suivaient

toujours le traitement à la fin de la première année. Par contre, pour les patients avec une hypertension artérielle établie (c'est-à-dire qu'ils n'étaient pas des nouveaux diagnostiqués) la persistance était de 97%. Les patients plus âgés étaient plus persistants que les plus jeunes, et les femmes l'étaient plus que les hommes. Une étude de cohorte rétrospective a été effectuée au Québec par Perreault *et al.* (2005) afin d'évaluer la persistance au traitement de 21 011 patients âgés entre 50 et 64 ans sans antécédents cardiovasculaires et nouvellement traités pour hypertension artérielle essentielle. Selon cette étude, la persistance était de 75% lors des 6 premiers mois et a continué de baisser pour atteindre 55% au bout de trois ans.

1.5. Pertinence de la recherche

La non adhésion et la non persistance aux traitements sont des phénomènes fréquents pour les maladies chroniques, donc nous ne pouvons pas négliger cet aspect. Il est donc important de développer et d'utiliser des méthodes pour mieux évaluer leurs impacts sur l'apparition des différentes maladies.

A l'heure actuelle et selon nos connaissances, aucune étude n'a évalué l'impact de la non-adhésion aux agents antihypertenseurs sur l'incidence d'évènements cardiovasculaires en tenant compte de la variabilité de l'exposition dans le temps et d'expositions différentes. Par contre, il semble logique de penser qu'un médicament mal utilisé peut avoir des conséquences non négligeables sur la santé. Des résultats antérieurs ont montré qu'un niveau d'adhérence aux médicaments antihypertenseurs de plus de 80% des doses prescrites réduit vraisemblablement le risque d'accident vasculaire cérébral non fatal et ce en dehors des contextes hautement contrôlés des essais cliniques (Kettani *et al.*, 2009). Ces réductions du risque ont été observées à partir d'une année d'exposition, confirmant l'importance d'une thérapie à long terme pour prévenir l'apparition des maladies cérébrovasculaires.

Notre étude donne une idée plus réelle de la façon dont sont utilisés les antihypertenseurs et de leur action qu'une étude où le moment de l'exposition n'est pas pris en considération.

Il est important à souligner le fait que les bases de données de la RAMQ (Régie de l'assurance maladie du Québec) sont celles des bénéficiaires du régime publique d'assurance médicament qui représentent près de 43% de la population québécoise. En utilisant ces banques de données nous omettons un segment important de la population, notamment ceux qui bénéficient d'un programme d'assurance privée de médicaments. Toutefois, cette pratique de plus en plus observée dans les études pharmaco-épidémiologiques, permet d'obtenir de grands échantillons et donc le biais de séléction devrait être négligeable. Par contre, un avantage de cette pratique est le fait que les banques de données permettent l'accès à l'historique médical et à la liste complète des médicaments sur ordonnance reçus par un patient sur une longue période de temps.

1.6. Aperçu du mémoire

Dans ce mémoire nous allons estimer une fonction de poids dans le cadre du modèle de Cox. Le chapitre deux introduit le modèle de Cox. Dans le chapitre trois nous présentons les splines avec un exemple. Nous finissons le chapitre trois avec la description du modèle de Cox qui inclut l'exposition cumulative pondérée par le passé récent. Le chapitre quatre décrit en détails le moyennage bayésien des modèles et son implémentation dans le modèle de Cox. Le chapitre cinq présente les résultats de l'application du modèle de Cox sur un devis cas-témoins (Kettani *et al.*, 2009), sur une cohorte sans ajustement pour le temps passé depuis l'exposition, sur la cohorte tenant compte de l'exposition cumulative pondérée par le passé récent et l'application du modèle de moyennage bayésien. Finalement, dans le chapitre six nous présentons nos conclusions.

Chapitre 2

MODÈLE DE COX

2.1. Analyse de survie

L'analyse de survie peut être utilisée pour faire des études sur les effets des traitements, des facteurs pronostiques, des expositions et d'autres covariables sur la fonction. Son objectif est de modéliser la survie d'une population exposée à un certain facteur et de comparer les survies de différentes populations.

Définition 2.1.1. *L'analyse de survie est une classe de procédures statistiques pour estimer la fonction de survie (fonction du temps, commençant par une population en vie à 100% à un moment donné et fournissant toujours le pourcentage en vie de la population aux instants suivants).*

Ainsi, l'analyse de survie étudie le délai d'apparition d'un événement. Dans ce genre de situation les sujets entrent dans l'étude au fur et à mesure qu'elle se déroule. Pour chaque sujet nous connaissons la date d'entrée dans l'étude, appelée « date index », et l'état par rapport à l'événement étudié.

Une des caractéristiques importantes des données de survie est la censure à droite.

Définition 2.1.2. *Un temps de survie est dit censuré à droite au temps t si nous savons seulement qu'il est plus grand que t.*

Dans le calcul de la vraisemblance il faut tenir compte des sujets censurés, de telle sorte que la régression linéaire ne peut pas être utilisée. La régression logistique ne peut pas être utilisée non plus à cause des temps inégaux de suivi, d'où la nécessité de l'analyse de survie.

Un résumé des données de survie est la fonction de survie.

Définition 2.1.3. *La fonction de survie au temps t, notée $S(t)$, est la probabilité de ne pas avoir eu un événement jusqu'au temps t, c'est-à-dire $S(t) = P(T > t)$.*

La fonction de survie est une fonction décroissante avec $S(0) = 1$ et $\lim_{t\to\infty} S(t) = 0$. La formule pour la fonction de survie est donnée par :

$$\begin{aligned} S(t) &= \mathrm{P}(X > t) \\ &= \int_t^\infty f(x)dx, \end{aligned}$$

où $f(x)$ est la densité de la durée de vie. De façon équivalente nous pouvons écrire :

$$f(t) = -S'(t).$$

Un autre résumé des données de survie très utile est la probabilité d'avoir l'événement avant t.

Définition 2.1.4. *La fonction d'incidence cumulative au temps t, $F(t)$, est la probabilité que l'événement a eu lieu avant t (ou bien que la probabilité que le temps de survie est plus grande ou égale à t).*

La fonction de distribution cumulative (ou bien la fonction d'incidence cumulative) de la variable aléatoire *temps de survie* T est donc la probabilité qu'un sujet choisi au hasard ait un temps de survie T plus petit ou égal à une valeur t :

$$F(t) = P(T \leq t). \tag{1}$$

Il est évident à partir de l'équation (1) que $F(t) = 1 - S(t)$.

Un autre concept important pour l'analyse de survie est la fonction de risque.

Définition 2.1.5. *La fonction de risque $h(t)$ (le risque instantané de l'événement) est le taux d'événements à court terme pour les sujets qui n'ont pas eu un événement avant t.*

La formule générale pour calculer la fonction de risque est la suivante (Lawless, 2003) :

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\mathrm{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{\mathrm{P}(\{t \leq T < t + \Delta t\} \bigcap \{T \geq t\})}{\mathrm{P}(T \geq t)\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{\mathrm{P}(t \leq T < t + \Delta t)}{S(t)\Delta t} \\ &= \frac{f(t)}{S(t)} = \frac{-d \log(S(t))}{dt}. \end{aligned}$$

La fonction de vraisemblance pour des données censurées est calculée selon la formule :

$$L(\beta) = \prod_{i=1}^{n} \left\{ [f(t_i, \beta, Z_i)]^{c_i} \times [S(t_i, \beta, Z_i]^{1-c_i} \right\},$$

où t_i est le temps d'événement, Z_i représente la covariable et $c_i = 0$ représente la censure, alors que $c_i = 1$ représente un événement.

Il est aussi utile de définir le concept de *taux de risque cumulatif*, aussi appelé taux de panne cumulatif :

$$\begin{aligned} H(t) &= \int_0^t h(x)dx \\ &= -\log(S(t)). \end{aligned} \tag{2}$$

À partir de l'équation (2) nous pouvons déduire facilement que

$$S(t) = \exp(-H(t)) = \exp\left(\int_0^t h(x)dx \right).$$

2.2. Le modèle de Cox

Le modèle de Cox est un des modèles les plus utilisés pour étudier le délai d'apparition d'un événement. Ce modèle permet d'exprimer le risque instantané de l'événement, c'est-à-dire la probabilité d'apparition de l'événement d'intérêt dans un intervalle de temps $[t, t + \Delta]$ en fonction de t et des autres variables, dites explicatives, sachant que l'événement ne s'est pas réalisé avant t. La variable t représente la différence entre la date de l'événement et la date index (date du début de l'observation).

La formule du risque instantané (taux de panne) est donnée par :

$$h(t|\underset{\sim}{Z}) = h_{\text{o}}(t) \times \exp(\underset{\sim}{\beta}'\underset{\sim}{Z}), \tag{3}$$

où $\underset{\sim}{\beta}$ est le vecteur des coefficients de régression et $\underset{\sim}{Z} = (Z_{\text{1}}, Z_{\text{2}}, \ldots, Z_q)'$ est le vecteur de q variables explicatives. Le risque instantané est donc le produit de deux termes : le premier qui dépend du temps et qui est commun pour tous les individus, $h_{\text{o}}(t)$, appelé risque de base (c'est-à-dire le risque au temps t pour un sujet dont tous les prédicteurs sont o) et le deuxième qui dépend des variables explicatives, $\exp(\underset{\sim}{\beta}'\underset{\sim}{Z})$.

Étant donné la forme pour le taux de panne instantané, le taux de panne cumulatif devient donc :

$$\begin{aligned} H(t|\underset{\sim}{Z}) &= \int_{\text{o}}^{t} h(x)dx \\ &= \int_{\text{o}}^{t} h_{\text{o}}(x) \times \exp(\underset{\sim}{\beta}'\underset{\sim}{Z})dx \\ &= \exp(\underset{\sim}{\beta}'\underset{\sim}{Z}) \times \int_{\text{o}}^{t} h_{\text{o}}(x)dx \\ &= H_{\text{o}}(t) \times \exp(\underset{\sim}{\beta}'\underset{\sim}{Z}). \end{aligned}$$

Le modèle de Cox est un outil flexible pour étudier la relation entre plusieurs prédicteurs et le temps jusqu'à l'événement d'intérêt ou jusqu'à la censure à droite.

Le modèle est basé sur deux hypothèses :

(1) il existe une relation log-linéaire entre la fonction de risque instantané et les covariables :

$$\log\left[h(t|\underset{\sim}{Z})/h_{\text{o}}(t)\right] = \underset{\sim}{\beta}'\underset{\sim}{Z};$$

(2) le rapport des fonctions de risque instantané pour deux sujets qui diffèrent seulement par la caractéristique Z_{1} (les autres caractéristiques étant pareilles) ne dépend pas du temps :

$$\begin{aligned} \frac{h(t|Z_{\text{11}})}{h(t|Z_{\text{12}})} &= \frac{h_{\text{o}}(t) \times \exp(\beta_{\text{1}} Z_{\text{11}})}{h_{\text{o}}(t) \times \exp(\beta_{\text{1}} Z_{\text{12}})} \\ &= \exp\left(\beta_{\text{1}}(Z_{\text{11}} - Z_{\text{12}})\right), \end{aligned}$$

en gardant tous les autres covariables constantes.

Il est donc à noter que le modèle de Cox appartient à la classe des modèles des risques proportionnels (ratio des risques, en anglais hasard ratio, noté par HR). Dans ces modèles, le prédicteur linéaire est lié au rapport des risques par la transformation logarithmique. Si le rapport des risques est constant alors nous pouvons écrire :

$$log(HR) = log\left(\frac{h(t|\underset{\sim}{Z})}{h_0(t)}\right) = \beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_q Z_q. \tag{4}$$

Ainsi, à partir de l'équation (4) le risque $h(t|\underset{\sim}{Z})$ peut être écrit aussi sous la forme :

$$\log[h(t|\underset{\sim}{Z})] = \log[h_0(t)] + \beta_1 Z_1 + \cdots + \beta_q Z_q,$$

qui est un modèle de régression log-linéaire, ce qui signifie que le logarithme du risque change linéairement avec les prédicteurs continus. De même une autre conclusion est évidente à partir de l'équation (3) : ce modèle est multiplicatif, c'est-à-dire que les effets des covariables multiplient le risque de base.

Exemple 2.2.1. *Considérons l'exemple suivant. Nous voulons estimer l'effet du traitement (« T ») sur le risque d'une maladie cardiovasculaire, en ajustant pour la pression artérielle systolique au début de l'étude (SBP0). Le temps de suivi est limité à cinq ans.*

Une cohorte de 300 sujets a été simulée. Les valeurs de pression artérielle systolique au début de l'étude ont été générées selon une distribution normale de moyenne 130 et d'écart-type égal à 15.

Le traitement est une variable binaire, prenant la valeur 1 si le sujet prend le traitement et 0 sinon. Le traitement peut être commencé et interrompu à tout moment pendant le suivi. Afin de générer l'historique du traitement, le traitement au début de l'étude (« T.1 ») a été d'abord généré de sorte que la probabilité d'être traité au départ dépend de la valeur de la pression artérielle systolique (telle que plus la pression est élevée, plus la probabilité d'être traité est grande) :

$$logit(\mathrm{P}[T.1_i = 1]) = -0{,}1 + 0{,}02 \times [SBP0_i - 130].$$

où $i = 1, 2, \ldots, 300$. *Le reste de l'historique du traitement a été généré en supposant que la probabilité d'être traité dépend du traitement précédent :*

$$logit(\mathrm{P}[T.j_i = 1]) = -0{,}5 + \log(1{,}5) \times T.(j-1)_i,$$

pour $j = 2, 3, \ldots, 60$, *donc une valeur à chaque deux mois pour un maximum de 5 ans donc 60 mois.*

Afin de générer des temps de suivi (temps jusqu'à l'événement) l'algorithme permutationnel (MacKenzie et Abrahamowicz, 2002) a été utilisé. Selon cet algorithme, les temps d'événement sont d'abord générés comme suit :

- *générer* t_i *selon une distribution exponentielle de paramètre* $\lambda = -\log(0{,}7)/5$ *;*
- *générer* s_i *à partir d'une distribution uniforme sur l'intervalle* $(0, 10)$ *;*
- *si* $t_i < s_i$ *et* $t_i < 10$, *alors* t_i *est un temps d'événement, sinon* $t_i = min(s_i, 10)$ *est un temps de censure.*

Ensuite, afin d'associer l'historique de traitement et la pression arterielle à un temps d'événement nous utilisons l'échantillonnage pondéré :

- *si* t_i *est un temps de censure nous utilisons des poids égaux pour chaque sujet (c'est-à-dire chaque historique de traitement) ;*
- *si* t_i *est un temps d'événement alors chaque sujet a un poids égal à :*

$$\lambda_i(t|T.t_i, SBP0_i) = \exp(\log(1{,}2) \times \sum_i w(u-t) \times T.t_i + 0{,}03 \times (SBP0_i - 130)).$$

La fonction $w(u-t)$ *est donnée par la relation :* $w(u-t) = 1-(u-t)/24$ *. Pour le sujet sélectionné (s)* $temps_s = t_{(j)}$ *et* $événement_s = 1$, *si* $t_{(j)}$ *est un événement et* $événement_s = 0$ *si* $t_{(j)}$ *est un temps de censure. L'échantillonage continue jusqu'à ce que tous les historiques sont appariés à des temps d'événement.*

Dans le modèle 1, nous appliquons le modèle de Cox pour prendre en considération la pression systolique et le traitement au début de l'étude.

Tableau 2.1: Résultats du modèle de Cox avec covariables fixes dans le temps

	$coef$	$\exp(coef)$	$se(coef)$	z	$\mathrm{P}(>\|z\|)$
$SBP0$	*0,04*	*1,04*	*0,01*	*6,15*	*0,00*
$T.1$	*0,03*	*1,03*	*0,19*	*0,18*	*0,86*

Si nous regardons le tableau 2.1, qui contient les résultats de ce modèle, nous pouvons dire qu'il n'y a pas d'effet significatif du traitement au début de l'étude sur le risque des maladies cardiovasculaires. Donc le rapport des risques pour 2 sujets ayant les caractéristiques $SBP0_1$, $T.1_1$, *respectivement* $SBP0_2$, $T.1_2$ *peut être écrit comme :*

$$HR = \exp(0{,}04 \times (SBP0_2 - SBP0_1) + 0{,}03 \times (T.1_2 - T.1_1)).$$

2.3. Estimation et interprétation des coefficients

Les coefficients de régression sont estimés par la méthode de vraisemblance partielle, calculée comme la probabilité que l'individu subisse l'événement au temps t_i, sachant que les individus à risque au temps t_i ont tous survécu jusqu'à ce moment et que l'un d'eux a subi l'événement au temps t_i. La fonction de vraisemblance est alors donnée par :

$$L(\underset{\sim}{\beta}) = \frac{\prod_{i=1}^{D} h(t|\underset{\sim}{Z_i})}{\sum_{j \in R_i} h(t|\underset{\sim}{Z_j})}, \tag{5}$$

où R_i est l'ensemble de sujets à risque au temps t_i et D représente le dernier temps d'événement. L'expression (5) peut s'écrire comme :

$$L(\underset{\sim}{\beta}) = \prod_{i=1}^{D} \frac{\exp(\underset{\sim}{\beta}\underset{\sim}{Z_i})}{\sum_{j \in R_i} \exp(\underset{\sim}{\beta}\underset{\sim}{Z_j})}.$$

Pour ce qui est de l'interprétation des coefficients, à partir de l'équation (4) nous pouvons dire que β_j est l'augmentation de $\log[h(t|\underset{\sim}{Z})]$ pour une augmentation d'une unité de la covariable Z_j, avec Z_l fixe pour $l \neq j$, ou bien que $exp(\beta_j)$ est le rapport des risques pour une augmentation d'une unité de Z_j.

Posons $HR_j = \exp(\beta_j)$. Si $HR_j < 1$ ($\beta < 0$), l'augmentation de la valeur de la covariable est associée à un risque plus petit, c'est-à-dire que le temps de survie est plus long. Par contre si $HR > 1$ l'augmentation de la valeur du prédicteur entraîne une augmentation du risque, donc une survie plus courte.

De façon générale, les coefficients $\underset{\sim}{\beta}$ représentent l'effet de la caractéristique sur l'apparition de l'événement ; si β_j est nul alors la caractéristique j n'a pas

d'influence sur l'évènement considéré ; si β_j est positif (négatif) et si deux sujets ne diffèrent que par la caractéristique j, des valeurs élevées (faibles) de la caractéristique sont associées à un risque plus élevé (faible).

Prenons maintenant le cas d'une variable binaire codée 0/1 (où 0 signifie l'absence du traitement et 1 signifie la présence du traitement). Dans ce cas, $\exp(\widehat{\beta}_j)$ est le risque d'événement estimé parmi les sujets exposés divisé par le risque d'événement chez les sujets non exposés au traitement (catégorie de référence) et garde l'interprétation de HR_j pour une augmentation d'une unité de Z_j. Si, par contre une variable a plusieurs catégories, par défaut la catégorie avec le score le plus bas est utilisée comme référence. Si dans la catégorie de référence il n'y a pas d'événements, alors le rapport de risques est infini et les tests et les intervalles de confiance sont difficiles à interpréter. La situation peut être corrigée en choisissant une autre catégorie de référence, mais les tests et l'intervalle de confiance pour la catégorie sans événement par rapport à la nouvelle catégorie de référence restent difficiles à interpréter.

Pour les prédicteurs continus, le HR est affecté par l'échelle de mesure et l'augmentation d'une unité peut être non importante. Si nous nous intéressons plutôt à une augmentation de k unités, alors :

$$HR_j = \frac{h_0(t) \times \exp(\beta_j(Z_j + k))}{h_0(t) \times \exp(\beta_j Z_j)} = \exp(\beta_j Z_j + \beta_j k - \beta_j Z_j) = \exp(\beta_j k). \qquad (6)$$

Donc un changement de k unités de la valeur de Z_j multiplie le rapport par $\exp(\beta_j k)$.

Le HR peut aussi être interprété comme un pourcentage de variation du risque, c'est-à-dire qu'un HR de 1,08 signifie une augmentation de 8% pour chaque augmentation d'une unité. De façon similaire, une augmentation de k unités dans le prédicteur implique une augmentation de

$$100 \times (\exp(\beta k) - 1)\%$$

du risque.

Exemple 2.3.1. *Reprenons l'exemple présenté dans la section 2.2. Nous avions obtenu un coefficient de 0,04 pour SBP_0. Nous pouvons donc dire que chaque*

unité supplémentaire dans la pression arterielle augmente le risque d'avoir un événement de 4 % $(\exp(0{,}04) - 1)$*, après ajustement pour le traitement.*

2.4. Estimation de taux de panne de base

Dans le cadre du modèle de Cox nous n'avons pas besoin de spécifier une forme paramétrique pour le taux de panne de base, mais nous exigeons que le modèle pour $\log(HR)$ soit celui spécifié par l'équation (4). Toutefois, il y a des situations (comme dans la validation externe du modèle par exemple qui doit inclure une évaluation de la calibration de la prédiction) qui nécessitent une estimation du taux de panne de base.

Pour estimer le taux de panne de base $h_0(t)$, nous utilisons l'estimateur de Breslow (Breslow, 1972), c'est-à-dire que nous maximisons une vraisemblance conditionnelle aux valeurs obtenues précédemment pour $\underset{\sim}{\beta}$:

$$\begin{aligned} L(\underset{\sim}{\widehat{\beta}}, h_0(t)) &= \prod_{j=1}^{n} \left(h(t_j|\underset{\sim}{Z_j})\right)^{\delta_j} \times S(t_j|\underset{\sim}{Z_j}) \\ &= \prod_{j=1}^{n} h_0(t_j)^{\delta_j} \times \exp(\underset{\sim}{\widehat{\beta}}'\underset{\sim}{Z_j})^{\delta_j} \times \exp\left(-H_0(t_j) \times \exp(\underset{\sim}{\widehat{\beta}}'\underset{\sim}{Z_j})\right) \\ &= \left[\prod_{i=1}^{D} h_0(t_i) \times \exp(\underset{\sim}{\widehat{\beta}}'\underset{\sim}{Z_i})\right] \times \exp\left(-\sum_{j=1}^{n} H_0(t_j) \times \exp(\underset{\sim}{\widehat{\beta}}'\underset{\sim}{Z_j})\right), \quad (7) \end{aligned}$$

où $\delta_j = 1$ si t_j est un temps d'événement et zéro sinon. Nous remarquons que si t_j n'est pas un temps d'événement, l'expression (7) est décroissante par rapport à $h_0(t_j)$. Ainsi, en maximisant la vraisemblance nous obtenons que

$$h_0(t_j) = 0, \text{ si } t_j \text{ n'est pas un temps d'événement} \quad (8)$$

De plus nous savons que

$$H_0(t) = \int_0^t h_0(t),$$

donc en utilisant (8) cela devient :

$$H_0(t) \approx \sum_{j=1}^{n} h_0(t_j) \approx \sum_{i=1}^{D} h_0(t_i)$$

Nous pouvons donc reformuler l'expression (7) :

$$L(\underset{\sim}{\widehat{\beta}}, h_{\mathrm{o}}(t)) = \left[\prod_{i=1}^{D} h_{\mathrm{o}}(t_i) \times \exp(\underset{\sim}{\widehat{\beta}}' \underset{\sim}{Z_i})\right] \times \exp\left[-\sum_{j=1}^{n}\left(\sum_{i=1}^{D} h_{\mathrm{o}}(t_i)\right)\exp(\underset{\sim}{\widehat{\beta}}' \underset{\sim}{Z_j})\right]$$
$$= \left[\prod_{i=1}^{D} h_{\mathrm{o}}(t_i) \times \exp(\underset{\sim}{\widehat{\beta}}' \underset{\sim}{Z_i})\right] \times \exp\left[-\sum_{i=1}^{D} h_{\mathrm{o}}(t_i) \times \left(\sum_{j\in R_i}\exp(\underset{\sim}{\widehat{\beta}}' \underset{\sim}{Z_j})\right)\right].$$

En maximisant cette expression par rapport à $h_{\mathrm{o}}(t_i)$ nous obtenons l'estimateur de Breslow pour le taux de panne de base :

$$\widehat{h}_{\mathrm{o}}(t_i) = \frac{1}{\sum_{j\in R_i}\exp(\underset{\sim}{\beta}' \underset{\sim}{Z_j})},$$

où R_i est l'ensemble des sujets à risque au temps t_i.

2.5. Tests d'hypothèses et intervalles de confiance

Généralement nous voulons tester si la valeur d'un coefficient est significativement differente de 0, c'est-à-dire si le facteur correspondant est significatif.

Le test de Wald est basé sur le fait que les estimateurs du maximum de vraisemblance convergent en loi vers une loi normale centrée autour des vraies valeurs des paramètres et dont la variance est l'inverse de l'information de Fisher :

$$\underset{\sim}{\widehat{\beta}} \to N(\underset{\sim}{\beta}, \Sigma),$$

où la matrice Σ est l'inverse de la matrice d'information de Ficher. L'information de Fisher peut être estimée en prenant la deuxième dérivée du logarithme de la vraisemblance par rapport à $\underset{\sim}{\beta}$. Nous pouvons donc écrire l'élément ij de la matrice comme :

$$(I(\widehat{\beta}))_{ij} = \frac{-\partial^2 \log L}{\partial\beta_i \partial\beta_j}$$
$$= \sum_{i=1}^{D}\left(\frac{\sum_{j\in R_i} exp(\underset{\sim}{\beta}' \underset{\sim}{z_j})\sum_{j\in R_i} z_{jh}z_{jk}exp(\underset{\sim}{\beta}' \underset{\sim}{z_j})}{\left[\sum_{j\in R_i} exp(\beta z_j)\right]^2}\right)$$
$$- \sum_{i=1}^{D}\left(\frac{\sum_{j\in R_i} z_{jh}\underset{\sim}{\beta}' \underset{\sim}{z_j}\sum_{j\in R_i} exp(\beta z_{jk}\underset{\sim}{\beta}' \underset{\sim}{z_j})}{\left[\sum_{j\in R_i} exp(\beta z_j)\right]^2}\right).$$

La statistique du test est donnée par :

$$W = \underset{\sim}{\widehat{\beta}}' I(\underset{\sim}{\widehat{\beta}})\underset{\sim}{\widehat{\beta}} \sim \chi^2_p,$$

où p est le nombre de paramètres explicatifs dans le modèle.

Le test du rapport de vraisemblance (LR) est aussi utilisé pour vérifier l'hypothèse $\underset{\sim}{\beta} = \underset{\sim}{0}$. Le test s'écrit comme :

$$LR = 2 \times \left[\log L(\widehat{\underset{\sim}{\beta}}) - \log L(\underset{\sim}{0})\right] \sim \chi^2_p,$$

où $\log L(\underset{\sim}{0})$ est le logarithme de la vraisemblance partielle evaluée en $\underset{\sim}{\beta} = \underset{\sim}{0}$.

2.6. Variables confondantes

Pour comprendre la notion de « variables confondantes » , il faut clarifier d'abord la notion « d'effet causal » .

Définition 2.6.1. *L'effet causal d'une exposition sur un résultat continu est la différence des valeurs moyennes de la population en présence par rapport à l'absence de l'exposition, tout en gardant les autres variables constantes. Si les moyennes diffèrent alors l'exposition est un déterminant causal du résultat.*

Supposons que nous voulons comparer les valeurs moyennes des deux populations dont l'une est exposée et l'autre est non exposée et considérons l'exemple suivant : soit $x_{1i} = 1$ si le sujet i a été exposé et $x_{1i} = 0$ sinon, et supposons l'existence d'une deuxième variable binaire x_{2i} qui a aussi un effet causal sur le résultat. Alors le modèle considéré dans la population exposée sera :

$$y_{1i} = \beta_0 + \beta_1^c x_{1i} + \beta_2^c x_{2i} + \varepsilon_{1i},$$

où β_0 est la moyenne du résultat si $x_{1i} = x_{2i} = 0$, β_j^c est l'effet causal de X_j ; x_{2i} est la valeur observée de X_2 pour l'individu i. De plus nous supposons que $\mathrm{E}[\varepsilon_{1i}] = 0$ et que les erreurs ne dépendent pas de X_1 et X_2. Ainsi, la moyenne de l'issue dans la population exposée ($x_{1i} = 1$) est donnée par :

$$\mathrm{E}[y_1] = \beta_0 + \beta_1^c + \beta_2^c \times \mathrm{E}_1[X_2],$$

où $\mathrm{E}_1[X_2]$ est la moyenne de X_2 parmi les exposés. De la même façon

$$\mathrm{E}[y_0] = \beta_0 + \beta_2^c \times \mathrm{E}_0[X_2].$$

Donc la différence des moyennes est :

$$\mathrm{E}[y_1] - \mathrm{E}[y_0] = \beta_1^c + \beta_2^c \times \left(\mathrm{E}_1[X_2] - \mathrm{E}_0[X_2]\right).$$

Définition 2.6.2. *Si la différence des moyennes d'un résultat entre deux populations définies par une potentielle variable d'intérêt est différente de l'effet causal sur l'issue (c'est-à-dire* $\mathrm{E}[y_1] - \mathrm{E}[y_0] \neq \beta_1^c$*) alors il s'agit des variables confondues.*

Pour traiter ce problème nous utilisons l'approche suivante : nous calculons d'abord le HR non ajusté, c'est-à-dire en utilisant un modèle qui prend en considération seulement une variable ; ensuite nous calculons le HR ajusté pour les autres variables. L'interprétation du HR ajusté est l'effet du changement d'une unité d'une variable en gardant les autres variables constantes. La baisse du HR non ajusté pour une variable dans le modèle ajusté est typique pour les variables confondues.

2.7. Covariables dépendantes du temps (CDT)

Dans les études pharmaco-épidémiologiques sur les effets des médicaments où l'exposition est une variable complexe (variable dans le temps, d'intensité et de périodes d'exposition variables dans le temps), son estimation est plus susceptible de biais. Il faut donc tenir compte de la façon dont l'effet s'accumule dans le temps, de l'importance des périodes d'exposition, mais aussi des doses utilisées durant le suivi.

Définition 2.7.1. *Une covariable dépendante du temps dans le modèle de Cox est un prédicteur dont les valeurs varient dans le temps.*

En pratique l'utilisation des CDT s'avère assez difficile. Une difficulté majeure est représentée par le fait que dans la majorité de cas, les prédicteurs sont mesurés occasionnellement et pour ce modèle nous avons besoin des valeurs à chaque temps d'événement. Une approche très utilisée est de prendre la dernière valeur de la covariable avant le temps d'intérêt comme valeur présente (Kettani *et al.*, 2009).

Exemple 2.7.1. *Reprenons l'exemple présenté à la section 2.2 et appliquons de nouveau le modèle de Cox en utilisant dans le modèle l'historique des traitements. Voici les résultats du modèle de Cox en utilisant la variable SBP0 et la variable « dose » (qui contient toute l'information sur le traitement d'un sujet). Nous constatons que selon ce modèle, l'effet du traitement devient significatif : la prise*

Tableau 2.2: Résultats du modèle de Cox avec covariable dépendante du temps

	$coef$	$\exp(coef)$	$se(coef)$	z	$\mathrm{P}(>\|z\|)$
SBPo	*0,04*	*1,04*	*0,01*	*6,22*	*0,00*
dose	*0,36*	*1,43*	*0,18*	*1,99*	*0,04*

du traitement augmente le risque des maladies cardiovasculaire de 43% (à tout temps).

Afin de déterminer l'effet cumulatif de l'exposition sur le risque d'un événement (maladie ou décès), nous avons besoin d'un seul proxy de l'exposition. Dans le but d'inclure toute l'information disponible dans un proxy nous pouvons utiliser la notion de dose cumulative (non pondérée) (Stranges, 2006) :

$$\sum_t X(t), \tag{9}$$

où $X(t)$ représente la valeur de la variable d'intérêt au moment t. Le problème avec ce genre de modèle est que la dose cumulative (non pondérée) ne prend pas en considération le moment de l'exposition. Ce modèle peut être amélioré pour prendre cet aspect en considération, en utilisant une combinaison linéaire :

$$\sum_t w(t) \times X(t), \tag{10}$$

où $w(t)$ est une fonction des poids. Si de l'information est disponible sur la forme de la fonction des poids, celle-ci peut être spécifiée *a priori*. Sinon elle peut être estimée en utilisant, par exemple, les B-splines. À la section 3.7, nous allons voir plus en détails l'utilisation des splines pour estimer la fonction des poids dans le cadre du modèle de Cox, modèle présenté par Sylvestre et Abrahamowicz (2009).

Chapitre 3

LES SPLINES

3.1. Introduction

L'idée de spline d'interpolation a ses origines dans l'industrie aéronautique et navale en utilisant le traçage conique. Le traçage conique a été remplacé par les splines au début des années 1960 avec le travail de Ferguson (1964) pour le Boeing et celui de M.A. Sabin pour British Aircraft Corporation.

Les splines ont été aussi utilisées dans l'industrie automobile autour de 1960 par plusieurs auteurs : Casteljau pour Citroën, Bézier pour Renault, Birkhoff, Garabedian et de Boor pour General Motors. Le travail du dernier s'est concrétisé en plusieurs articles incluant du travail fondamental sur les B-splines.

La première mention mathématique des splines est attribuée à Schoenberg (1946), qui est aussi la première personne à avoir utilisé le mot « spline » pour désigner l'approximation d'une fonction par des polynômes définis par morceaux.

Les fonctions splines sont des polynômes de degré p continues par morceaux. Le but des fonctions splines est de remplacer une seule fonction f, définie sur $[L, U]$, par plusieurs polynômes (splines) d'ordre inférieur définis sur des sous-intervalles de [L,U].

Définissons la partition suivante de [L,U] en k sous-intervalles :

$$I_i = [\tau_{i-1}, \tau_i) \ , \ i = 1, ..., k-1 \text{ et } I_k = [\tau_{k-1}, \tau_k] \,, \tag{11}$$

où $\tau = (\tau_0, ...\tau_k)^T$ est une séquence de points tels que : $L = \tau_0 < \tau_1 < ... < \tau_k = U$.

Définition 3.1.1. *Une courbe $s(t)$ est appelée « spline de degré p » avec les nœuds $\tau_0, ...\tau_k$, où $\tau_i \leq \tau_{i+1}$ et $\tau_i < \tau_{i+p+1}$ pour tout* i *si :*

- $s(t)$ *est un polynôme de degré* $\leq$ p *sur tout sous-intervalle* $[\tau_i, \tau_{i+1}]$;
- $s(t)$ *est* $p-\nu_i$ *fois dérivable* (ν_i *étant le nombre des fois qu'un nœud apparaît) pour* $i = 1, ..., k-1$.

Il faut mentionner que c'est usuel de parler d'une spline de degré p comme étant une spline d'ordre $p+1$.

Dans les problèmes d'interpolation, les splines sont souvent préférées à l'interpolation polynomiale car les résultats sont similaires et nous évitons les problèmes d'oscillation aux limites d'un intervalle qui sont présentes lors d'une interpolation polynomiale.

Parmi les premiers à utiliser les splines dans le cadre de l'analyse de survie est Hauptmann *et al.*(2000). Ils ont utilisé les B-splines dans le cadre d'un modèle généralisé pour estimer la fonction de poids. Plus récemment, Sylvestre et Abrahamowicz (2009) ont considéré une méthode flexible pour la modélisation des effets cumulatifs des expositions variables dans le temps. Ils ont utilisé la régression spline cubique pour estimer la fonction qui assigne les poids à des doses prises dans le passé.

Dans les sections suivantes, les B-splines seront introduites de façon plus détaillée. Une attention particulière sera aussi accordée aux splines cubiques et aux splines de régression. Finalement, les M-splines et les splines monotones (I-splines) seront présentées.

3.2. Les splines cubiques

Les splines cubiques sont utilisées en raison de leur grande flexibilité. De plus, comme toutes les splines, elles sont lisses et incluent moins de paramètres que les splines d'ordres supérieurs.

En définissant des polynômes sur chaque sous-intervalle I_i (tel que défini à l'équation (11), nous obtenons l'espace des polynômes par morceaux aux points τ_i pour $i = 0, ..., k$:

$$P_3(\tau) = \{f : f(x) = p_i(x)\mathbf{1}_i(x), p_i(x) \in P_3, i = 1, ..., k\},$$

où $\mathbf{1}_i$ est 1 si nous sommes dans l'intervalle I_i et 0 ailleurs. L'espace ainsi obtenu est beaucoup plus flexible, mais pas nécessairement lisse. Pour assurer la

continuité, nous imposons que les dérivées d'ordre un et deux soient continues :

$$S_3(\tau) = P_3(\tau) \cap C^2[L, U],$$

où $C^2[L, U] = \{f : f^{(i)}$ est continue sur $[L, U], i = 1, 2\}$. L'espace $S_3(\tau)$ est appelé espace des fonctions splines d'ordre trois avec k nœuds de multiplicité un. La notion de « multiplicité » fait référence au nombre de fois qu'un nœud apparaît dans l'ensemble.

Nous disposons d'un jeu des données de la forme :

$$(x_i, y_i),\ i = 1, ..., n,$$

où n représente le nombre total de couples dont nous disposons. Donc nous voulons ajuster une courbe sur les points (x_i, y_i), où $x_1 \leq x_2 \leq ... \leq x_n$. Nous allons considérer un polynôme cubique de la forme :

$$a \times x^3 + b \times x^2 + c \times x + d$$

entre chaque paire de points, représentant les nœuds, consécutifs de sorte que :

(1) chaque polynôme passe par les points extrêmes de l'intervalle sur lequel le polynôme est défini ;

(2) au point de rencontre de 2 polynômes, leurs dérivées premières et deuxièmes sont égales.

Notons que la spline cubique est appelé « spline de degré 3 » ou « spline d'ordre 4 » .

3.3. Les splines de régression

Les splines de régression sont des splines calculées selon un modèle de régression. En effet, nous considérons le modèle suivant :

$$y_i = f(x_i) + \varepsilon_i,\ i = 1, \ldots, n,$$

où f est une fonction que nous voulons estimer et ε représentent les termes d'erreur. Nous supposons que les erreurs sont indépendantes et identiquement distribuées selon une loi normale de moyenne 0 et variance σ^2, noté $\varepsilon_i \sim N(0, \sigma^2)$. En utilisant le théorème de Taylor, Eubank (1988) a réécrit ce modèle comme :

$$y_i = \sum_{j=0}^{3} \beta_j {x_i}^j + R(x_i) + \varepsilon_i,$$

où

$$R(x) = [(m-1)!]^{-1} \times \int_a^b f^{(m)}(x) \times (x-z)_+^{m-1} \tag{12}$$

et $u_+ = max(u, 0)$. L'expression (12) peut être approximée par la somme

$$\sum_{j=1}^{k} \theta_j \times (x - \tau_j)_+^3$$

pour un ensemble de points τ_j qui représente les nœuds et des coefficients θ_j, ce qui nous permet d'estimer la fonction f par

$$\sum_{j=0}^{3} \beta_j x^j + \sum_{j=1}^{k} \theta_j \times (x - \tau_j)_+^3.$$

Les coefficients $\{\beta_j\}_{j=0}^3$ et $\{\theta_j\}_{j=1}^k$ peuvent être estimés par la méthode des moindres carrés. Généralement, si nous avons k nœuds nous aurons $(p+k+1)$ coefficients de régression pour une spline de degré p.

Exemple 3.3.1. *Considérons que dans le cadre du même exemple de la section 2.2, nous disposons de la variable « âge », qui a été générée selon :*

$$\exp(SBP0/50) + 45 + erreur,$$

où l'erreur suit une distribution normale de moyenne zéro et d'écart type égal à deux. Nous voulons estimer SBP à l'aide de splines. Comme nous pouvons le constater en regardant la figure 3.1, l'approximation est très bonne.

3.4. Les B-splines

Dans cette section nous allons d'abord définir les B-splines– qui sont des splines de régression– et nous allons ensuite présenter quelques-unes de leurs propriétés. À travers cette section nous allons utiliser la même notation p pour indiquer soit le degré, soit l'ordre de la spline.

Le terme « B-spline » signifie simplement « spline de base » . En effet, Curry et Schoenberg (1966) ont montré que les B-splines sont linéairement indépendantes, donc elles forment une base.

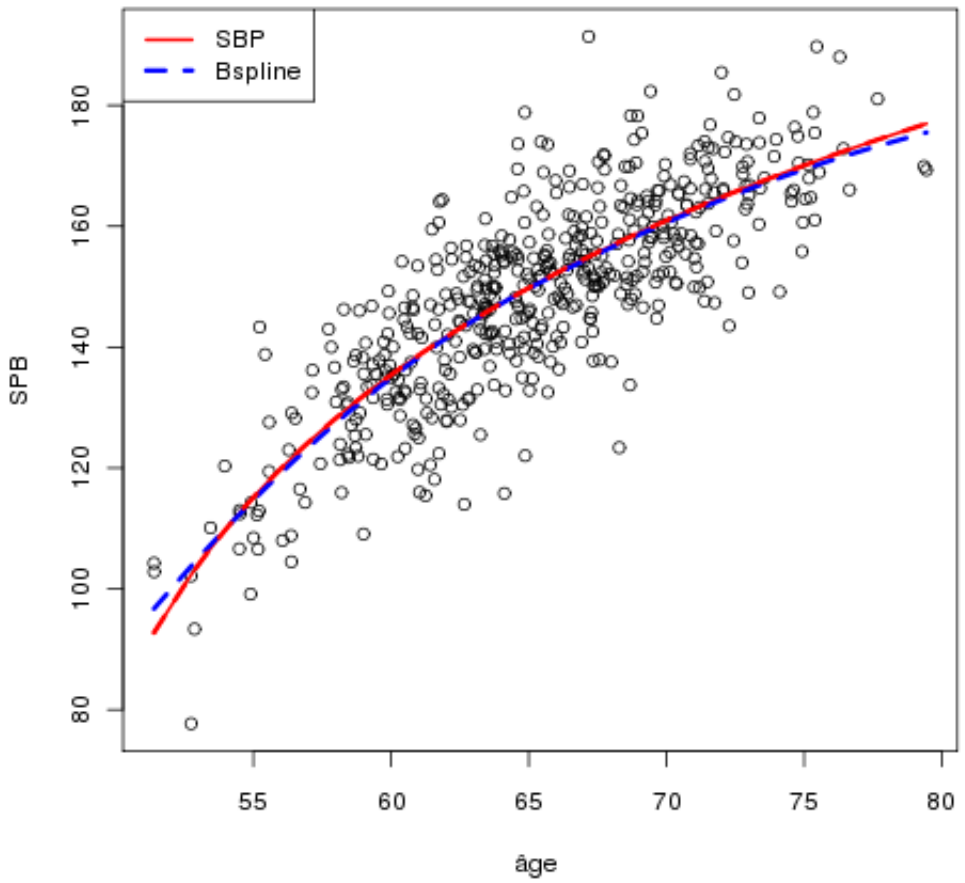


Figure 3.1: Approximation d'une courbe par des splines quadratiques

Théorème 3.4.1. *Pour une suite strictement croissante* $\underset{\sim}{C} = (c_1, ...c_{l+1})'$ *et une suite entière, non négative* $\underset{\sim}{\nu}=(\nu_2, .., \nu_l)'$ *avec* $\nu_i \leq p$ *pour tout* i *nous posons*

$$k = p + \sum_{i=2}^{l} (p - \nu_i).$$

Soit la suite non décroissante t *telle que :*

- $t_1 \leq t_2 \leq ... \leq t_p \leq c_1$ *et* $c_{l+1} \leq t_{k+1} \leq ... \leq t_{k+p}$ *;*
- *pour* $i = 2, ...l$, c_i *apparaît* $p - \nu_i$ *fois.*

Alors la suite $b_1, ..., b_k$ *de B-splines de degré* $p - 1$ *pour la suite de nœuds* t *est une base pour* $P_{p,c,\nu}$ *(l'espace linéaire des polynômes d'ordre* p*, avec les points de cassures* C *qui satisfont aux critères de continuité spécifiés par* ν *).*

Le théorème 3.4.1 nous permet aussi de définir les nœuds à partir d'une suite strictement croissante de points de cassures $\underset{\sim}{C} = (c_1, ..., c_{l+1})'$ et une suite entière, non négative $\underset{\sim}{\nu} = (\nu_2, ..\nu_l)'$ avec $\nu_i \leq p$ pour tout i. Ainsi, le vecteur des nœuds $\underset{\sim}{t}$ est tel que nous avons p nœuds inférieurs à c_1, p nœuds supérieurs à c_{l+1} et les nœuds intérieurs $c_2, c_3, \ldots, c_l$ apparaîssent $p-\nu_i$ fois. Nous allons voir à la section 3.6 comment choisir la suite $\underset{\sim}{\nu}$.

Dans le but de définir formellement le concept de B-spline, il faut d'abord clarifier la notion de « différence divisée » .

Définition 3.4.1. (de Boor, 1978)
La p^e différence divisée d'une fonction g aux points $c_i, \ldots, c_{i+p}$ est le coefficient dominant (coefficient de x^p) du polynôme de degré p qui coïncide avec g aux points $c_i, \ldots, c_{i+p}$. Nous notons $[c_i, \ldots, c_{i+p}]g$.

Il est important de noter que $[t_i, ..., t_{i+p}]g = 0$ si g est un polynôme d'ordre inférieur ou égal à p. De même, remarquons la possibilité d'écriture récursive des différences divisées :

$$[t_i, ...t_{i+p}]g = \begin{cases} \dfrac{[t_{i+1}, ...t_{i+p}]g - [t_i, ...t_{i+p-1}]g}{t_{i+p} - t_i} & \text{si } \ t_i \neq t_{i+p}, \\ \dfrac{g^{(p)}(t_i)}{p!} & \text{si } t_i = ... = t_{i+p}, \end{cases}$$

où $[t_i]g = g(t_i)$.

Nous sommes maintenant capable de définir les B-splines d'ordre p par la p^e différence divisée de fonctions de puissance tronquées.

Définition 3.4.2. (Curry et Schoenberg, 1966)
Soit t, un vecteur de nœuds. La i^e B-spline d'ordre p est définie par

$$b_{i,p}(x) = (t_{i+p} - t_i)[t_i, ..., t_{i+p}](\cdot - x)_+^{p-1}.$$

La notation $(\cdot - x)_+^{p-1}$ est utilisée pour accentuer le fait que la différence divisée de la fonction $(t-x)_+^{p-1}$ sera calculée comme fonction de t, en fixant x. Il faut aussi mentionner que le terme $(t_{i+p} - t_i)$, permet d'obtenir une base de fonctions normalisée. Dans la litérature, les B-splines définies comme ci-dessus sont aussi notées $N_{i,p}(x)$.

La définition donnée dans Curry et Schoenberg (1966) a été utilisé par de Boor et Cox (1972) pour développer une formule récurrente pour définir les B-splines, beaucoup plus faciles à implémenter numériquement.

Définition 3.4.3. (de Boor et Cox, 1972)
Soit la suite strictement croissante de nœuds t_i. Pour définir les B-splines nous utilisons la formule récursive suivante :

$$b_{i,0}(x) = \begin{cases} 1 & si \;\; x \in [t_i, t_{i+1}) \,, \\ 0 & sinon \,, \end{cases}$$

et

$$b_{i,p}(x) = \frac{x - t_i}{t_{i+p} - t_i} \times b_{i,p-1}(x) + \frac{t_{i+p+1} - x}{t_{i+p+1} - t_{i+1}} \times b_{i+1,p-1}(x).$$

Par convention, un quotient $\frac{0}{0}$ est défini comme 0.

Exemple 3.4.1. *Soit le vecteur des noeuds $\underset{\sim}{t} = (0, 0, 0, 0, 1, 3, 5, 6, 6, 6, 6)'$. Nous désirons évaluer la quatrième B-spline de degré 3 de la base, $b_{4,3}(x)$ au point x=4. En utilisant la définition 3.4.3 voici les calculs :*

$$\begin{aligned}
b_{4,3}(x) &= \frac{x - t_4}{t_7 - t_4} \times b_{4,2}(x) + \frac{t_8 - x}{t_8 - t_5} \times b_{5,2}(x) \\
&= \frac{x - t_4}{t_7 - t_4} \times \left[\frac{x - t_4}{t_6 - t_4} \times b_{4,1}(x) + \frac{t_7 - x}{t_7 - t_5} \times b_{5,1}(x) \right] \\
&+ \frac{t_8 - x}{t_8 - t_5} \times \left[\frac{x - t_5}{t_7 - t_5} \times b_{5,1}(x) + \frac{t_8 - x}{t_8 - t_6} \times b_{6,1}(x) \right] \\
&= \frac{(x - t_4)^2}{(t_7 - t_4)(t_6 - t_4)} \times \left[\frac{x - t_4}{t_5 - t_4} \times b_{4,0}(x) + \frac{t_6 - x}{t_6 - t_5} \times b_{5,0}(x) \right] \\
&+ \left[\frac{(x - t_4)(t_7 - x)}{(t_7 - t_4)(t_7 - t_5)} + \frac{(t_8 - x)(x - t_5)}{(t_8 - t_5)(t_7 - t_5)} \right] \times \left[\frac{x - t_5}{t_6 - t_5} \times b_{5,0}(x) + \frac{t_7 - x}{t_7 - t_6} \times b_{6,0}(x) \right] \\
&+ \frac{(t_8 - x)^2}{(t_8 - t_5)(t_8 - t_6)} \times \left[\frac{x - t_6}{t_7 - t_6} \times b_{6,0}(x) + \frac{t_8 - x}{t_8 - t_7} \times b_{7,0}(x) \right].
\end{aligned}$$

Nous voulons évaluer cette expression pour $x = 4$. *Comme* $x = 4 \in [t_6, t_7)$, *seulement* $b_{6,0}(x)$ *sera non nulle et donc l'expression devient :*

$$\begin{aligned} b_{4,3}(x) &= \left[\frac{(4-t_4)(t_7-4)}{(t_7-t_4)(t_7-t_5)} + \frac{(t_8-4)(4-t_5)}{(t_8-t_5)(t_7-t_5)}\right] \times \left(\frac{t_7-4}{t_7-t_6}\right) \times b_{6,0}(4) \\ &\quad + \frac{(t_8-4)^2}{(t_8-t_5)(t_8-t_6)} \times \left(\frac{4-t_6}{t_7-t_6}\right) \times b_{6,0}(4) \\ &= \frac{(4-0)(5-4)^2}{(5-0)(5-1)(5-3)} + \frac{(6-4)(4-1)(5-4)}{(6-1)(5-1)(5-3)} + \frac{(6-4)^2(4-3)}{(6-1)(6-3)(5-3)} \\ &= 0{,}3833. \end{aligned}$$

La définition récurrente met en évidence quelques propriétés des B-splines :

(1) $b_{i,p}(x)$ est un polynôme de degré p par morceaux et donc toutes les propriétés des polynômes notamment leur caractère relativement lisse et la facilité des manipulations algébriques se retrouvent chez les splines ;

(2) $b_{i,p}(x)$ est positive si $t_i < x < t_{i+p+1}$; à partir de cette propriété et de la façon dont nous avons défini les B-splines de degré p comme des fonctions normalisées, nous pouvons conclure qu'elles forment une partition de l'unité (Marsden et Schoenberg, 1966) :

$$\sum_i b_{i,p}(x) = \sum_{i=j+1-p}^{j} b_{i,p}(x) = 1;$$

(3) $b_{i,p}(x)$ est 0 si $x \notin [t_i, t_{i+p+1}]$, ce qui signifie que la fonction $b_{i,p}$ a un support compact ; de plus Curry et Schoenberg (1966) ont démontré qu'au plus $(p+1)$ B-splines peuvent être différentes de 0 sur un intervalle $[t_j, t_{j+1}]$, notamment $b_{j-p,p}(x), ..., b_{p,p}(x)$; l'intérêt du support compact est lié à la flexibilité de l'estimateur spline ;

(4) si $p > 0$, une spline de degré p, $b_{i,p}(x)$, est une combinaison linéaire des deux fonctions de base de degré $p-1$;

(5) $b_{i,p}(x)$ est continue à droite.

De même, il est facile de voir que $b_{i,0}(x)$ est une fonction constante égale à 0 partout sauf sur l'intervalle $[t_i, t_{i+1})$. Une autre propriété des B-splines découle de la définition des splines. Aux nœuds $b_{i,p}(x)$ est $p - \nu$ fois continûment dérivable, où ν représente la multiplicité d'un nœud.

La dérivation et l'intégration ne posent pas de problèmes, les B-splines étant des polynômes. Nous présentons ici la formule démontrée par de Boor (1972) pour le calcul de la dérivée de la i^e B-spline :

$$\frac{d}{dx}b_{i,p}(x) = \frac{p}{t_{i+p} - t_i} \times b_{i,p-1}(x) - \frac{p}{t_{i+p+1} - t_{i+1}} \times b_{i+1,p-1}(x).$$

Il faut souligner que la dérivée est une combinaison linéaire de B-splines, donc elle est aussi une spline. En conséquence, pour calculer un ensemble de fonctions de base il faut spécifier le degré p et un vecteur des nœuds $\underset{\sim}{t}$.

Il est à noter que toute fonction spline peut être écrite comme une combinaison linéaire de B-splines.

Définition 3.4.4. *Une fonction spline d'ordre p avec la séquence de nœuds t est toute combinaison linéaire de B-splines d'ordre p pour les mêmes nœuds. L'ensemble de toutes ces fonctions est noté $E_{p,t}$ (espace linéaire des splines d'ordre p, avec le vecteur des nœuds t)*

Dans la section suivante, nous allons définir la base des splines monotones.

3.5. M-splines et splines intégrés

La famille des M-splines est une base spline (pour l'espace linéaire des polynômes continus par morceaux) qui est très intéressante pour les statisticiens. La M-spline de degré $p, M_{i,p}$, pour $i = 1, \ldots, n$, est définie de telle façon qu'elle soit positive sur l'intervalle (t_i, t_{i+p+1}) et 0 ailleurs et qu'elle satisfasse la normalisation

$$\int_{t_i}^{t_{i+p+1}} M_{i,p}(x)dx = 1$$

(Curry et Schoenberg,1966). Comme les B-splines, les M-splines peuvent être définies en utilisant la notion de différence divisée, mais du point de vue computationnel il est préférable d'utiliser une formule récursive :

$$M_{i,p}(x) = \begin{cases} \dfrac{1}{t_{i+1} - t_i}, & \text{si } \ p = 0, \\ \dfrac{(p+1) \times [(x - t_i) \times M_{i,p}(x) + (t_{i+p+1} - x) \times M_{i+p+1,p}(x)]}{p \times (t_{i+p+1} - t_i)}, & \text{sinon,} \end{cases}$$

pour tout $x \in [t_i, t_{i+p+1}]$.

Il est à noter que chaque $M_{i,p}$ est une spline et par conséquent conserve toutes les propriétés énumérées à la section 3.4. De plus nous remarquons que cette base est fortement liée à la base des B-splines par la relation :

$$B_{i,p} = \frac{(t_{i+p+1} - t_i) \times M_{i,p}}{(p+1)}$$

qui utilise la normalisation $\sum_i B_{i,p}(x) = 1$ pour tout x.

Les splines monotones sont dérivées des précédentes. Puisque les M-splines sont non négatives, Ramsay (1988) a proposé de les intégrer pour obtenir des splines monotones. Les splines intégrées (I-splines) sont donc définies par $I_{i,p}(x) = \int_L^x M_{i,p-1}(x)$. Une formule beaucoup plus facile à implémenter est la suivante :

$$I_{i,p}(x) = \begin{cases} 0, & x < t_i, \\ \dfrac{\sum_{m=i}^{j}(t_{m+p+1} - t_m) \times M_{m,p}(x)}{p+1}, & x \in [t_j, t_{j+1}], \\ 1, & x > t_{i+p+1} \end{cases}$$

pour $j \in [i, i+p-1]$.

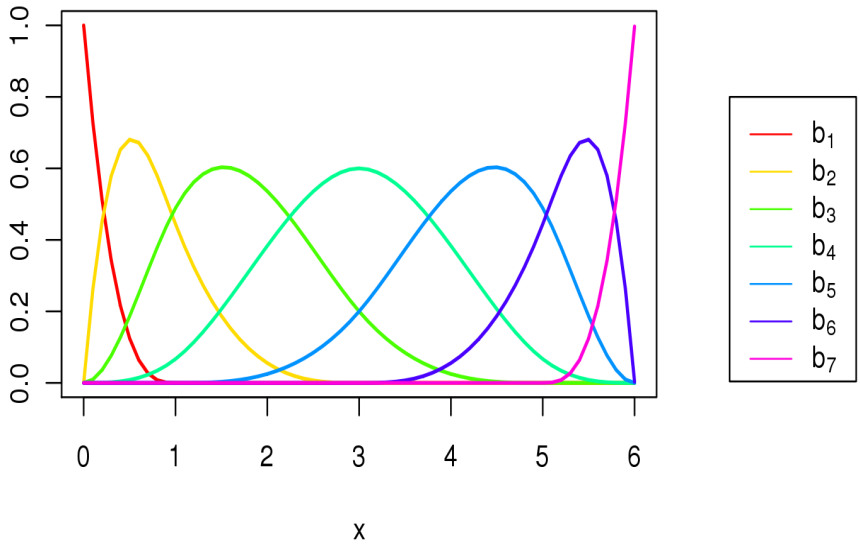
La figure 3.5 nous permet de visualiser les bases de fonctions B-splines et celle de I-splines cubiques. Les deux sont définies sur l'intervalle $[0, 6]$ avec trois nœuds intérieurs placés à un, trois et cinq respectivement.

Il faut mentionner que le nombre de splines avec lesquelles nous travaillons et le positionnement des nœuds diffèrent selon le type de modélisation. Dans la section suivante, nous présentons quelques critères pour choisir le nombre et l'emplacement des nœuds.

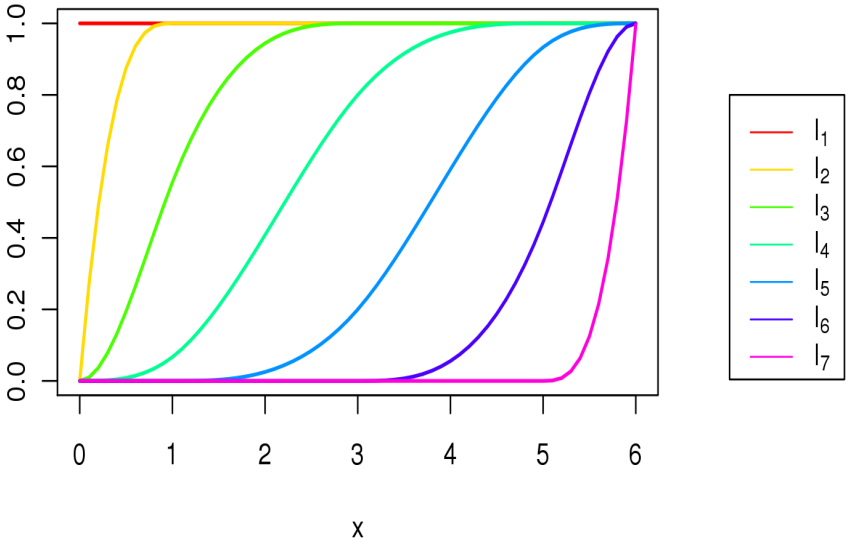
3.6. Le nombre et l'emplacement des nœuds

Le théorème 3.4.1 propose une façon de choisir la séquence de nœuds. Le niveau désiré de lissage à un point de cassure implique un nombre correspondant de nœuds à ce point, selon le principe que moins il y a de nœuds, plus il y a des conditions de continuité telles que :

multiplicité+flexibilité=ordre,



(a)



(b)

Figure 3.2: Exemple de fonctions splines (a) :B-splines, (b) : I-splines.

c'est-à-dire que le nombre de conditions de continuité à c_i (flexibilité) plus le nombre de nœuds à c_i (multiplicité) doit être égal à l'ordre p. Donc il y aura $\sum_{i=2}^{l}(p-\nu_i)$ nœuds intérieurs :

$$c_2 \leq t_{p+1} \leq ... \leq t_k \leq c_l$$

(où k est défini tel que dans le théorème 3.4.1) et en plus p nœuds initiaux et p finaux. Le théorème ne spécifie pas les modalités de sélection des nœuds initiaux et finals, mais un choix possible serait de prendre c_1 et c_{l+1}, ce qui est équivalent à prendre $\nu_1 = \nu_{l+1} = 0$, donc sans conditions de continuité aux extrêmes, ce qui est en accord avec le fait que les B-splines donnent une bonne description des données seulement sur l'intervalle $[t_p, t_{k+1}]$. Le tableau 3.1 présente une façon de convertir les points de cassure et les conditions de continuité en vecteur de nœuds, en prenant comme exemple la spline cubique (ordre $p = 4$) et où k est tel que spécifié par le théorème 3.4.1 :

Tableau 3.1: Combinaisons de points de cassures et conditions de continuité pour l'obtention des nœuds

Points de cassures	c_1	$c_2, ..., c_l$	c_{l+1}
Nombre de conditions de continuité	ν_1	$\nu_2, ..., \nu_l$	ν_{l+1}
Multiplicité des nœuds	4	$4-\nu_i, i = 2, \ldots l$	4
nœuds	$t_1, ..., t_4$	$t_5, ..., t_k$	$t_{k+1}, ..., t_{k+4}$

En ce qui concerne le nombre de nœuds, plusieurs auteurs (Stone, 1986 ; Durrleman et Simon, 1989 ; Hess, 1994) recommandent un petit nombre de nœuds (trois à cinq) en considérant que c'est suffisant pour bien décrire les données. Pour un échantillon assez grand nous pouvons choisir un nombre de nœuds plus grand si nous pensons qu'il y a des changements brusques dans la courbe, mais il est à noter que la variance de l'estimateur de la spline et le risque de surparamétrisation augmentent aussi.

Il faut aussi se rappeler que le choix des nœuds doit prendre en considération deux éléments. Premièrement,un plus grand nombre de nœuds dans une région augmente la flexibilité de la courbe dans cette région. Deuxièmement, lorsqu'il y

a un plus grand nombre des points entre deux nœuds consécutifs, la courbe est mieux définie (Ramsay, 1988).

Au sujet de l'emplacement des nœuds intérieurs, les observations empiriques indiquent que les résultats sont insensibles à l'emplacement des nœuds sauf s'ils sont placés d'une façon non uniforme. Nous présentons ici quelques approches utilisées par divers auteurs :

- une première approche consiste à placer les nœuds aux $(k+1)^e$ quantiles de la distribution de la variable X ; de cette façon nous assurons un nombre égal de données pour l'estimation entre les nœuds. Il est à noter qu'un minimum de quatre ou cinq observations entre les nœuds sont nécessaires (Abrahamowicz 1992) ;
- He and Shi (1998) ont proposé une autre approche, notamment de commencer avec un ensemble de nœuds équidistants placés tel que $t_j = x_{[\frac{jn}{k}]}$, le $(\frac{j}{k})^e$ quantile de X et enlever des nœuds jusqu'à l'obtention d'un emplacement optimal (flexibilité et nombre de nœuds minimal) selon le critère :

$$IC(k) = log\left(\sum_{i=1}^{n} |y_i - \hat{g}(x_i)|\right) + 2 \times (k+2)/n,$$

  où n est le nombre d'observations et k est le nombre de paramètres estimés, tel que défini par le théorème 3.4.1 ;
- une autre possibilité (Stone, 1986) serait de placer des nœuds selon la règle suivante : trois nœuds placés aux $5^e, 50^e$ et 95^e percentiles et deux nœuds supplémentaires aux percentiles $\frac{1}{1+\sqrt{n}}$ et $\frac{1}{1+\frac{1}{\sqrt{n}}}$ (où n est la taille de l'échantillon) ;
- pour $k = 3$ nœuds, Durrleman et Simon (1989) suggèrent de fixer t_1 au 5^e percentile et t_3 au 95^e percentile et d'essayer pour t_2 tous les percentiles et choisir le nœud optimal en utilisant la régression par étapes.

En conclusion, nous désirons des nœuds qui sont :

- près, mais pas dans les extrêmes ; si nous les plaçons dans les extrêmes les valeurs aberrantes peuvent avoir une influence excessive sur la spline ;
- à peu près uniformes sur les quantiles ;
- en nombre limité.

Comme mentionné à la section 2.7, les splines peuvent être utilisées pour estimer une fonction de poids, afin d'inclure le moment de l'exposition dans la valeur de l'exposition cumulative. Dans la section suivante, nous présentons le modèle de Cox avec exposition cumulative pondérée par le passé récent où la fonction de poids est estimée en utilisant les B-splines cubiques.

3.7. Exposition cumulative pondérée

Dans le cadre de ce modèle, les B-splines cubiques sont utilisées principalement en raison de leur caractère relativement lisse et de leur facilité de manipulations algébrique. De plus, elles sont assez flexibles pour pouvoir représenter une grande diversité de formes.

Comme mentionné précédemment, une fonction de poids peut être utilisée afin de tenir compte du moment de l'exposition :

$$\sum_{t}^{u} w(u-t) \times X(t),$$

où t représente des temps d'exposition qui précèdent u et $w(u-t)$ est la fonction de poids pour l'exposition passée selon le passé récent. Cette fonction est par la suite estimée en utilisant les B-splines cubiques :

$$w(u-t) = \sum_{j=1}^{k} \theta_j \times b_j(u-t), \tag{13}$$

où $b_j, j = 1, \ldots, k$ sont les k fonctions de la base spline et $\theta_j, j = 1, \ldots, k$ sont les coeficients de la combinaison linéaire des splines.

Le modèle de Cox incluant la fonction de poids s'écrit comme :

$$h(u|\underset{\sim}{X}(u), \underset{\sim}{Z}(u)) = h_0(u) \times \exp\left[\beta_1 \sum_{t}^{u} w(u-t) \times X(t) + \sum_{s=2}^{q} \beta_s Z_s(u)\right],$$

où $\underset{\sim}{X}(u)$ est le vecteur des expositions passées pour la variable dépendante du temps et $Z_s(u), s = 2, \ldots, q$ sont les valeurs des autre variables indépendantes du

temps u. En utilisant l'équation (13) nous pouvons réécrire le modèle de Cox :

$$h(u|\underset{\sim}{X}(u), Z(u)) = h_{\text{o}}(u) \times \exp\Big[\beta_1 \sum_{t}^{u} \sum_{j=1}^{k} \theta_j b_j(u-t) \times X(t) + \sum_{s=2}^{q} \beta_s Z_s(u)\Big]$$
$$= h_{\text{o}}(u) \times \exp\Big[\sum_{j=1}^{k} \beta_1 \theta_j \Big[\sum_{t}^{u} b_j(u-t) \times X(t)\Big] + \sum_{s=2}^{q} \beta_s Z_s(u)\Big].$$

Nous posons $\gamma_j = \beta_1 \theta_j,$ pour $j = 1, \ldots, k$ et nous remplaçons le terme

$$\sum_{t}^{u} b_j(u-t) \times X(t)$$

par des covariables dépendantes du temps $T_j(u),$ pour $j = 1, \ldots, k$.

Le modèle de hasards proportionnels devient alors :

$$h(u|\underset{\sim}{X}(u), \underset{\sim}{Z}(u)) = h_0(u) \times \exp\Big[\sum_{j=1}^{k} \gamma_j T_j(u) + \sum_{s=2}^{q} \beta_s Z_s(u)\Big].$$

L'étape suivante consiste à calculer les covariables dépendantes du temps $T_j(u),$ pour $j = 1, \ldots, k$. Finalement le modèle de Cox est appliqué pour estimer les paramètres $\beta_s,$ pour $s = 2, \ldots, q$ et γ_j pour $j = 1, \ldots, k$ (voir l'équation (3)).

Même si β_1 ne peut pas être explicitement calculé, nous pouvons calculer le HR pour différents modèles d'exposition à partir des estimés obtenus dans le modèle plus haut selon l'équation (6). Ainsi nous pouvons comparer par exemple les individus avec vecteurs d'exposition passée X_1 et X_2 en gardant les autre covariables constantes, en utilisant :

$$HR = exp\Big[\sum_{j=1}^{k} \hat{\gamma}_i \sum_{t}^{u} b_j(u-t) \times \Big[X_1(t) - X_2(t)\Big]\Big]. \tag{14}$$

Il faut mentionner que dans le cadre du modèle une fenêtre d'exposition est utilisée (plusieurs peuvent être essayer afin de déterminer le meilleur modèle). Cette fenêtre représente la longueur de la période de l'effet résiduel après l'arrêt du médicament étudié. Une fois cette fenêtre déterminée, les nœuds sont établis aux intervalles équidistants dans la fenêtre d'exposition.

Exemple 3.7.1. *Revenons à l'exemple présenté précédemment. Nous pouvons penser que si une personne est exposée à un certain traitement, l'effet de cette exposition s'accumule dans le temps. Nous allons par la suite considérer l'effet de l'exposition cumulative.*

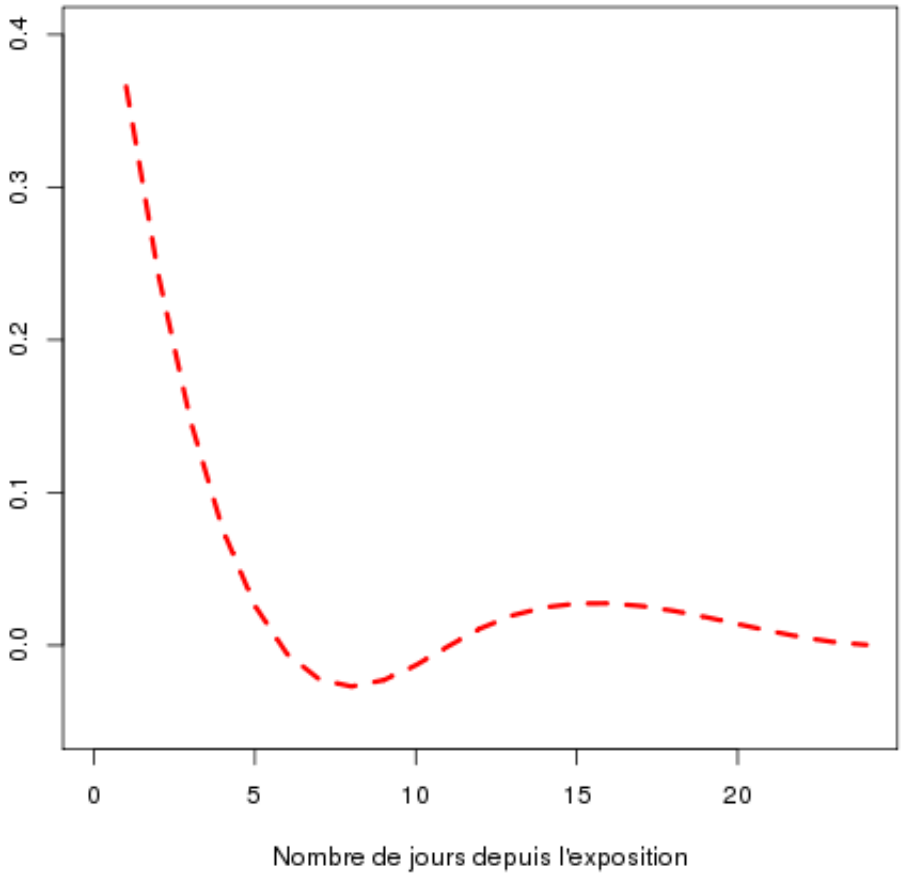


Figure 3.3: Approximation des poids pour le calcul de l'effet cumulatif du traitement

La figure 3.3 montre le résultat obtenu selon le modèle présenté ci-dessus.À la lecture de ce graphique, nous pouvons constater que l'effet cumulatif du traitement décroît rapidement pour atteindre zéro après 7-8 jours.

Comme mentionné ci-dessus, un avantage de cette méthode est qu'elle permet de calculer le rapport de risques pour différents modèles d'exposition. En utilisant l'équation (14)*, dans le tableau 3.2 nous présentons quelques exemples pour différents historiques du traitement :*

Tableau 3.2: Résultats pour différents historiques du traitement

Historique du traitement	*Référence*	*HR*
Traité pour les 24 derniers jours	*Jamais traité dans le 24 derniers jours*	*2,66*
Traité pour les 7 derniers jours	*Jamais traité dans le 24 derniers jours*	*2,29*
Traité 7 jours, 7 jours auparavant	*Jamais traité dans le 24 dernieres jours*	*1*

Nous pouvons dire qu'une personne ayant pris le traitement pendant les 24 derniers jours a 2,66 fois plus de chances d'avoir un événement cardiovasculaire, comparativement à une personne qui n'a pas pris le traitement pendant ce temps. Par contre, pour un sujet qui a été exposé sept jours auparavant pour une semaine, il n'y a plus d'effet du traitement.

Dans le chapitre cinq, ce modèle est appliqué pour estimer l'effet de l'adhésion à la thérapie avec antihypertenseurs sur l'incidence des accidents vasculaires cérébraux.

Chapitre 4

MOYENNAGE BAYÉSIEN DE MODÈLES

Dans ce chapitre nous allons d'abord présenter le modèle bayésien général. Dans la deuxième section, nous allons exposer les principes généraux du moyennage bayésien de modèles et continuer avec les détails de son implémentation (Volinski *et al.*,1997). Nous allons par la suite introduire l'interprétation des résultats du moyennage bayésien de modèles. À la fin du chapitre nous présentons l'approche bayésienne pour l'exemple introduit à la section 2.2. Dans le chapitre cinq nous allons reprendre le modèle dans un contexte nouveau, notamment avec une variable dépendante du temps.

4.1. Modèle bayésien général

Dans cette section nous décrivons un modèle bayésien et nous donnons quelques définitions.

La fonction de vraisemblance est la loi des observations conditionnée sur le paramètre $\underset{\sim}{\theta}$. La différence par rapport à l'approche classique réside dans le fait que le paramètre $\underset{\sim}{\theta}$ est considéré aléatoire. Nous faisons donc l'hypothèse que le paramètre $\underset{\sim}{\theta}$ suit une certaine loi de probabilité que nous appelons loi *a priori*.

Définition 4.1.1. *La loi a posteriori du paramètre $\underset{\sim}{\theta}$ est sa loi étant donnée les observations $\underset{\sim}{x}$:*

$$\pi(\underset{\sim}{\theta}|\underset{\sim}{x}) = \frac{f(\underset{\sim}{x}|\underset{\sim}{\theta}) \times \pi(\underset{\sim}{\theta})}{m(\underset{\sim}{x})},$$

où $f(\underset{\sim}{x}|\underset{\sim}{\theta})$ est la fonction de vraisemblance des observations sachant le paramètre $\underset{\sim}{\theta}$, $\pi(\underset{\sim}{\theta})$ est la loi a priori du paramètre $\underset{\sim}{\theta}$ et $m(\underset{\sim}{x})$ est la loi marginale des observations définie comme $m(\underset{\sim}{x}) = \int f(\underset{\sim}{x}|\underset{\sim}{\theta}) \times \pi(\underset{\sim}{\theta}) d\underset{\sim}{\theta}$.

La loi *a priori* est souvent choisie de façon à faciliter les calculs de la densité marginale des observations. Une telle loi fait partie d'une classe de densités dites conjuguées.

Définition 4.1.2. *Une famille F sur $\underset{\sim}{\theta}$ est dite conjuguée si pour toute loi a priori $\pi(\underset{\sim}{\theta}) \in F$, la loi a posteriori $\pi(\underset{\sim}{\theta}|\underset{\sim}{x})$ appartient également à F.*

En utilisant une loi *a priori* conjuguée, le calcul de $\pi(\underset{\sim}{\theta}|\underset{\sim}{x})$ revient à faire une mise-à-jour des paramètres. L'utilisation des densités conjuguées est justifiée par le fait que l'information sur $\underset{\sim}{\theta}$ contenue dans l'échantillon $\underset{\sim}{x}$ est une quantité finie et nous ne pouvons changer la forme de la densité *a priori* (Raiffa et Schlaifer, 1961).

La mesure de l'erreur lorsque nous estimons $\underset{\sim}{\theta}$ par $\underset{\sim}{\widehat{\theta}}$ est donnée par la fonction de perte. En estimation ponctuelle, il existe des fonctions de perte standards : quadratique, absolue.

Définition 4.1.3. *La fonction de perte quadratique $L(\underset{\sim}{\theta}, \underset{\sim}{\widehat{\theta}}) : R_p \times R_p \xrightarrow{L(.,.)} R_p^+$ est de la forme :*

$$L(\underset{\sim}{\theta}, \underset{\sim}{\widehat{\theta}}) = (\underset{\sim}{\theta} - \underset{\sim}{\widehat{\theta}})' \times \boldsymbol{Q} \times (\underset{\sim}{\theta} - \underset{\sim}{\widehat{\theta}}),$$

où $\boldsymbol{Q}$ est une matrice connue, symétrique et définie positive.

En général, la fonction de perte est supposée positive. Nous pouvons donc dire que la fonction de perte représente le coût (ou l'erreur) dû à une mauvaise évaluation de la fonction de θ et donc que même la meilleure évaluation de cette fonction peut donner au mieux un coût nul.

Il est généralement impossible de minimiser uniformément (en $\widehat{\theta}$) la fonction de perte. Pour obtenir un critère de comparaison utilisable, l'approche fréquentiste propose de considérer la perte moyenne (risque fréquentiste) :

$$R(\theta, \delta) = E_\theta[L(\theta, \delta(x))] = \int_X L(\theta, \delta(x)) \times f(x|\theta) dx,$$

où $\delta(x)$ est la règle de décision, c'est-à-dire l'attribution d'une décision à chaque résultat x de l'expérience aléatoire.

Comme θ est inconnu dans l'approche bayésienne, nous intégrons sur l'espace Θ. Nous obtenons alors la perte *a posteriori* :

$$\rho(\pi, \widehat{\theta}|x) = E^\pi[L(\theta, \widehat{\theta})|x] = \int_\Theta L(\theta, \widehat{\theta}) \times \pi(\theta|x) d\theta,$$

qui est la moyenne de la fonction de perte selon la distribution *a posteriori* de θ, conditionnellement à la valeur observée x.

En se donnant une distribution *a priori* π, il est possible de définir le risque intégré qui est le risque fréquentiste moyenné sur les valeurs de θ selon leur distribution *a priori* :

$$r(\pi, \delta) = E^{\pi}[R(\theta, \delta)] \tag{15}$$
$$= \int_{\Theta} \left(\int_{X} L(\theta, \delta(x)) \times f(x|\theta) dx \right) \times \pi(\theta) d\theta.$$

Le concept est important parce qu'il attribue un nombre réel à chaque estimateur, ce qui permet une comparaison directe entre ces estimateurs.

Théorème 4.1.1. *Un estimateur minimisant le risque intégré $r(\pi, \delta)$ est obtenu par sélection, pour chaque $x \in X$, de la valeur $\delta(x)$ qui minimise la perte* a posteriori, *$\rho(\pi, \widehat{\theta}|x)$, puisque :*

$$r(\pi, \delta) = \int_{X} \rho(\pi, \delta(x)|x) \times m(x) dx.$$

En utilisant ce résultat nous pouvons définir le concept d'estimateur de Bayes.

Définition 4.1.4. *Un estimateur de Bayes associé à une distribution* a priori π *et une fonction de perte L est un estimateur δ^{π} minimisant $r(\pi, \delta)$. Pour chaque $x \in X$, ce dernier est donné par*

$$\delta^{\pi}(x) = arg\ min_{\theta} \rho(\pi, \theta|x).$$

La valeur $r(\pi) = r(\pi, \delta^{\pi})$ est alors appelée risque de Bayes.

4.2. Moyennage bayésien des modèles (MBM)

Une des premières mentions de la comparaison de modèles sur un même ensemble de données est dans la littérature spécifique au contrôle de la qualité. En 1963, Barnard compare les prévisions du modèle proposé par Box-Jenkins aux modèles standards en utilisant des données sur des réservations pour des vols aériens. Bates et Granger (1969) ont démontré que la moyenne des deux prévisions est meilleure que chacune des deux prévisions prises individuellement. De plus ils ont obtenu des poids optimaux à cet effet.

Une des premières mentions de la notion de moyennage des modèles dans la litérature statistique appartient à Roberts (1965) qui a suggéré une distribution qui combine l'opinion de deux experts (modèles). Le paradigme du moyennage bayésien de modèles a été présenté par Leamer (1978). Il a developpé des expressions pour la valeur espérée et pour la variance de l'estimateur moyen. De plus, il a aussi mis en évidence le fait que le moyennage bayésien de modèles prend en considération l'incertitude dans le choix desdits modèles.

Newbold et Granger (1974) ont analysé plusieurs procédures servant à prédire une série temporelle à partir de ses valeurs passées et présentes. Ils ont également étudié des façons possibles de combiner des prévisions. Leurs résultats empiriques ont montré une amélioration de la prévision lors de l'utilisation de cette procédure.

Souvent, l'espace sous lequel nous voulons faire le moyennage peut être très grand. Si nous disposons des q variables explicatives, alors la dimension de cet espace sera d'au moins 2^q. Cet aspect explique la popularité des méthodes MCMC. Madigan et Xork (1995) ont introduit les « Markov chain Monte Carlo model composition (MC^3) ». Cette méthode a été utilisée par Raftery *et al.*(1997) afin de faire une sélection de modèle dans la régression linéaire hiérarchique. La même méthode a été utilisée dans le cadre du modèle linéaire généralisé par Raftery *et al.* (1996).

Même si beaucoup d'auteurs considèrent que l'incertitude dans le choix des modèles doit être prise en considération, il n'y a pas eu beaucoup de progrès en raison de la difficulté d'implémentation informatique du moyennage bayésien de modèles. Hoeting *et al.* (1999) ont développé des logiciels pour l'implémentation du moyennage bayésien de modèles dans le cadre de la régression linéaire, du modèle linéaire généralisé, de l'analyse de survie et pour les modèles graphiques.

Dans ce mémoire, le MBM est appliqué au modèle de Cox et cela en prenant en considération l'exposition qui varie dans le temps.

4.2.1. Principes généraux

Du point de vue conceptuel, le moyennage bayésien des modèles fait l'inférence par une moyenne pondérée de tout l'espace-modèle et de cette façon nous prenons en considération l'incertitude dans les prévisions et dans les estimations.

Supposons l'existence de q prédicteurs. Alors il y aura jusqu'à $K = 2^q$ modèles possibles (si nous supposons qu'il n'y a pas d'interactions entre les facteurs) définis tels que chaque prédicteur peut être inclus ou non dans le modèle. Nous notons ces modèles par $M_1, \ldots, M_K$. Le moyennage bayésien propage l'incertitude quant au choix du modèle vers l'inférence sur toute quantité d'intérêt en utilisant la loi des probabilités totales.

Notons Q une quantité d'intérêt qui a la même interprétation dans chaque modèle. Sa distribution *a posteriori* prenant en considération l'incertitude dans le choix du modèle est donnée par :

$$\mathrm{P}(Q|D) = \sum_{k=1}^{K} \mathrm{P}(Q|D, M_k) \times \mathrm{P}(M_k|D), \tag{16}$$

où D représente les données, $\mathrm{P}(Q|D, M_k)$ est la distribution *a posteriori* de Q pour le modèle M_k et $\mathrm{P}(M_k|D)$ est la probabilité *a posteriori* du modèle M_k sachant les données. Donc nous pouvons dire que la distribution *a posteriori* de Q est une moyenne pondérée de la distribution *a posteriori* spécifique aux différents modèles et les poids sont donnés par les probabilités *a posteriori* du chacun des modèles.

Les probabilités *a posteriori* pour chacun des modèles peuvent être calculées en utilisant la relation :

$$\mathrm{P}(M_k|D) \propto \mathrm{P}(D|M_k) \times \mathrm{P}(M_k), \tag{17}$$

où $\mathrm{P}(M_k)$ est la probabilité *a priori* du modèle M_k ; ces probabilités sont souvent choisies comme étant égales. La quantité $\mathrm{P}(D|M_k)$ est donnée par :

$$\mathrm{P}(D|M_k) = \int \mathrm{P}(D|\theta_k, M_k) \times \mathrm{P}(\theta_k|M_k) d\theta_k,$$

où $\underset{\sim}{\theta_k}$ est le vecteur des paramètres pour le modèle M_k, $\mathrm{P}(D|\theta_k, M_k)$ est sa vraisemblance et $\mathrm{P}(\theta_k|M_k)$ sa distribution *a priori* sous le modèle M_k.

La constante de proportionalité dans l'équation (17) est choisie de telle façon que les probabilités *a posteriori* des modèles somment à un (Lee, 1989).

La technique MBM donne en moyenne une meilleure prévision qu'un seul modèle. De plus l'inférence est bien calibrée, en ce sens que les intervalles de confiance, par exemple, ont en moyenne le bon taux de couverture. Le modèle utilisé dans le cadre de ce mémoire est le modèle qui a été proposé par Volinsky *et al.* (1997).

4.2.2. Implémentation du moyennage bayésien des modèles

Le cadre standard pour le modèle de moyennage bayésien est donné par l'équation (16).

Comme mentionné à la section 4.2.1, si nous supposons l'existence de q prédicteurs le nombre total de modèles (K dans l'équation (16)) sera très grand. Pour cette raison, Madigan et Raftery (1994) ont proposé de réduire l'espace de moyennage en prenant un sous-ensemble des meilleurs modèles selon leur probabilité *a posteriori*. Ils ont montré que si nous utilisons l'ensemble :

$$A = \left\{ M_k : \frac{\max_j \mathrm{P}(M_j|D)}{\mathrm{P}(M_k|D)} \leq C \right\},$$

avec $C = 20$ nous obtenons une bonne approximation de la moyenne sur tout l'espace.

L'équation (16) présente trois difficultés :

(1) La distribution prédictive de Q nécessite l'intégration par rapport au paramètre θ_k :

$$\mathrm{P}(Q|M_k, D) = \int \mathrm{P}(Q|\theta_k, M_k, D) \times \mathrm{P}(\theta_k|M_k, D) d\theta_k.$$

Cette intégrale n'a pas de forme analytique pour la majorité des modèles de survie. Taplin (1993) et Taplin et Raftery(1994) ont utilisé l'approximation par maximum de vraisemblance :

$$\mathrm{P}(Q|M_k, D) \approx \mathrm{P}(\theta|M_k, \widehat{\theta}^k, D),$$

et ils ont trouvé que c'était une très bonne approximation.

(2) La probabilité *a posteriori* du modèle M_k est donnée par :

$$\mathrm{P}(M_k|D) \propto \mathrm{P}(D|M_k) \times \mathrm{P}(M_k), \tag{18}$$

où

$$\mathrm{P}(D|M_k) = \int \mathrm{P}(D|\theta_k, M_k) \times \mathrm{P}(\theta_k|M_k) d\theta_k. \tag{19}$$

et $\mathrm{P}(\theta_k|M_k)$ est la densité *a priori* du paramètre θ_k sous le modèle M_k. Premièrement, l'évaluation de l'intégrale (19) a rarement une solution analytique ; l'approximation à l'aide de la méthode de Laplace peut être utilisée (Raftery, 1996) :

$$\log \mathrm{P}(D|M_k) = \log \mathrm{P}(D|\widehat{\theta}_k, M_k) - d_k \times \log n + O(1), \tag{20}$$

où n est le nombre de sujets et d_k est la dimension de θ_k. Dans le cadre de l'analyse de survie, n peut représenter soit le nombre total d'individus, soit le nombre d'individus non censurés. Pour l'analyse effectuée dans ce mémoire nous avons utilisé le nombre d'individus non censurés. Volinski (1997) a montré que cette expression représente une bonne approximation pour le BIC (Bayesian information criterion). Deuxièmement, dans l'équation (18), nous avons besoin de spécifier la probabilité *a priori* pour chacun des modèles M_i. S'il n'existe pas d'information supplémentaire alors nous pouvons supposer des probabilités *a priori* égales pour tous les modèles. Par contre, si nous disposons d'information supplémentaire la probabilité *a priori* pour le modèle M_i peut être écrite comme :

$$\mathrm{P}(M_i) = \prod_{j=1}^{p} \pi_j^{\delta_{ij}} \times (1 - \pi_j)^{1-\delta_{ij}}, \tag{21}$$

où $\pi_j \in [0; 1]$ est la probabilité *a priori* que $\theta_j \neq 0$ et δ_{ij} est un marqueur qui indique la présence de la variable j dans le modèle M_i.

(3) Les modèles qui se trouvent dans l'ensemble A doivent être identifiés et évalués de façon efficace.

Comme mentionné ci-dessus, pour cette approche nous avons besoin d'identifier les modèles qui ont une probabilité *a posteriori* au moins $\frac{1}{C}$ de la

probabilité *a posteriori* du meilleur modèle. Le meilleur modèle est considéré celui avec la plus grande probabilité *a posteriori*. Pour ne pas avoir besoin d'ajuster tous les modèles possibles, nous utilisons l'algorithme développé par Furnival et Wilson (1947) pour la régression linéaire. Cet algorithme est basé sur le fait que pour 2 modèles de régression, M_1 et M_2, si $M_1 \subset M_2$, alors la somme des carrés des résidus pour le modèle M_1 est plus grande que la somme des carrés des résidus pour le modèle M_2. Supposons que nous sommes intéressés à trouver le meilleur modèle à $p-i$ variables. Le point de départ de l'algorithme est le modèle complet qui contient p variables. À l'étape suivante, nous ajustons les p modèles avec $p-1$ variables. Parmi ces modèles nous regardons celui qui est le meileur (plus petite somme des carrés des résidus), ensuite nous continuons en regardant seulement les $p-2$ sous-modèles de celui-là. Nous arrêtons le processus lorsque nous avons trouvé le meilleur modèle à $p-i$ variables.

Toute l'information qui est contenue dans les données peut être exprimée sous la forme matricielle suivante :

$$\begin{pmatrix} X'X & X'y \\ y'X & y'y \end{pmatrix}, \tag{22}$$

Cette matrice est souvent utilisée comme point de départ pour l'algorithme.

Lawless et Singhal (1978) ont modifié l'algorithme présenté ci-dessus pour les modèles de régression non linéaires (qui donne aussi une approximation pour le test du rapport de vraisemblance). Kuk (1984) a appliqué cet algorithme au modèle de Cox. Si nous supposons que $\underset{\sim}{\theta}$ est le vecteur des paramètres du modèle complet et $\underset{\sim}{\theta_k}$ est le vecteur des paramètres du modèle k, nous pouvons réécrire $\underset{\sim}{\theta_k} = (\theta_1, \theta_2)$. Soit V l'inverse de la matrice d'information :

$$V = I^{-1} = \begin{pmatrix} V_{11} & V_{12} \\ V'_{12} & V_{22} \end{pmatrix}.$$

La statistique du rapport de vraisemblance pour le test du modèle le plus petit contre le modèle complet est donnée par :

$$\Lambda = -2 \times [\log L(\tilde{\theta}) - \log L(\widehat{\theta})],$$

où $L(\tilde{\theta})$ est la vraisemblance maximisée pour le modèle complet et $\log L(\widehat{\theta})$ est la vraisemblance maximisée si $\theta_2 = 0$. La statistique du rapport de vraisemblance, Λ, peut être approximée par :

$$\Lambda' = \widehat{\theta}_2' V_{22}^{-1} \widehat{\theta}_2.$$

Finalement la matrice (22) est remplacée par :

$$\begin{pmatrix} I & I\widehat{\theta} \\ \widehat{\theta}' I & \widehat{\theta}' I \widehat{\theta} \end{pmatrix},$$

et l'algorithme est appliqué à cette matrice.

Si le nombre de modèles est assez grand, cette technique retourne tous les modèles dans l'ensemble A de même que d'autres modèles qui ne sont pas nécessairement dans A. Afin de réduire le nombre de modèles, dans un premier temps, sont gardés les modèles dont la probabilité *a posteriori* est au moins $\frac{1}{C^2}$. Ensuite tous les modèles restants sont ajustés avec un programme standard pour l'analyse de survie, la valeur exacte du BIC est calculée et les modèles qui ne sont pas dans A sont éliminés. Finalement les probabilités *a posteriori* sont normalisées. Les probabilités *a posteriori* des modèles sont utilisées comme des poids dans le moyennage des modèles et donnent de plus une mesure de l'incertitude.

4.2.3. Interprétation des résultats

La probabilité *a posteriori* que le coefficient de régression pour une variable soit non nulle est calculée en additionnant les probabilités *a posteriori* de tous les modèles contenant la variable. Cette probabilité peut être une mesure de l'importance de la variable. Kass and Raftery (1995) précisent les règles générales pour interpréter les probabilités *a posteriori* :

– $< 50\%$ évidence contre l'effet ;

– [50%; 75%] évidence faible de l'effet ;

– [75%; 95%] évidence positive ;

– [95%; 99%] forte évidence ;

– > 99% très forte évidence.

La probabilité *a posteriori* du paramètre a une interprétation très importante et qui est différente de l'interprétation de la valeur p. L'approche bayésienne répond aux questions plus intéressantes du point de vue des chercheurs : « Quelle est la probabilité que le modèle soit vrai ? » et « Quelle est la probabilité que ce coefficient soit non nul ? ».

En utilisant (16) nous pouvons calculer la moyenne *a posteriori* des coefficients de régression :

$$\begin{aligned}\widehat{\theta}_{MBM} &= \mathrm{E}_M(\widehat{\theta}) = \sum_{i=1}^{K} \widehat{\theta}_i \, \mathrm{P}(M_i|D) \\ &= \frac{\sum_{i=1}^{K} \widehat{\theta}_i \, \mathrm{P}(M_i|D)}{\sum_{i:\theta_i \in M_i} \mathrm{P}(M_i|D)} \times \sum_{i:\theta_i \in M_i} \mathrm{P}(M_i|D) \\ &= \mathrm{E}(\widehat{\theta}|\theta_i \in M_i) \times \mathrm{P}(\theta \neq 0),\end{aligned}$$

donc la moyenne conditionnelle *a posteriori* multipliée par la probabilité *a posteriori*. Une autre possibilité qui mèrite d'être explorée ca serait de calculer les HRs pour chaque modèle et d'appliquer le moyennage sur eux, en prenant comme poids la probabilité *a posteriori* du modèle.

La variance du coefficient de régression est égale à :

$$\begin{aligned}V(\widehat{\theta}) &= \mathrm{E}(\widehat{\theta}^2) - \left(\sum_{i=1}^{K} \widehat{\theta}_i \, \mathrm{P}(M_i|D) \right)^2 \\ &= \sum_{i=1}^{K} \mathrm{P}(M_i|D) \times \left(Var(\widehat{\theta}|M_i, D) + \widehat{\theta}_i^2 \right) - \left(\sum_{i=1}^{K} \mathrm{P}(M_i|D) \widehat{\theta}_i \right)^2 \\ &= \sum_{i=1}^{K} \mathrm{P}(M_i|D) \times Var(\widehat{\theta}|M_i, D) + \sum_{i=1}^{K} \widehat{\theta}_i^2 \, \mathrm{P}(M_i|D) - \left(\sum_{i=1}^{K} \widehat{\theta}_i \, \mathrm{P}(M_i|D) \right)^2 \\ &= \sum_{i=1}^{K} \mathrm{P}(M_i|D) \times Var(\widehat{\theta}|M_i, D) + \sum_{i=1}^{K} \mathrm{P}(M_i|D) \times \left(\widehat{\theta}_i - \sum_{i=1}^{K} \widehat{\theta}_i \, \mathrm{P}(M_i|D) \right)^2 .\end{aligned}$$

La variance est donc composée par deux parties : une partie qui représente une variance pondérée par la probabilité *a posteriori* du modèle et une autre

partie qui dépend de la stabilité de la variable à travers les modèles. Plus les estimés sont semblables, plus la variance *a posteriori* sera petite.

Exemple 4.2.1. *Reprenons l'example traité à la section 2.2 et appliquons le moyennage bayésien des modèles. Dans le tableau 4.1, la notation « MP » signifie moyenne a posteriori et la notation « SD » représente l'écart-type a posteriori.*

Tableau 4.1: Résultats du moyennage bayésien

	$P(\theta \neq 0)$	MP	SD	$modèle1$	$modèle2$
$SBP0$	*1*	*0,040*	*0,006*	*0,040*	*0,040*
$T.1$	*0,084*	*0,003*	*0,055*	*.*	*0,033*
$nVar$				*1*	*2*
BIC				*-36,055*	*-31,266*
$postprob$				*0,916*	*0,084*

Le meilleur modèle est le modèle qui contient seulement SBP0, avec une probabilité a posteriori de 0,916. La probabilité a posteriori pour la variable SBP0 est 1, ce qui implique une très forte évidence que son coefficient est non nul. Nous pouvons constater que les estimés pour SBP0 et T.1 dans le modèle 2 sont les mêmes que dans le modèle ordinaire de Cox. De plus comme la probabilité a posteriori pour SBP0 est 1, même la moyenne a posteriori est pareille– étant donné que la moyenne a posteriori est le produit de la moyenne conditionnelle a posteriori et de la probabilité a posteriori moyenne pondérée où les poids sont données par la probabilité a posteriori. Par contre, la moyenne a posteriori pour T.1 est 0,003 ce qui est très loin de l'estimé initial. La probabilité a posteriori pour la variable T.1 est de 0,084 ce qui implique que le paramètre est nul.

Généralement dans le cadre du modèle de Cox, la pratique commune est de présenter la valeur p comme un indicateur de l'importance de la variable. En appliquant le moyennage bayésien des modèles nous prenons en considération l'incertitude dans le choix du modèle. La probabilité *a posteriori* que le paramètre soit non nul est beaucoup plus informative qu'une valeur p.

Ce modèle sera mis en pratique au chapitre cinq afin de comparer les résultats de cette approche aux résultats de l'approche classique.

Chapitre 5

APPLICATION

Afin d'illustrer les méthodes décrites aux chapitres précédents, nous introduisons un exemple, qui est l'exemple présenté dans Kettani *et al.*(2009). Dans cet exemple, un devis cas-témoins emboîté dans une cohorte de nouveaux utilisateurs d'antihypertenseurs est utilisé pour évaluer l'association entre l'adhésion à la thérapie avec antihypertenseurs et l'incidence des accidents vasculaires cérébraux. Les résultats obtenus avec un devis cas-témoins sont comparés :

- aux résultats obtenus dans un devis cohorte, en utilisant la variable exposition qui varie dans le temps, sans faire d'ajustement pour le temps passé depuis l'exposition ;
- aux résultats obtenus en utilisant l'exposition cumulative pondérée par le passé récent ;
- aux résultats obtenus selon la méthode bayésienne.

La cohorte utilisée consiste en de nouveaux utilisateurs d'antihypertenseurs entre le 1^{er} janvier 1999 et 31 décembre 2004, âgés entre 45 et 85 ans (pour plus de détails voir l'annexe A). Dans ce chapitre nous présentons d'abord les résultats obtenus en utilisant le devis cas-témoins. Ensuite nous allons présenter les résultats des analyses effectuées en utilisant le devis cohorte (sans et avec l'ajustement pour le temps passé depuis l'exposition). Nous terminons ce chapitre par la présentation de l'implémentation de la méthode bayésienne.

5.1. Devis cas-témoins

La cohorte utilisée dans cette étude consiste en 83 267 nouveaux utilisateurs d'antihypertenseurs entre le 1^{er} janvier 1999 et 31 décembre 2004, âgés entre 45 et 85 ans, avec une moyenne d'âge de 65 ans. Parmi ces patients, 37,3% étaient des hommes, 11,6% bénéficiaient de l'aide sociale, 8,6% étaient diabétiques et 19,5% dyslipidémiques. Dans le tableau 5.1, nous trouvons la description complète des patients de la cohorte utilisée.

Une étude cas-témoins emboîtée à l'intérieur de la cohorte a été effectuée afin d'évaluer le risque d'accident vasculaire cérébral non fatal en relation avec le niveau d'adhésion aux agents antihypertenseurs. Tous les sujets ayant développé un premier événement cérébrovasculaire (cas) pendant la période de suivi ont été identifiés. Pour chaque cas, au plus 15 témoins ont été sélectionnés aléatoirement à partir de la population source (c'est-à-dire la cohorte) parmi les sujets à risque de développer un événement cérébrovasculaire au moment où survient le cas. Les témoins devaient donc avoir au moins le même temps de suivi que les cas. La probabilité d'être témoin est proportionnelle au temps qu'un sujet contribue au dénominateur de son taux. La date de l'événement du cas et celle de sélection du témoin a été définie comme date index. Notons qu'un sujet sélectionné comme témoin sera toujours éligible pour devenir un cas et un témoin peut être utilisé plus d'une fois. Les cas et les témoins ont été appariés pour l'âge et la durée de suivi.

Tableau 5.1: Caractéristiques des patients. Voir l'annexe A pour les abréviations.
(Moyenne ± écart-type)

		Classe d'antihypertenseur					
	Cohorte	Diurétiques	β-bloquants	CCB	ACE	ARB	Thérapie combinée
No. patients	83 267	21 542	8 601	11 257	21 552	16 644	3671
No. jours de suivi	1 160±614	1 159±625	1 256±630	1 198±615	1 196±614	1 088±583	938±540
Age moyenne	65±10	66±9,9	62±10	66±10	64±10	64±10	63±10
Hommes	37,3%	28,0%	33,9%	40,8%	42,4%	40,8%	42,9%
Sécurité sociale	11,6%	10,6%	14,7%	11,4%	12,1%	10,2%	13,1%
Diabète	8,6%	3,2%	3,5%	4,4%	19,2%	7,7%	6,1%
Dyslipidémia	19,5%	17,0%	16,8%	16,4%	23,7%	20,7%	18,7%

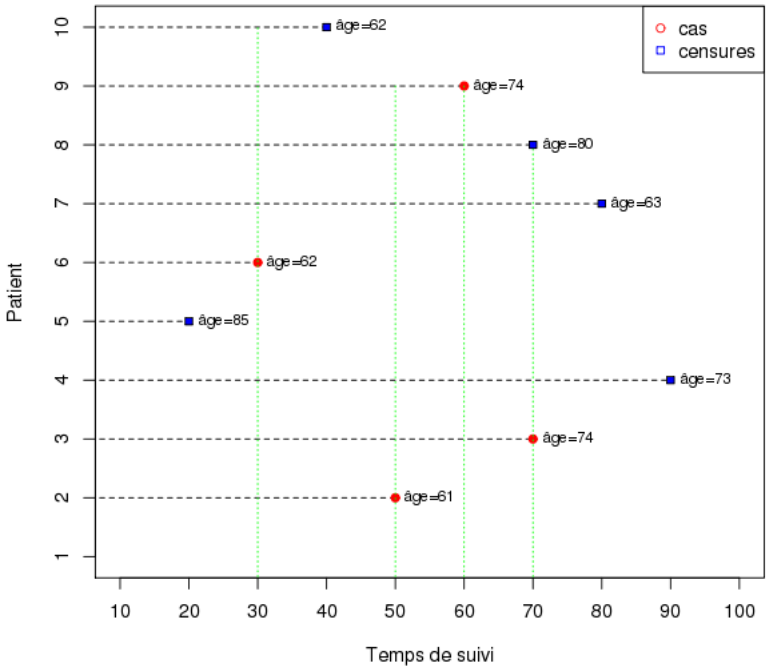


Figure 5.1: Ensemble à risque à l'apparition d'un cas

Sur la figure 5.1, le patient six, par exemple, est un cas et pour lui nous allons d'abord regarder ceux qui sont à risque au temps 30, notamment les patients deux, trois, quatre, sept, huit, neuf et dix. Ensuite, parmi eux nous allons choisir ceux ayant le même âge (plus ou moins un an). Donc, les témoins possibles pour le patient six sont les patients deux, sept et dix.

Dans le tableau 5.2, nous présentons les résultats des analyses effectuées. Les résultats « bruts » sont des résultats des analyses univariées. Les résultats « ajustés » sont les résultats d'une analyse multivariée.

L'analyse multivariée a mis en évidence que pour les sujets ayant un suivi plus long que 365 jours l'« adhésion aux AH $\geq$80 % » (voir annexe A pour les abréviations) est associée avec une baisse du risque d'un événement cérébrovasculaire (RR=0,78 ; IC :(0,70 ;0,86)), comparativement à l'« adhésion aux AH <80 % ». De même, le modèle ajusté a montré que le fait d'être un homme, l'« assistance sociale », l'usage d'« antiplaquettaires », le « cds », la « dyslipidémie »

Tableau 5.2: Risque relatif d'une maladie cérébrovasculaire. Voir l'annexe A pour les abréviations.

	Temps de suivi <= 365		Temps de suivi > 365	
	Brut	Ajusté	Brut	Ajusté
Adhésion aux AH < 80 %	Référence	Référence	Référence	Référence
Adhésion aux AH ≥ 80 %	0,87	0,86(0,72-1,03)	0,75	0,78(0,70-0,87)
Monothérapie	Référence	Référence	Référence	Référence
Bithérapie	1,57	1,35(1,17-1,57)	1,09	1,03(0,94-1,14)
Trithérapie	1,82	1,47(0,92 ;2,35)	1,35	1,09(0,90-1,33)
Sexe	1,48	1,39(1,24-1,56)	1,28	1,17(1,06-1,28)
Assistance sociale	1,19	1,18(0,92-1,50)	1,66	1,47(1,23-1,76)
CAD	2,88	2,68(2,47-4,22)	1,37	2,04(1,73-2,40)
IC	2,80	3,57(2,07-6,15)	1,56	2,29(1,70-3,09)
PAD	9,47	8,40(7,56-14,93)	3,18	4,56(3,67-5,67)
Autres MC	2,79	3,25(2,50-4,22)	1,62	2,36(2,04-2,74)
≥ 2 MC	5,50	7,05(5,50-9,83)	2,73	3,54(3,02-4,15)
Antiplaquettaires	1,69	2,06(1,68-2,53)	1,15	1,70(1,48-1,97)
Pas de diabète	Référence	Référence	Référence	Référence
Diabète non traité	1,19	1,11(0,84-1,48)	1,17	1,12(0,93-1,34)
Nouveau diabète	1,13	0,95(0,68-1,33)	1,52	1,34(0,97-1,86)
Adhésion aux AA < 80 %	1,47	1,31(0,81-2,11)	1,60	1,22(0,94-1,44)
Adhésion aux AA ≥ 80 %	1,24	1,17(0,87-1,159)	1,35	1,22(1,03-1,44)
Pas de dyslipidémie	Référence	Référence	Référence	Référence
Dyslipidémie non traité	0,75	0,76(0,51-1,12)	0,78	0,89(0,69-1,15)
Nouvelle Dyslipidémie	1,65	1,19(1,00-1,42)	1,83	1,40(1,18-1,66)
Adhésion aux AHyp < 80 %	1,18	1,17(0,86-1,61)	1,30	1,10(0,93-1,29)
Adhésion aux AHyp ≥ 80 %	0,66	0,68(0,53-0,87)	0,89	0,80(0,70-0,91)
Cds	1,13	1,13(0,90-1,41)	1,47	1,23(1,06-1,43)

nouvellement diagnostiqué et l'« adhésion aux AA ≥80 % » augmentent de façon significative le risque d'un événement cérébrovasculaire.

Comme dans l'étude cas-témoins l'exposition devient protectrice seulement après au moins une année d'exposition, dans les analyses suivantes nous allons examiner seulement les patients qui ont un suivi de plus de 365 jours.

5.2. Analyse de cohorte sans ajustement pour le temps passé depuis l'exposition

Pour divers raisons, la cohorte utilisée pour l'étude de Kettani *et al.*(2009) ne pouvait pas être reconstituée, mais une cohorte similaire a été construite à partir des bases de données électroniques de la Régie de l'Assurance Maladie du Québec (RAMQ). La cohorte utilisée dans cette étude consiste en 82 585 nouveaux utilisateurs d'antihypertenseurs entre le 1er janvier 1999 et le 31 décembre 2004, âgés entre 45 et 85 ans, avec une moyenne d'âge de 65 ans. Parmi ces patients 37,1% étaient hommes, 11,3% bénéficiaient de l'aide sociale, 8,6% étaient diabétiques et 19,5% étaient dyslipidémiques. Dans le tableau 5.3 nous trouvons la description complète des patients de la cohorte utilisée. En regardant les 2 tableaux contenant les caractéristiques (5.3 et 5.1) nous constatons que les deux cohortes sont assez similaires par rapport aux caractéristiques présentées. Par contre, nous ne sommes pas sûrs que l'historique d'exposition est semblable.

Pour cette partie, la cohorte telle que décrite à l'annexe B est utilisée. Pour chaque individu de la cohorte, l'historique de l'exposition est déterminé (c'est-à-dire les périodes d'exposition et de non-exposition). Donc, les variables « Adhésion aux AH ≥80 % » et « Adhésion aux AH <80 % » sont remplacées par les variables « Exposé », respectivement « Non exposé ». Nous rappelons que dans l'étude cas-témoins les cas et les témoins étaient appariés pour l'âge, ce qui n'est pas le cas dans le cadre de cette analyse (et celles qui suivent) où l'« âge » sera ajusté comme une covariable dans le modèle. Dans le tableau 5.4, nous retrouvons les résultats de l'analyse réalisée sur la cohorte sans ajustement pour le temps passé depuis l'exposition. Nous pouvons constater que les estimateurs ponctuels sont

Tableau 5.3: Caractéristiques des patients dans la cohorte utilisée pour les analyses.Voir l'annexe A pour les abréviations. (Moyenne ± écart-type)

		Classe d'antihypertenseur					
	Cohorte	Diurétiques	β-bloquants	CCB	ACE	ARB	Thérapie combinée
No. patients	82585	21 449	8 490	11 210	21 310	16 477	3649
No. jours de suivi	1 183±623	1 188±638	1 282±637	1 222±623	1 218±623	1 103±592	956±553
Âge	65±9,6	66±9,6	63±9,6	66±9,6	65±9,4	65±9,4	64±9,5
Hommes	37,1%	27,8%	33,6%	40,7%	42,3%	40,7%	42,6%
Sécurité sociale	11,3%	10,4%	14,4%	11,2%	11,8%	10,0%	12,9%
Diabète	8,6%	3,2%	3,5%	4,4%	19,3%	7,8%	6,0%
Dyslipidémie	19,5%	17,0%	17,0%	16,4%	23,8%	20,8%	18,8%

similaires à ceux obtenus en utilisant le devis cas-témoins pour certaines variables. Par contre, pour d'autre, ils sont très différents.

Tableau 5.4: Taux de risque d'une maladie cérébrovasculaire.
Voir l'annexe A pour les abréviations.

	Taux de risque (IC95%) Temps de suivi plus grand qu'un an	
	Brut	Ajusté
Âge	1,05	1,05(1,04-1,05)
Non exposé	Référence	Référence
Exposé	0,70	0,70(0,64-0,77)
Monothérapie	Référence	Référence
Bithérapie	1,09	0,77(0,71-0,84)
Trithérapie	1,35	0,82(0,70-0,96)
Sexe	1,14	1,18(1,09-1,27)
Assistance sociale	0,77	1,35(1,17-1,57)
CAD	1,12	1,22(1,06-1,41)
IC	1,42	1,33(1,05-1,69)
PAD	2,84	2,71(2,27-3,24)
Autres MC	1,07	1,11(0,97-1,28)
$\geq$ 2 MC	1,59	1,50(1,34-1,69)
Antiplaquettaires	0,64	0,75(0,66-0,85)
Pas de diabète	Référence	Référence
Diabète non traité	1,17	1,04(0,90-1,22)
Nouveau diabète	1,52	1,26(0,97-1,65)
Adhésion aux AA $<$ 80 %	1,60	1,27(1,04-1,56)
Adhésion aux AA $\geq$ 80 %	1,35	1,20(1,04-1,39)
Pas de dyslipidémie	Référence	Référence
Dyslipidémie non traité	0,60	0,70(0,56-0,87)
Nouvelle Dyslipidémie	1,64	1,64(1,42-1,86)
Adhésion aux AHyp. $<$ 80 %	0,84	0,83(0,72-0,95)
Adhésion aux AHyp. $\geq$ 80 %	0,59	0,62(0,55-0,69)
Cds	1,36	1,14(1,03-1,27)

Nous remarquons tout d'abord une baisse dans l'estimé ponctuel de l'exposition. En effet, nous obtenons maintenant un HR de 0,70, ce qui est équivalent à une baisse de 5% de celui obtenu dans l'étude cas-témoins. Pour les variables démographiques (le sexe et l'assistance sociale), de même que pour les covariables liées au diabète et à la dyslipidémie, les points estimés sont similaires. Pour les variables indiquant les comorbidités, tous les estimés sont diminués. Ces estimés

sont loins des estimés obtenus avec le devis cas-témoin et des résultats de la littérature spécialisée. Nous constatons aussi que certains estimés qui n'étaient pas significatifs dans l'étude cas-témoins le deviennent maintenant. Ce fait pourrait être expliqué par une manque de puissance dans l'étude cas-témoins (Wacholder, 2009). Par contre, pour quelques variables les résultats de l'étude de cohorte sont surprenantes. Il s'agit ici des variables « bithérapie », « trithérapie » et « antiplaquéttaires » que nous pouvons penser être des facteurs de risque (une personne ayant une bithérapie est plus à risque d'avoir un événement cérébrovasculaire si elle ne prend pas son traitement comparativement à une personne ayant une monothérapie) et dont les résultats indiquent qu'elles sont des facteurs protectifs.

Les résultats ci-haut sont surprenants. Les différences observées pourraient être causées par diverses raisons. Premièrement, dans le cadre du devis cas-témoins, les cas et les témoins ont été appariés pour l'âge et la durée de suivi, alors que dans la cohorte nous n'avons pas fait d'appariemment pour l'âge ce qui pourrait expliquer les résultats surprenants de l'analyse sur la cohorte. Nous pensons que cet aspect devraient être investigué dans une étude subséquente en faisant un appariement pour l'âge dans la cohorte. Deuxièmement, les différences observées pourraient aussi être causées par l'utilisation de définitions différentes pour les covariables dans l'étude cas-témoins et dans la cohorte (voir annexe B). Donc le contexte de l'étude cas-témoins n'a pas été respecté. Une autre explication possible serait l'utilisation des bases de données différentes pour les analyses (même si les caractéristiques des patients étaient les mêmes, nous ne sommes pas sûrs que l'historique d'exposition était semblable). Finalement, il peut y avoir d'autres problèmes non identifiés et ceci doit d'être investigué car, en moyenne, le résultat obtenu avec un devis cas-témoins devrait être similaire au résultat obtenu avec un devis de cohorte en prenant en considération une exposition variable dans le temps (Essebag *et al.*, 2005 ; Langholz et Richardson, 2009 ; Wacholder, 2009). Prentice et Breslow (1978) ont montré que la vraisemblance conditionnelle utilisée dans la régression logistique conditionnelle a la même forme que la vraisemblance partielle utilisée dans le modèle de Cox, à l'exception du dénominateur, qui inclut dans le premier modèle seulement un nombre limité de témoins et dans

le deuxième tous les sujets à risque. Généralement, le seul désavantage de l'analyse sur un dévis cas-témoins est la précision et la puissance (nous choisissons 15 témoins par rapport à tous les sujets à risque).

Étant donné les différences observées entre les estimés obtenus avec le devis cas-témoins et ceux obtenus avec le devis de cohorte, les modèles suivants seront comparés seulement au modèle ajusté sur la même base des données.

5.3. Analyse de cohorte avec exposition cumulative pondérée par le passé récent

Rappelons que pour l'utilisation de la spline cubique nous allons considérer un polynôme cubique de la forme :

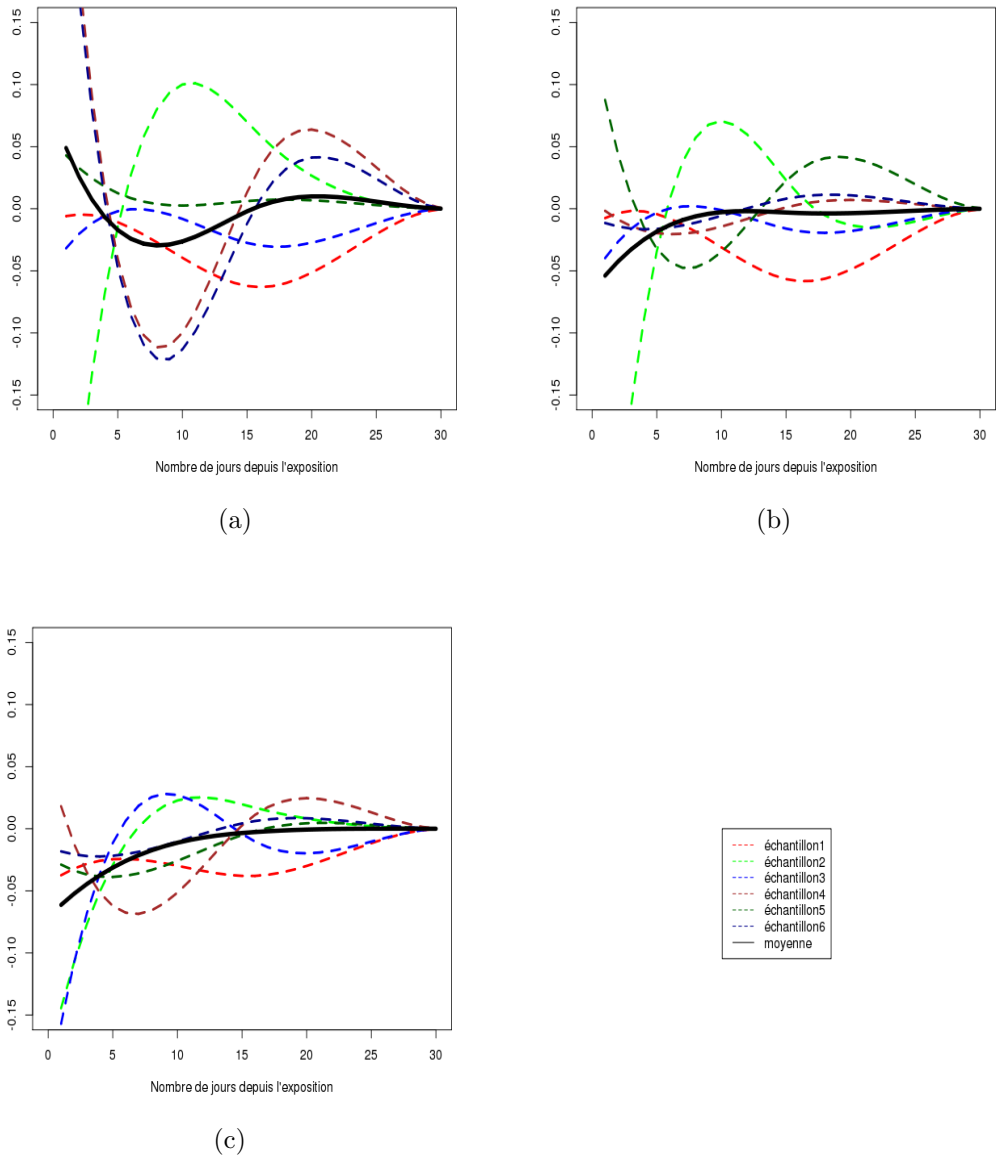
$$a \times x^3 + b \times x^2 + c \times x + d$$

entre deux nœuds consécutifs de sorte que :

(1) chaque polynôme passe par les points extrêmes de l'intervalle sur lequel le polynôme est défini ;

(2) au point de rencontre de 2 polynômes, leurs dérivées première et deuxième sont égales.

Pour cette partie de l'analyse, le modèle de l'exposition cumulative pondérée tel que présenté à la section 3.7 a été utilisé. Il faut mentionner que même si la mémoire vive sur la machine utilisée dans les analyses était de 32 Go, le logiciel R avec lequel a été ajusté ce modèle a une limite de mémoire adressable beaucoup plus petite. L'utilisation du modèle incluant l'exposition cumulative implique la création, pour chaque individu, de l'historique d'exposition aux agents antihypertenseurs (c'est-à-dire l'état de l'exposition pour chaque jour du suivi). Nous avons exploré des librairies créées pour pouvoir stocker et manipuler de grosses matrices (« big memory », « ff ») mais aucune n'était appropriée pour l'utilisation par la suite dans la fonction décrite à la section 3.7. Ces limites d'ordre informatique ont fait en sorte que nous n'avons pas été capable d'appliquer le modèle sur toute la cohorte. Nous avons plutôt regardé des échantillons de 3 500, 7 000 et 15 000 sujets.

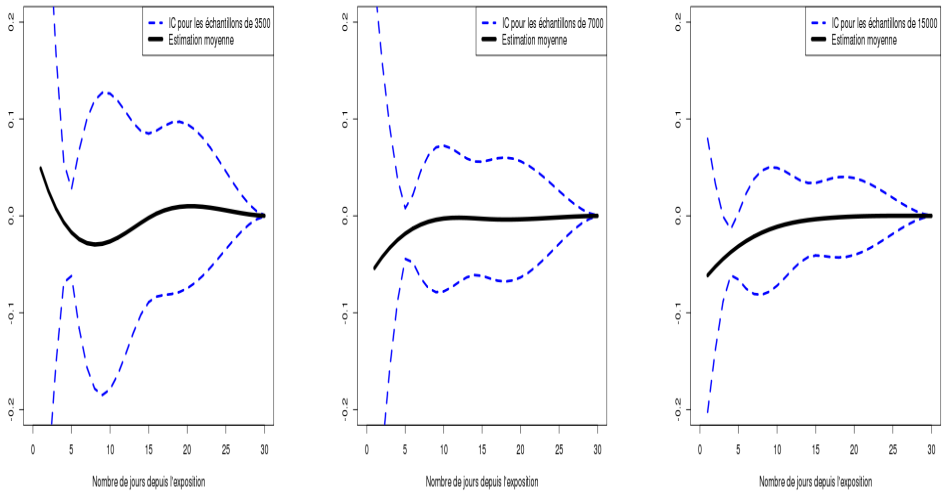


(a) (b)

(c)

Figure 5.2: Fonction de poids estimée sur (a) : Échantillons de 3 500 sujets ; (b) : Échantillons de 7 000 sujets ; (c) : Échantillons de 15 000 sujets.

Pour les échantillons de 3 500 sujets, la fonction de poids semble avoir une forme différente de celle des échantillons de 7 000 et 15 000 sujets. Ce fait pourrait être expliqué par le nombre faible de cas ou bien par l'historique de l'exposition des sujets sélectionnés. À partir des échantillons de 7 000 sujets ces faits posent moins de problèmes, la fonction de poids semble moins variable. Étant donné les



(a) (b) (c)

Figure 5.3: Fonction de poids et intervalle de confiance ponctuel pour (a) : Échantillons de 3 500 sujets ; (b) : Échantillons de 7 000 sujets ; (c) : Échantillons de 15 000 sujets.

différences assez grandes entre les courbes, en déterminant (*a priori*) la forme de la fonction des poids à partir des données et ensuite utiliser celle-là dans le modèle de Cox pour estimer les coefficients pourrait améliorer la précision des estimation.

Sur la figure 5.2(c) nous pouvons constater que la fonction moyenne de « poids » augmente pour atteindre 0 autour de 10 jours après l'arrêt et reste constante à 0 après. Ainsi il semble que dans les 10 premiers jours après l'arrêt il y a un effet protecteur plus faible des agents antihypertenseurs que si nous regardions seulement la journée courante et 10 jours après l'exposition il n'y a plus d'effet d'antihypertenseurs.

Afin de vérifier le comportement de la fonction de poids dans les trois situations, nous avons calculé des intervalles de confiance ponctuels présentés dans la figure 5.3.

Une observation évidente est que l'intervalle de confiance rétrécit avec l'augmentation de la taille de l'échantillon. Nous pouvons aussi constater l'instabilité

au début de la fenêtre d'exposition. Cela peut être expliqué par une des caractéristiques des splines, soit l'instabilité aux extrêmes (à la fin de la fenêtre d'exposition la fonction de poids est contrainte à être zéro, donc ça explique l'absence de variabilité). Nous pensons que la variabilité des estimés serait beaucoup plus petite si les données étaient utilisées pour déterminer la forme de la fonction de poids et ensuite l'introduire dans le modèle pour estimer les coefficients.

Dans le tableau 5.5, nous retrouvons le rapport de risques pour différents historiques d'exposition. Il faut se rappeler que nous avons considéré seulement les sujets ayant été exposés au traitement au moins 365 jours et nous examinons les 30 derniers jours de leur historique du traitement.

Tableau 5.5: Résultats pour différents historiques du traitement

Historique du traitement	Référence	HR
Traité pour les 30 derniers jours	Jamais traité dans le 30 derniers jours	0,70
Traité pour 29 jours, 1 journée avant	Jamais traité dans le 30 derniers jours	0,75
Traité pour 27 jours, 3 journée avant	Jamais traité dans le 30 derniers jours	0,82
Traité pour 20 jours,10 jours plutôt	Jamais traité dans le 30 derniers jours	0,96
Traité pour les 30 derniers jours	Traité pour 20 jours 10 jours avant	0,73

Il semble qu'en utilisant l'exposition cumulative nous obtenons le même estimé qu'en utilisant l'exposition courante (0,70). Comme suggéré par la forme de la fonction de poids nous pouvons dire qu'à partir de dix jours depuis l'exposition il n'y a plus d'effet d'antihypertenseurs (HR=0,96). La dernière ligne du tableau comparant un sujet qui a été exposé pendant les 30 derniers jours à un sujet qui à été exposé pendant 20 jours, mais qui a arrêté le traitement dix jours auparavant met en évidence l'importance des dix derniers jours d'exposition(c'est-à-dire que si la comparaison d'un sujet traité pour les 30 derniers jours est faite avec un sujet non exposé ou avec un sujet exposé mais dix jours auparavant les résultats sont similaires : 0,70 comparativement à 0,73).

Il ne faut pas oublier que ces rapports de risques pour différents historiques du traitement sont calculés à partir des estimés obtenus en utilisant les splines. Il faudra vérifier ces résultats dans une prochaine étude qui devrait investiguer la

possibilité de déterminer *a priori* à partir des données la forme de la fonction des poids et ensuite utiliser celle-là dans le modèle de Cox pour estimer les coefficients.

Afin de pouvoir comparer l'estimation des covariables, dans la figure 5.4 nous présentons les estimés ponctuels obtenus sur la cohorte sans ajustement pour le passé récent et les estimés moyens obtenus sur les échantillons de 3 500, 7 000 et 15 000 sujets (obtenus selon le modèle incluant l'exposition cumulative).

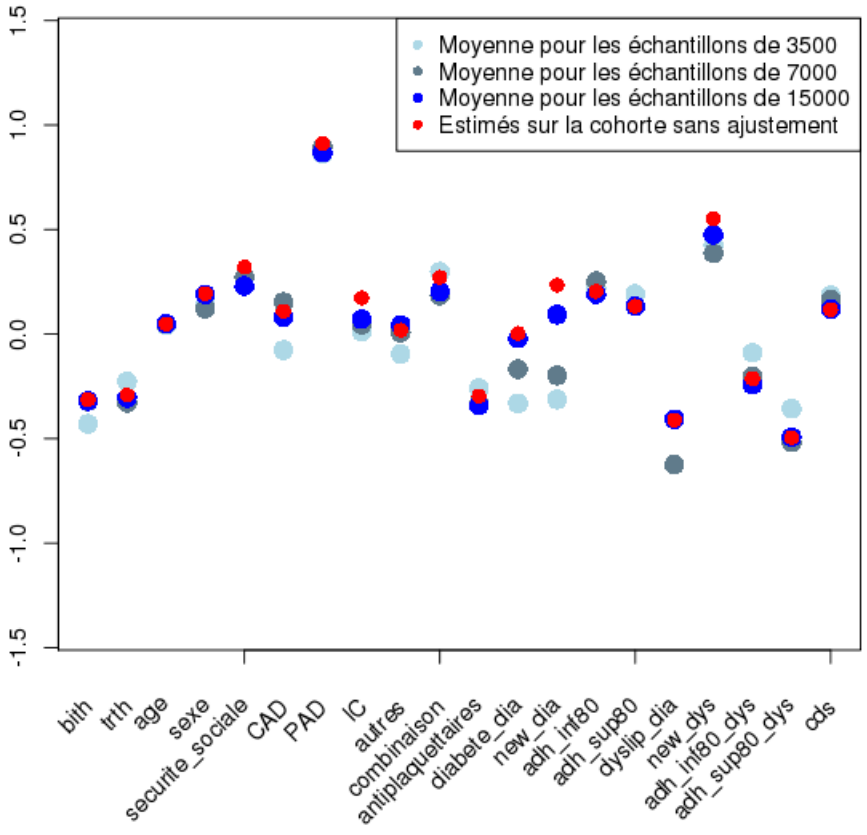


Figure 5.4: Estimation des covariables

Nous pouvons constater qu'en moyenne les estimés sont similaires. Nous pouvons aussi remarquer que pour la variable « âge », l'estimation est très bonne peu importe la taille de l'échantillon, ce qui pourrait être expliqué par le fait que les échantillons utilisés représentaient bien la population totale quant à leur moyenne d'âge. Par contre pour quelques variables (comme « diabete dia » ou « new dia »)

il semble y avoir une différence. Ceci pourrait être expliqué par le petit nombre d'échantillons ou bien par le nombre très petit de sujets ayant ces diagnostiques.

5.4. Analyse utilisant le moyennage bayésien de modèles

Dans cette section, nous allons présenter les résultats obtenus en appliquant le moyennage bayésien de modèles. La comparaison dans cette section vise le modèle de Cox avec l'exposition qui varie dans le temps sans faire d'ajustement pour le passé récent et le moyennage bayésien de modèles, ajustés sur la même base des données.

Dans le tableau 5.4 nous retrouvons les résultats de l'analyse bayésienne (nous avons omis d'inclure les variables qui n'étaient pas sélectionnées dans aucun modèle).

Les cinq meilleurs modèles ont une probabilité *a posteriori* cumulative de un. Il semble que le meilleur modèle soit le modèle qui contient les variables : « exposition », « type de thérapie », « age »,« sexe », « assistance sociale », « PAD », « autres MC », « $\geq$ 2 MC »,« antiplaquettaires », et « dyslipidémie », avec une probabilité *a posteriori* de 0,591. La probabilité *a posteriori* pour les variables mentionnées est 1, ce qui implique une très forte évidence que leurs coefficients sont non nuls.

Dans le tableau 5.8 nous présentons les estimations pour les HRs et les intervalles de confiance obtenus selon le modèle de Cox sans ajustement pour le passé récent et ceux obtenus selon le moyennage bayésien des modèles pour les variables sélectionnées au moins une fois dans le modèle selon l'approche bayésienne (si la variable n'a jamais été sélectionnée nous ne pouvons pas obtenir une estimation *a posteriori*). Nous pouvons constater que l'estimé de l'exposition en utilisant le moyennage bayésien des modèles est HR= 0,73 ce qui est très proche de l'estimé obtenu avec la cohorte sans ajustement pour le passé récent (0,70). Les estimés ponctuels de même que les intervalles de confiance pour les covariables sont similaires aux estimés obtenus en utilisant le modèle de Cox sans ajustement pour le passé récent pour les variables sélectionnées dans chacun des cinq modèles et très différents pour les variables qui apparaissent dans un ou deux modèles, ce

Tableau 5.6: Résultats du moyennage bayésien de modèles. Voir l'annexe A pour les abréviations.

	$P(\theta \neq 0)$ en %	Moyenne *a posteriori*	SD	modèle 1	modèle 2	modèle 3	modèle 4	modèle 5
exposition	100,0	-0,31	0,05	-0,31	-0,31	-0,31	-0,31	-0,31
âge	100,0	0,05	0,002	0,05	0,05	0,05	0,05	0,05
sexe	100,0	0,18	0,04	0,18	0,18	0,18	0,18	0,18
monothérapie	100,0							
bithérapie		-0,280	0,04	-0,28	-0,28	-0,28	-0,28	-0,28
trithérapie		-0,23	0,08	-0,23	-0,23	-0,24	-0,23	-0,23
assistance sociale	100,0	0,32	0,07	0,33	0,31	0,32	0,32	0,31
CHD	13,1	0,02	0,06			0,15		0,15
IC	5,9	0,01	0,06				0,23	
PAD	100,0	0,92	0,09	0,92	0,92	0,94	0,93	0,94
$\geq$ 2 MC	100,0	0,32	0,06	0,32	0,30	0,34	0,33	0,33
antiplaquettaires	100,0	-0,29	0,06	-0,30	-0,29	-0,28	-0,29	-0,27
Pas de dyslipidémie	100.0							
Dyslipidémie non-traitée		-0,38	0,11	-0,38	-0,38	-0,39	-0,38	-0,38
Nouvelle dyslipidémie		0,53	0,07	0,54	0,54	0,53	0,53	0,53
Adhésion aux AHyp. $<$ 80 %		-0,16	0,07	-0,16	-0,16	-0,17	-0,16	-0,17
Adhésion aux AHyp. $\geq$ 80 %		-0,45	0,05	-0,45	-0,45	-0,46	-0,45	-0,46
cds	25,3	0,03	0,06		0,13			0,13
nVar				9	10	10	10	11
BIC				-821,13	-819,15	-817,54	-816,52	-815,35
Probabilité *a posteriori* du modèle				0,59	0,22	0,10	0,06	0,03

(voir annexe A pour les abréviations)

qui peut être expliqué par le fait que la moyenne *a posteriori* est une moyenne pondérée où le poids est la probabilité *a posteriori* du modèle.

Tableau 5.8: Comparaison entre les estimations sur la cohorte sans ajustement et le moyennage bayésien. Voir l'annexe A pour les abréviations.

	HR ajusté et IC95%)	
	Modèle de Cox sans ajustement pour le passé récent	Moyennage bayésien des modèles
Âge	1,05(1,04-1,05)	1,05(1,04-1,05)
Non exposé	Référence	Référence
Exposé	0,70(0,64-0,77)	0.73(0,66-0,82)
Monothérapie	Référence	Référence
Bithérapie	0,77(0,71-0,84)	0,76(0,70-0,82)
Trithérapie	0,82(0,70-0,96)	0,80(0,68-0,93)
Sexe	1,18(1,09-1,27)	1,19(1,10-1,29)
Assistance sociale	1,35(1,17-1,57)	1,38(1,19-1,60)
CAD	1,22(1,06-1,41)	1,02(0,91-1,14)
IC	1,33(1,05-1,69)	1,01(0,90-1,14)
PAD	2,71(2,27-3,24)	2,52(2,11-3,00)
$\geq$ 2 MC	1,50(1,34-1,69)	1,38(1,23-1,54)
Antiplaquettaires	0,75(0,66-0,85)	0,75(0,66-0,85)
Pas de dyslipidémie	Référence	Référence
Dyslipidémie non traité	0,70(0,56-0,87)	0,68(0,55-0,85)
Nouvelle Dyslipidémie	1,64(1,42-1,86)	1,71(1,48-1,96)
Adhésion aux AHyp. $<$ 80 %	0,83(0,72-0,95)	0,85(0,74-0,98)
Adhésion aux AHyp. $\geq$ 80 %	0,62(0,55-0,69)	0,64(0,57-0,71)
Cds	1,14(1,03-1,27)	1,03(0,92-1,16)

Au vue des grandes différences entre les estimés obtenus avec le devis cas-témoins et ceux obtenus avec le devis de cohorte (tous les modèles), nous pouvons s'interroger sur le fait que le contexte de l'étude cas-témoins n'est pas respecté. Il peut aussi y avoir d'autres problèmes non identifiés et cela doit d'être investigué car, tel que mentionné auparavant, le résultat obtenu avec un devis cas-témoins devrait être similaire au résultat obtenu avec un devis de cohorte en prenant en considération une exposition variable dans le temps.

Chapitre 6

CONCLUSION

Dans ce mémoire, nous avons comparé les résultats obtenus dans le cadre d'une étude cas-témoins nichée dans la cohorte avec les résultats d'une étude utilisant un devis de cohorte. La comparaison a été faite :

- avec les résultats obtenus dans un devis cohorte, en utilisant la variable exposition qui varie dans le temps sans faire d'ajustement pour le temps passé depuis l'exposition ;
- avec les résultats obtenus en utilisant l'exposition cumulative pondérée par le passé récent ;
- avec les résultats obtenus selon la méthode bayésienne.

La comparaison avec les résultats de la cohorte sans ajustement pour le temps passé depuis l'exposition a mis en évidence que pour l'exposition les résultats étaient semblables, de même que pour les variables démographiques (le sexe et l'assistance sociale) et pour les covariables liées au diabète (« diabète non traité », « nouveau diabète », « Adhésion aux AA $\geq$80% », « Adhésion aux AA $<$ 80% ») et à la dyslipidémie (« dyslipidémie non traité », « nouvelle dyslipidémie », « Adhésion aux AHyp $\geq$80% », « Adhésion aux AHyp $<$ 80% »). Pour les variables indiquant les comorbidités, tous les estimés sont plus petits. Ces estimés sont loins des estimés obtenus avec le devis cas-témoin et des résultats de la littérature spécialisée. Nous constatons aussi que certains estimés qui n'étaient pas significatifs dans l'étude cas-témoins le deviennent maintenant. Ce fait pourrait être expliqué par une manque de puissance dans l'étude cas-témoins (Wacholder,

2009). Par contre, pour quelques variables les résultats de l'étude de cohorte sont surprenantes.

Les différences observées pourraient être causées par divers raisons. Premièrement, dans le cadre du devis cas-témoins, les cas et les témoins ont été appariés pour l'âge et la durée de suivi, alors que dans la cohorte nous n'avons pas fait d'appariemment pour l'âge ce qui pourrait expliquer les résultats surprenants de l'analyse sur la cohorte. Nous pensons que cet aspect devraient être investigué dans une étude subséquente en faisant un appariement pour l'âge dans la cohorte. Deuxièmement, les différences observées pourraient aussi être causées par l'utilisation de définitions différentes pour les covariables dans l'étude cas-témoins et dans la cohorte (voir annexe B). Donc le contexte de l'étude cas-témoins n'a pas été respecté. Troisièmement, une autre explication possible serait l'utilisation des bases de données différentes pour les analyses (même si les caractéristiques des patients étaient les mêmes, nous ne sommes pas sûrs que l'historique d'exposition était semblable). Finalement, il peut y avoir d'autres problèmes non identifiés et ceci doit d'être investigué car, en moyenne, le résultat obtenu avec un devis cas-témoins devrait être similaire au résultat obtenu avec un devis de cohorte en prenant en considération une exposition variable dans le temps (Essebag *et al.*, 2005 ; Langholz et Richardson, 2009 ; Wacholder, 2009).

Étant donné les différences observées entre les estimés obtenus avec le devis cas-témoins et ceux obtenus avec le devis de cohorte en prenant en considération l'exposition au traitement qui varia dans le temps, les modèles subséquents ont été comparés seulement au modèle ajusté sur la même base de données.

L'utilisation du modèle incluant l'exposition cumulative pondérée par le passé récent implique la création pour chaque individu de l'historique de son exposition aux agents antihypertenseurs (c'est-à-dire le statut « exposé » ou « non exposé » pour chaque jour du suivi). Comme le suivi moyen était de 1183 jours, ceci impliquait la création et la manipulation d'une matrice avec plus de 95 millions de lignes et 28 colonnes. Même si la mémoire vive sur la machine utilisée dans cette partie de l'analyse était de 32 Go, le logiciel R qui a été utilisé pour réaliser ce modèle a une limite de mémoire adressable beaucoup plus petite. Nous avons

exploré des librairies spécialement créées pour pouvoir stocker et manipuler de grosses bases de données (« big memory », « ff ») mais aucune n'était appropriée pour l'utilisation par la suite dans la fonction « WCE ».

Ces limites d'ordre informatique ont fait en sorte que nous n'avons pas été capable d'appliquer le modèle sur toute la cohorte, mais nous avons regardé quelques échantillons de 3 500, de 7 000 et de 15 000 sujets. Les résultats sur les échantillons de 3 500 sujets sont différents de ceux sur les échantillons de 7 000 et 15 000 sujets. Ce fait pourrait être expliqué par le nombre faible de cas ou bien par l'historique de l'exposition des sujets sélectionnés. À partir des échantillons de 7 000 sujets la fonction de poids semble moins variable (excepté dans les extrêmes, faiblesse reconnue des splines). Étant donné les différences assez grandes entre les courbes, en déterminant (*a priori*) la forme de la fonction des poids à partir des données et ensuite utiliser celle-là dans le modèle de Cox pour estimer les coefficients pourrait améliorer la précision des estimations. Nous avons constaté que la fonction de « poids » augmente pour atteindre zéro autour de dix jours après l'arrêt et reste constante à zéro après. Ainsi, il semble que dans les dix premiers jours après l'arrêt il y a un effet protecteur plus petit des agents antihypertenseurs que si nous regardions seulement la journée courante et dix jours après l'exposition il n'y a plus d'effet d'antihypertenseurs. La comparaison des estimés ponctuels obtenus pour les covariables selon la cohorte sans ajustement pour le passé récent avec ceux obtenus en utilisant l'exposition cumulative pondérée n'a pas montré de différences importantes. Nous considérons qu'une prochaine étude devrait investiguer la possibilité de déterminer *a priori* à partir des données la forme de la fonction des poids et ensuite utiliser celle-là dans le modèle de Cox pour estimer les coefficients.

Le moyennage bayésien de modèles a mis en évidence des résultats semblables aux deux approches précedentes. L'incertitude dans le choix du modèle n'était pas trop grande, étant donné que les cinq meilleurs modèles avait une probabilité *a posteriori* cumulative de un.

En conclusion, l'utilisation de l'exposition cumulative pondérée par le passé récent semble être un bon choix, surtout si nous sommes intéressés à comparer des historiques d'exposition spécifiques. De plus, ce modèle peut donner des informations très utiles sur la façon dont l'effet s'accumule (ou diminue) dans le temps, mais il faut être prudent quand il s'agit de grosses bases de données. Une autre limite se réfère à l'impossibilité d'imposer *a priori* aux « poids » qu'ils soient positifs. Ceci peut poser des problèmes dans le cadre des études où il est jugé nécessaire que les estimés pour la fonction de poids soient non négatifs. Dans l'étude présentée dans ce mémoire les « poids » négatifs reflètent l'effet protecteur des agents antihypertenseurs.

Annexe A

ABRÉVIATIONS UTILISÉES DANS LE MÉMOIRE

AH: antihypertenseur

AA: agent antidiabètique

AHyp: agent antihyperlipidémique

CAD: maladie coronarienne

IC: agent antihyperlipidémique

MC: maldie cardiovasculaire

PAD: maladie artérielle périphérique

$\geq$ 2**MC:** combinaison d'au moins deux maladies cardiovasculaires

BB: bêta-bloqueurs

BCC: bloqueurs des canaux calciques

IECA: inhibiteurs de l'enzyme de conversion de l'angiotensine

ARA: antagonistes des récepteurs à l'angiotensine

RR: risque relative

HR: hazard ratio

OR: odds ratio

IM: infarctus du myocarde

Annexe B

DÉFINITION DE LA COHORTE

B.1. Sources des données

L'étude présente est effectuée à partir des bases de données électroniques de la Régie de l'Assurance Maladie du Québec (RAMQ) et celles de MED-ECHO. Les banques de la RAMQ, organisme administrateur des programmes d'assurance de la province, contient des informations issues de trois types de fichiers.

Le premier fichier fournit les données démographiques, tels que l'âge, le sexe, l'année de décès et le code postal, de toutes les personnes couvertes par le régime d'assurance-maladie.

Le second fichier est celui des données médicales. Il contient toutes les informations relatives aux services médicaux reçus rémunérés à l'acte, plus précisement leur nature, la date à laquelle ils ont été rendus, le type d'établissement (hospitalier ou ambulatoire) où ils ont été effectués, et leurs codes diagnostiques, identifiés selon la classification internationale des maladies (CIM-9). Les codes de procédures y figurent également et ils sont déterminés selon la classification canadienne des procédures diagnostiques, thérapeutiques et chirurgicales.

Enfin, le troisième fichier dit pharmaceutique comprend les données concernant les médicaments prescrits qui sont dispensés en milieu communautaire et remboursés par le régime provincial d'assurance-médicaments. Ces données sont celles recueillies par le pharmacien lors de l'exécution d'une ordonnance à savoir le nom (dénomination commune internationale), le dosage, la quantité de médicament remise au patient, la date et la durée de la prescription ainsi que la spécialité

du médecin prescripteur. Ce fichier contient aussi les dates de début et de fin de couverture de l'individu couvert par le régime d'assurance-médicaments. Les deux premiers fichiers rassemblent les données sur toutes les personnes couvertes par le régime d'assurance-maladie, c'est-à-dire sur la population totale du Québec. Les données du fichier pharmaceutique sont celles des bénéficiaires du régime public d'assurance-médicaments qui représentent près de 43 % de la population québécoise. Ces bénéficiaires sont les individus prestataires de l'assurance-emploi, les personnes âgées de plus de 65 ans et depuis 1997, les personnes de moins de 65 ans n'ayant pas accès à un régime privé d'assurance collective.

Les banques de données de MED-ECHO contiennent les informations sur les patients hospitalisés dans les établissements québécois de soins généraux et spécialisés. Elles sont complétées pour les soins de courte durée et les chirurgies d'un jour. Nous y retrouvons la date d'admission, le diagnostic principal et jusqu'à 15 diagnostiques secondaires ainsi que la durée d'hospitalisation. Les quatre fichiers ont pour clé commune les numéros d'assurance-maladie des sujets, lesquels sont brouillés afin de préserver la confidentialité des données. Ces fichiers ont déjà été utilisés à des fins de recherche pharmacoépidémiologiques et le fichier pharmaceutique, en particulier, a été validé.

B.2. Définition de la cohorte

Les patients inclus dans la cohorte sont ceux âgés entre 45 et 85 ans, ayant un diagnostic d'hypertension artérielle essentielle (codes CIM-9 401) et ayant nouvellement initié un traitement avec un médicament appartenant à une des cinq principales classes d'antihypertenseurs : diurétiques, bêta-bloqueurs, bloqueurs des canaux calciques, inhibiteurs de l'enzyme de conversion de l'angiotensine ou antagonistes des récepteurs de l'angiotensine II, ou une combinaison d'au moins deux agents de ces classes, entre le 1er janvier 1999 et le 31 décembre 2004. La date de la nouvelle intention de traitement a été définie comme date d'entrée dans la cohorte. Pour être considéré comme nouvel utilisateur, le patient ne doit pas avoir reçu de prescription d'antihypertenseurs dans les deux années précédant la date d'entrée dans la cohorte. Les sujets doivent d'ailleurs avoir été inscrits au régime

de l'assurance-médicaments pendant au moins 24 mois avant la date d'entrée dans la cohorte afin de vérifier l'absence de médicaments antihypertenseurs dans le fichier pharmaceutique et de pouvoir valider le critère de nouvelle prescription d'agents antihypertenseurs. En plus, seulement les patients qui ont eu au moins trois ordonnances dans les 6 mois suivants la date d'entrée ont été inclus. Pour s'assurer que le patient nécessite un traitement pharmacologique, il doit voir son médecin et remplir au moins une ordonnance pour chaque période de 1,5 ans.

Les patients présentant des antécédents de maladies cardio-vasculaires ont été exclus de l'étude, ce qui a été vérifié par l'absence de codes diagnostiques et des codes des procédures médicales reliées à ces maladies dans le fichier des services médicaux cinq ans avant la date d'entrée dans la cohorte, ainsi que l'absence de médicaments marqueurs des maladies cardio-vasculaires dans le fichier pharmaceutique dans les deux années précédant cette même date.

Ces pathologies sont énumérées ci-dessous :

(1) cardiopathie ischémiques, définie par l'un des éléments suivants : infarctus du myocarde, angor (angine de poitrine) ou autres formes de cardiopathies ischémiques ; recours à une procédure médicale telle que la pose d'une endoprothèse vasculaire, une angioplastie ou un pontage coronarien ; utilisation de vasodilatateurs de type nitrates ;

(2) maladie vasculaire cérébrales : recours à des procédures médicales associées telles que l'endartérectomie ; prescription de nimodipine ;

(3) insuffisance cardiaque : utilisation des médicaments suivants : furosémide seul ou associé à i) digoxine, ii) IECA, iii) spironolactone, iv) ß-bloqueurs ou carvédilol ;

(4) arythmies : recours à une procédure médicale impliquant un stimulateur ou un fibrillateur électrique ; utilisation de médicaments antiarythmiques (amiodarone, digoxine, quinidine, disopyramide, flécaïnide, méxilétine, procaïnamide, propafénone ou sotalol) ;

(5) maladie artérielle périphérique : procédures médicales de revascularisation non coronarienne ; utilisation de pentoxifylline ;

(6) utilisation de médicaments antiplaquettaires, d'acide acétylsalicylique à faibles doses ou d'anticoagulants dans les deux ans précédents la date d'entrée dans la cohorte.

Le suivi des patients s'est fait de la date d'entrée dans la cohorte jusqu'à la survenue d'un premier AVC ou la fin de l'étude (30 décembre 2005). Les patients dont la couverture par le régime d'assurance-médicaments de la RAMQ a pris fin et qui sont décédés avant la fin de l'étude de même que ceux qui n'ont pas renouvellé leur dernière prescription d'AH dans un délai de 1,5 années ont été censurés. Seul le premier événement cérébrovasculaire qui est survenu durant le suivi a été analysé.

B.3. Définition des covariables dans le devis cas-témoin et dans la cohorte

Pour l'étude cas-témoins emboîtée à l'intérieur de la cohorte tous les sujets ayant développé un premier événement cérébrovasculaire (cas) pendant la période de suivi ont été identifiés. Pour chaque cas, les témoins ont été sélectionnés parmi les sujets à risque de développer un événement cérébrovasculaire au moment où survient le cas. Les témoins devaient donc avoir au moins le même temps de suivi que les cas et le même âge (plus ou moins un an). Il est important à noter que la date de l'événement du cas et celle de sélection du témoin a été définie comme date index.

Toutes les analyses ont tenu compte du sexe des patients, identifié à la date d'entrée dans la cohorte. Le statut socio-économique a été estimé approximativement à partir du type de programme provincial de remboursement des médicaments dont bénéficie un patient à la date d'entrée dans la cohorte. Le diabète considéré comme une variable dichotomique (oui/non) a été identifié à partir des codes diagnostiques CIM-9 250 ou de l'utilisation des médicaments hypoglycémiants un an avant la date d'entrée dans la cohorte jusqu'à la date index. La dyslipidémie, également considérée comme une variable dichotomique (oui/non), a été identifiée grâce aux codes diagnostiques CIM-9 272 ou l'utilisation d'hypolipémiants un an avant la date d'entrée dans la cohorte jusqu'à la date index.

Les patients diagnostiqués avec un diabète ou une dyslipidémie un an avant la date index étaient considérés comme nouvellement diagnostiqués. Pour les autres patients, le niveau d'adhésion aux antidiabétiques et aux hypolipidémiants a été mesuré 1 an avant la date index et il a été catégorisé en 2 groupes : adhérents à $> 80\%$ des doses prescrites et adhérents à $< 80\%$ des doses prescrites. Le score de maladie chronique (CDS) a été mesuré pour les cas et les témoins un an avant la date index. Ce score est pondéré en fonction du nombre de maladies traitées et de leur sévérité. Enfin, toute MC (autre que l'ACV) qui se serait développée entre la date d'entrée dans la cohorte et la date index a été considérée comme variable potentiellement confondante.

Pour l'étude de cohorte tous les sujets ayant développé un premier événement cérébrovasculaire (cas) pendant la période de suivi ont été identifiés. Il est à noter que la date index est considérée dans ce cas comme la date de l'événement pour les cas et comme celle de censure pour les autres.

Le sexe des patients et le statut socio-économique ont été identifiés à la date d'entrée dans la cohorte, comme pour le devis cas-témoins. Le diabète et la dyslipidémie considérées comme des variables dichotomiques (oui/non) ont été identifiées à partir des codes diagnostiques ou de l'utilisation de médicaments un an avant la date d'entrée dans la cohorte jusqu'à la date index. Les patients diagnostiqués avec un diabète ou une dyslipidémie un an avant la date index étaient considérés comme nouvellement diagnostiqués. Pour les autres patients, le niveau d'adhésion aux antidiabétiques et aux hypolipidémiants a été mesuré 1 an avant la date index et il a été catégorisé en 2 groupes : adhérents à $> 80\%$ des doses prescrites et adhérents à $< 80\%$ des doses prescrites. Le score de maladie chronique (CDS) a été mesuré pour les cas et les témoins un an avant la date index. Enfin, toute MC (autre que l'ACV) qui se serait développée entre la date d'entrée dans la cohorte et la date index a été consdérée comme variable potentiellement confondante.

Même si en apparences les définitions sont pareils, la façon de définir les dates index fait en sorte que les variables sont différentes.

Annexe C

CODE R

```
#####################################################
# Data1 est la matrice des données qui contient au moins les variables suivantes:
# indice, temps d'événement, temps de censure (s), événement (1 ou 0), variable
# fixe en temps (var_fixe), T1 et l'historique du traitement (vecteur de longueur
# 60, une valeur à chaque 2 mois).
# Nous comptons le nombre de lignes nécessaires pour la nouvelle matrice,
# en calculant une ligne chaque fois que le temps d'événement d'un sujet
# est plus petit que 2*(j+1) mois, pour j=0, 1, ..., 59.
#####################################################

for(i in 1:nrow(data1)){
for(j in 0:59){
if(data1$temps[i]<2*(j+1)) data1[i,(9+j)]<-NA}}
n4<-sum(!is.na(data1[,8:67]))

#####################################################
# Nous créons la nouvelle matrice data2 telle que la variable "Start"
# prend des valeurs entre 0 et 59, la variable "Stop" est "Start"+1
# et les autres variables sont copiées telles qu'elles (inclusivement
# la valeur pour le  "traitement" au temps j, pour j allant de 1:60);
#####################################################

data2<-matrix(0,n4,10)
```

```
colnames(data2) <- c("Start", "Stop", names(data1)[1:7], "TDvar")
row=0
for(i in 1:nrow(data1)){
    for(j in 8:67){
if (is.na(data1[i, j])) next
else{
     row=row+1
     Start<-j-8
     Stop<-Start+1
     data2[row,]<-c(Start, Stop, unlist(data1[i, c(1:7, j)]))}}}
data2 <- as.data.frame(data2)

##################################################
# Nous enlevons les variables qui ne sont pas nécessaires par la suite.
# Si nous supposons que nous avons une valeur pour la variable qui dépend
# du temps, à chaque 2 mois (comme dans l'exemple présenté dans la cadre
# de ce mémoire),les variables "Start" et "Stop" sont multipliées par 2.
##################################################
data2<-data2[,-c(4,6,8)]
data2$Start<-2*data2$Start
data2$Stop<-2*data2$Stop

##################################################
# Définition de la variable "event1" qui prend la même valeur que dans la
# matrice data1 sur la dernière ligne pour chaque sujet et 0 ailleurs;
# ensuite nous remplaçons l'ancienne variable "event" avec "event1";
##################################################

n1<-dim(data2)[1]
event1<-rep(0,n1)
for(i in 1:n1){
if (data2$Stop[i]>=data2$temps[i])event1[i]<-data2$event[i]
else event1[i]<-0}
data2<-data2[,-5]
```

```
data2<-data.frame(data2,event1)

################################################
# Si nous avons besoin d'une valeur à chaque mois le code suivant
# transforme la matrice de façon que les valeurs pour toutes les variables
# sont gardées constantes sauf pour "Start" et "Stop";
################################################

n2<-2*(dim(data2)[1])
Id<-rep(0,n2)
Start<-rep(0,n2)
Stop<-rep(0,n2)
Event<-rep(0,n2)
TDvar<-rep(0,n2)
var_fixe<-rep(0,n2)
row=0
for(i in 1:n1){
row<-row+1
Id[row]<-data2$indice[i]
Start[row]<-data2$Start[i]
Stop[row]<-data2$Start[i]+1
Event[row]<-0
TDvar[row]<-data2$TD_var[i]
cov1[row]<-data2$var_fixe[i]
row<-row+1
Id[row]<-data2$indice[i]
Start[row]<-data2$Start[i]+1
Stop[row]<-data2$Stop[i]
Event[row]<-data2$event1[i]
TDvar[row]<-data.final2$TDvar[i]
var_fixe[row]<-data.final2$var_fixe[i]}
data3<-data.frame(Id,Start,Stop,Event,TDvar,var_fixe)
```

BIBLIOGRAPHIE

Abrahamowicz, M., Bartlett, G., Tamblyn, R., Berger, R. (2006). Modeling cumulative dose and duration provided insights regarding the association between benzodiazepines and injuries, *Journal of Clinical Epidemiology*, **59**, 393–403.

Abrahamowicz, M., Ciampi, A., Ramsay, J.O. (1992). Nonparametric density estimation for censored survival data : Regression-spline approach, *The Canadian Journal of Statistics*, **20**, 171–185.

The ALLHAT Officers and Coordinators for the ALLHAT Collaborative research Group (2002). Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretics. The Antihypertensive and Lipid-Lowering treatment to Prevent Heart Attack Trial (ALLHAT), *Journal of the American Medical Association*, **288**,2981–2997.

Barnard, G.A. (1963). New methods of quality control, *Journal of the Royal Statistical Society*, **A 126**, 255–258.

Bates, J.M., Granger, C.W.J. (1969). The combination of forecasts, *Operational Research Quarterly*, **20**, 451–468.

Blood Pressure Lowering Treatment Trialists' Collaboration (2003). Effects of different blood-pressure-lowering drugs on major cardiovascular events : results of prospectively-designed overviews of randomized trials, *Lancet*, **362**, 1527–1535.

Breslow, N. E. (1972). Discussion following « Regression models and life tables » by D. R. Cox, *Journal of the Royal Statistical Society*, **B 34**, 187–220.

Caro, J.J., Salas M., Speckman, J.L., Raggio, G., Ja, J.D. (1999). Persistence with treatment for hypertension in actual practice, *Canadian Medical Association Journal*, **160**, 31–37.

Collins, R., MacMahon, S. (1994). Blood pressure, antihypertensive drug treatment and the risks of stroke and of coronary heart disease, *British Medical Bulletin*, **50**, 272–298.

Cox, M. G. (1972). The numerical evaluation of B-splines, *Journal of the Institute of Mathematics and Its Applications*, **10**, 134–149.

Curry, H.B., Schoenberg, I.J. (1966). On Polya frequency functions.IV.The fundamental spline functions and their limits, *Journal d'analyse mathémathique*, vol **17**.

De Boor, C. (1978). *A practical guide to spline*, **Springer-Verlag**.

De Boor , C. (1972). On calculating with B-splines, *Journal of Approximation Theory*, **6**, 50–62.

DiMatteo, M.R., Giordani, P.J., Lepper, H.S., Croghan, T.W. (2002). Patient adherence and medical treatment outcomes : a meta-analysis, *Medical Care*, vol **40**, 794–811.

Durrleman, S., Simon, R. (1989). Flexible regression models with cubic splines, *Statistics in Medicine*, **8**, 551–561.

Elliott, W.J., Plauschinat, C.A., Skrepnek, G.H., Gause, D. (2007). Persistence, Adherence, and Risk of Discontinuation Associated with Commonly Prescribed Antihypertensive Drug Monotherapies, *Journal of the American Board of Family Medicine*, **20**, 72–80.

Essebag, V., Platt, R.W., Abrahamowicz, M., Pilote, L. (2005).Comparison of nested case-control and survival analysis methodologies for analysis of time-dependent exposure, *BioMedCentral Medical Research Methodology*, **5**, 5–10

Eubank, R.L. (1988). *Spline smoothing and nonparametric regression*, **New York : M. Dekker**.

Ferguson, J.C.(1964), Multi-variable curve interpolation, *Journal of the Association for Computing Machinery* , **11**, 221–228.

Goldstein, L.B., Adams R., Alberts M.J., Appel L.J., Brass L.M., Bushnell C.D.*et al.*(2006). Primary prevention of ischemic stroke : a guideline from the American Heart Association/American Stroke Association Stroke Council : cosponsored by the Atherosclerotic Peripheral Vascular Disease Interdisciplinary Working Group ; Cardiovascular Nursing Council ; Clinical Cardiology Council ; Nutrition, Physical Activity, and Metabolism Council ; and the Quality of Care and Outcomes Research Interdisciplinary Working Group, *Circulation*, **113**, 873–923.

Goldstein, L.B., Bushnell, C.D., Adams, R.J., Appel, L.J., Braun, L.T.,Chaturvedi, S., Creager, M.A., Culebras, A., Eckel, R.H., Hart, R.G., Hinchey, J.A., Howard, V.J., Jauch, E.C., Levine, S.R., Meschia, J.F., Moore, W.S., Nixon, J.V., Pearson, T.A. ; on behalf of the American Heart Association Stroke Council, Council on Cardiovascular Nursing, Council on Epidemiology and Prevention, Council for High Blood Pressure Research, Council on Peripheral Vascular Disease, and Interdisciplinary Council on Quality of Care and Outcomes Research(2011). Guidelines for the primary prevention of stroke : a guideline for healthcare professionals from the American Heart Association/American Stroke Association, *Stroke*, **42**. 517–584.

Hauptmann, M., Wellmann, J., Lubin, J.H., Rosenberg, P.S., Kreienbrock, L.(2000). The analysis of exposure-time-response relationships using a spline weight function, *Biometrics*, **56**, 1105–1108.

He, X. et Shi, P. (1998). Monotone B-spline smoothing, *Journal of the American Statistical Association*, **93**, 643–650.

Hess, K.R. (1994). Assessing time-by covariate interactions in proportional hazard regression models using cubic spline functions, *Statistics in Medicine*, **13**, 1045–1062.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T. (1999), Bayesian model averaging : A tutorial(with discussion), *Statistical Science*, vol **14**, 382–417.

HOPE (Heart Outcomes Prevention Evaluation) Study Investigators (2000). Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high risk patients, *The New England Journal of Medicine*, **342**, 145–153.

Joffres, M.R., Hamet, P., MacLean, D.R., L'italien, G.J., Fodor, G. (2001). Distribution of blood pressure and hypertension in Canada and the United States, *American Journal of Hypertension*, **14**, 1099–1105.

Kettani, F.Z., Dragomir, A., Côté, R., Roy, L., Bérard, A., Blais, L., Lalonde, L., Moreau, P., Perreault, S. (2009). Impact of a better adherence to antihypertensive agents on cerebrovascular disease for primary prevention, *Stroke*, **40**, 213–220.

Langholz, B., Richardson, D. (2009). Are nested case-control studies biased ?, *Epidemiology*, **20**, 321–329.

Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*, **Wiley**.

Lawless,J . F., Singhal,K . (1978). Efficients screening of nonnormal regression models, *Biometrics*, **34**, 318–327.

Leamer, E.E. (1978). *Specification Searches*, **Wiley**.

Lee, P. (1989). Bayesian statistics : An introduction, **Oxford Universitry Press**.

MacKenzie, T., Abrahamowicz, M. (2002). Marginal and hazard ratio specific random data generation : applications to semi-parametric bootstrapping, *Statistics and Computing*, **12**, 245–252.

Madigan, D., York, J. (1995). Bayesian graphical models for discrete data, *International Statistical Review*, **63**, 215–232.

Madigan, D. et Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's Window, *Journal of the American Statistical Association*, vol **89**, 1335–1346.

Marsden, M.J., Schoenberg, I.J. (1966). On variation diminishing spline approximation methods, *Mathematica*, **31**, 61–82.

Newbold, P., Granger, C.W.J.(1974). Experience with forecasting univariate time series and the combination of forecasts, *Journal of the Royal Statistical Society*, **A 137**, 131–232.

Perreault, S., Lamarre, D., Blais, L., Dragomir, A., Berbiche, D., Lalonde, L., Laurier, C., St-Maurice, F., Collin, J.(2005). Persistence and determinants with antihypertensive agents in middle-aged newly treated patients, *The Annals of Pharmacotherapy*, **39**, 1401–1408.

Pitt, B., Byington, R., Furberg, C., Hunninghake, D.B., Mancini, J.G.B., Miller, M.E., Riley, W.(2000). Effects of amlodipine on the progression of atherosclerosis and the occurrence of clinical events, *Circulation*, **102**, 1503–1510.

Prentice, R.L., Breslow, N.E. (1978). Retrospectives sudies and failure time models, *Biometrika*, **65**, 153–158.

Raiffa H., Schlaifer, R.(1961). *Applied statistical decision theory*, **Boston, Division of Research, Graduate School of Business Administration, Harvard University** .

Raftery, A.E., Madigan, D., Hoeting, J.A. (1997). Bayesian model averaging for linear regression models, *Journal of the American Statistical Association*, **92**, 179–191.

Ramsay J.O. (1988). Monotone regression splines in action, *Statistical Science*, **3**, 425–461.

Roberts, H.V. (1965). Probabilistic prediction, *Journal of the American Statistical Association*, **60**, 50–62.

Schoenberg, I.J. (1946). Contributions to the problem of approximation of equidistant data by analytic functions, *Quarterly of Applied Matematics*, **4**, 45–99.

Staessen J.A., Wang, J.G., Birkenhäger, W.H., Fagard. R. (1999). Treatment with beta-blockers for the primary prevention of the cardiovascular complications of hypertension, *European Heart Journal*, **20**, 11–25.

Statistique Canada (2012). CANSIM Tableau 102-0529 : Deaths, by cause, Chapter IX : Diseases of the circulatory system (I00 à I99), age group and sex, Canada, 2005 à 2009. Publié le 31 mai 2012.

Stone, C.J. (1986). Generalized additive models : Comment on Hastie and Tibshirani, *Statistical Science*, **1**, 312–314.

Sylvestre, M.P., Abrahamowicz, M. (2009). Flexible modeling of the cumulative effects of time-dependent exposures on the hazard, *Statistics in Medicine*, **28**, 3437–3453.

Thom, T., Haase, N., Rosamond, W. *et al.* (2006). Heart disease and stroke statistics-2006 update : a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee, *Circulation*, **113**, 85–151.

Thomas D.C.(1988). Models for exposure-time-response relationships with applications to cancer epidemiology, *Annual Reviews of Public Health*, *9*, 451–482.

Volinsky C.T., Madigan D., Raftery A.E., Kronmal M.A.(1997). Bayesian model averaging in proportional hazard models : Assessing the risk of a stroke, *Applied Statistics*, **46**, 433–448.

Volinsky C.T.(1997). Bayesian model averaging for censored survival models, *University of Washington Statistics Department Ph.D. Dissertation.*

Wacholder, S. (2009). Bias in full cohort and nested case-control studies, *Epidemiology*, **20**, 339–340.