Estimation du risque attribuable et de la fraction préventive dans les études de cohorte

Malamine Gassama

NNT : 2016SACLV131

THESE DE DOCTORAT
DE
L'UNIVERSITE PARIS-SACLAY
PREPAREE A
L'UNIVERSITE VERSAILLES SAINT-QUENTIN EN YVELINES

ÉCOLE DOCTORALE N°570
SANTÉ PUBLIQUE

Spécialité de doctorat : Biostatistiques

Par

M. Malamine GASSAMA

Estimation du risque attribuable et de la fraction préventive dans les études de cohorte

Thèse présentée et soutenue à l'Institut Pasteur de Paris, le 9 décembre 2016 :

Composition du Jury :

M. Jean BOUYER, Directeur de Recherche, Inserm, Université Paris-Sud, Président
Mme Catherine HUBER-CAROL, Professeur, Université Paris Descartes, Rapporteur
M. Vivian VIALLON, Maître de Conférences – HDR, Université Claude Bernard - Lyon 1, Rapporteur
Mme Mounia HOCINE, Maître de Conférences, Conservatoire National des Arts et Métiers, Examinatrice
Mme Anne THIÉBAUT, Chargée de Recherche, Inserm, UVSQ, Directrice de thèse
M. Jacques BÉNICHOU, Professeur, Université de Rouen, Directeur de thèse

Remerciements

Mes premiers remerciements s'adressent naturellement à mes deux directeurs de thèse Anne Thiébaut et Jacques Bénichou. Ils sont à l'origine de mon intérêt pour les sujets traités dans cette thèse autour du risque attribuable et de la fraction préventive, ils ont dirigé, encadré, accompagné de très près chaque étape de ces trois années de thèse. Je tiens par ces lignes à leur exprimer toute ma gratitude. Merci à vous de m'avoir fait confiance et de m'avoir donné la possibilité de réaliser ce travail de thèse.

Je remercie également les membres du Jury d'avoir accepté d'évaluer mon travail : Merci à Jean Bouyer d'avoir accepté de présider mon Jury, merci à Catherine Huber-Carol et à Vivian Viallon de m'avoir fait l'honneur d'accepter d'être mes rapporteurs, merci à Philippe Saint Pierre et à Mounia Hocine de m'avoir fait l'honneur d'accepter d'être mes examinateurs et de l'intérêt que vous portez à mon travail. J'ai eu la chance de suivre le cours sur l'analyse de survie à l'UPMC et le cours de statistique mathématique à Paris Descartes de Philippe Saint Pierre et de Catherine Huber-Carol respectivement.

Je remercie également Didier Guillemot de m'avoir accueilli dans son unité de recherche à l'Institut Pasteur pour mon stage de Master 2 mais aussi pour mes trois années de thèse.

Je remercie également l'Université Versailles Saint Quentin en Yvelines de m'avoir financé pendant ces trois années de thèse. Je souhaite également remercier Audrey Bourgeois pour son soutien, ses orientations durant ces trois années de thèse, ton aide m'a vraiment été précieuse dans mes démarches administratives au niveau de l'UVSQ mais aussi au niveau de la préfecture. Je remercie également l'agence nationale de sécurité du médicament et des produits de santé (ANSM) qui a subventionné notre projet « Estimation du risque attribuable en pharmaco-épidémiologie et dans le contexte de la résistance bactérienne », ce qui a permis ma participation à des congrès ainsi que le financement de mes trois derniers mois de thèse.

Je remercie les investigateurs des cohortes E3N et 3C qui m'ont permis l'accès à leurs données. En particulier Agnès Fournier, Marie-Christine Boutron-Ruault, Pascale

Gerbouin-Rérolle et Wilna Tello pour E3N, Christophe Tzourio, Annick Alpérovitch et Catherine Helmer pour 3C.

Je remercie vivement Pascale Tubert-Bitter d'avoir accepté de relire mon premier travail et de m'avoir fait part de ses précieux commentaires. Merci également à Mohammed Sedki pour ton aide précieuse sur le logiciel `R` mais aussi dans la partie théorique des probabilités et statistiques. Merci à Michael et Maya pour vos différentes corrections sur mon anglais.

Merci à tous les membres du B2PHI : Annick, Lulla, Didier, Elisabeth, Bich-Tram, Anne T, Laurence, Lénaig, Thomas, Anne F, Matthieu, Anna, Felix, Hélène, Michael, Jeanne, Laura, Juliette, Ismail de m'avoir offert l'occasion de me rendre au Sénégal pour voir ma mère qui était malade. Votre geste m'a profondément touché et je vous en serai toujours reconnaissant.

Merci à Thomas pour ton aide qui m'a été vraiment précieuse dans mes recherches de logement et les bons moments que nous avons passés ensemble à courir dans le parc Georges-Brassens. Merci à Matthieu de m'avoir proposé de continuer le sport après le départ de Thomas, j'ai particulièrement adoré les exercices dans le Champ de Mars. Merci à mes collègues de bureau et ancien collègues Hélène, Michael, Maya, Xuefeng et Clotilde (ancien thésard) pour votre bonne humeur et la bonne ambiance dans le bureau des thésards. Je remercie également à Adrien (ancien stagiaire d'Anne) pour les différentes échanges sur le risque attribuable.

Je tiens également à remercier l'équipe Tutorat de l'Institut Pasteur et mon tuteur Mathieu Picardeau pour son écoute et ses orientations ainsi que Hugo Varet et Stevenn Volant du C3BI pour votre accueil et votre aide sur le logiciel R.

Je remercie également mes ami(e)s et copain(e)s : Bachir, Moussa, Alhassane, Birane, Iba, Jean, Félémou, Alpha Boubou et sa femme, Idy, Tamsir, Regis, Alaya, Manthita, Maryam, Couwaly, Fatou Bintou, etc.

Un grand merci aux doctorants avec qui j'ai partagé quelques repas de midi et quelques retrouvailles autour du Franglish qui ont rendu ces années si agréables : Jean, Pauline, Catherine, Manik, Lucie, Eve, etc.

Merci à la famille Dieng : Momar, Mounasse et Amina. Merci à mes grands Mbagnick, Mor, Mbaye et Papa Moussa pour les différentes fêtes célébrées ensemble. Je tiens également à remercier la famille FAYE qui représentait ma famille d'accueil à Dakar. Merci à mon ancien professeur de Physique-Chimie et de Mathématiques, Ibou Guéye, pour ses conseils, orientations et aux différentes intéressantes discussions au téléphone, tu es pour moi un ami, un grand, un exemple pour ma génération.

Une pensée pour mes parents, mes frères et ma sœur pour votre amour, votre affection, votre soutien et vos prières. Merci à ma belle-famille également pour votre confiance, vos prières, votre affection et votre soutien. Merci à ma famille en France : Oncle Abdoulaye qui m'a accueilli et soutenu durant ma première année à Clermont, Pa Sekou qui tient beaucoup à moi, mes cousins Abasse, Alassane et Mouhamet pour leur accueil durant mon Master 2 à Paris 6. Je remercie également Edwidge et Guillaume Kalonji et leurs enfants (ma deuxième famille en France) pour vos soutiens, orientations et encouragements. Merci pour l'attention que vous portez à ma famille.

Enfin, je remercie ma femme Mame Coumba de m'avoir supporté, encouragé, aimé, soutenu sans faille durant toutes ces années. Ta présence à mes côtés m'a beaucoup aidé à surmonter les difficultés de la vie. Merci pour le beau cadeau que tu m'as offert dont les sourires m'encourageaient énormément tous les jours.

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique

Titre : Estimation du risque attribuable et de la fraction préventive dans les études de cohorte

Mots clés : risque attribuable; fraction préventive; étude de simulation; Kaplan-Meier pondéré; modèle à risques constants par morceaux; modèle de Cox; étude de cohorte; cancer du sein; accident vasculaire cérébral; statine; traitement hormonal de la ménopause

Résumé : Le risque attribuable (RA) mesure la proportion de cas de maladie qui peuvent être attribués à une exposition au niveau de la population. Plusieurs définitions et méthodes d'estimation du RA ont été proposées pour des données de survie. En utilisant des simulations, nous comparons quatre méthodes d'estimation du RA dans le contexte de l'analyse de survie : deux méthodes non paramétriques basées sur l'estimateur de Kaplan-Meier, une méthode semi-paramétrique basée sur le modèle de Cox à risques proportionnels et une méthode paramétrique basée sur un modèle à risques proportionnels avec un risque de base constant par morceaux. Nos travaux suggèrent d'utiliser les approches semi-paramétrique et paramétrique pour l'estimation du RA lorsque l'hypothèse des risques proportionnels est vérifiée. Nous appliquons nos méthodes aux données de la cohorte E3N pour estimer la proportion de cas de cancer du sein invasif attribuables à l'utilisation de traitements hormonaux de la ménopause (THM). Nous estimons qu'environ 9 % des cas de cancer du sein sont attribuables à l'utilisation des THM à l'inclusion.

Dans le cas d'une exposition protectrice, une alternative au RA est la fraction préventive (FP) qui mesure la proportion de cas de maladie évités. Cette mesure n'a pas été considérée dans le contexte de l'analyse de survie. Nous proposons une définition de la FP dans ce contexte et des méthodes d'estimation en utilisant des approches semi-paramétrique et paramétrique avec une extension permettant de prendre en compte les risques concurrents. L'application aux données de la cohorte des Trois Cités (3C) estime qu'environ 9 % de cas d'accident vasculaire cérébral peuvent être évités chez les personnes âgées par l'utilisation des hypolipémiants. Notre étude montre que la FP peut être utilisée pour évaluer l'impact des médicaments bénéfiques dans les études de cohorte tout en tenant compte des facteurs de confusion potentiels et des risques concurrents.

Title: Estimation of attributable risk and prevented fraction in cohort studies

Keywords: attributable risk; prevented fraction; simulation study; weighted Kaplan-Meier estimator; piecewise constant hazards model; Cox model; cohort studies; breast cancer; stroke; statins; menopausal hormone therapy

Abstract: The attributable risk (AR) measures the proportion of disease cases that can be attributed to an exposure in the population. Several definitions and estimation methods have been proposed for survival data. Using simulations, we compared four methods for estimating AR defined in terms of survival functions: two nonparametric methods based on Kaplan-Meier's estimator, one semiparametric based on Cox's model, and one parametric based on the piecewise constant hazards model. Our results suggest to use the semiparametric or parametric approaches to estimate AR if the proportional hazards assumption appears appropriate. These methods were applied to the E3N women cohort data to estimate the AR of breast cancer due to menopausal hormone therapy (MHT). We showed that about 9% of cases of breast cancer were attributable to MHT use at baseline.

In case of a protective exposure, an alternative to the AR is the prevented fraction (PF) which measures the proportion of disease cases that could be avoided in the presence of a protective exposure in the population. The definition and estimation of PF have never been considered for cohort studies in the survival analysis context. We defined the PF in cohort studies with survival data and proposed two estimation methods: a semiparametric method based on Cox's proportional hazards model and a parametric method based on a piecewise constant hazards model with an extension to competing risks. Using data of the Three-City (3C) cohort study, we found that approximately 9% of cases of stroke could be avoided using lipid-lowering drugs (statins or fibrates) in the elderly population. Our study shows that the PF can be estimated to evaluate the impact of beneficial drugs in observational cohort studies while taking potential confounding factors and competing risks into account.

Valorisation scientifique

Articles

M. Gassama, J. Bénichou, L. Dartois, A.C.M. Thiébaut. *Comparison of methods for estimating the attributable risk in the context of survival analysis*, en révision (Annexe D)

M. Gassama, A.C.M. Thiébaut, C. Tzourio, J. Bénichou. *Use of the prevented fraction to estimate the proportion of stroke cases that could be avoided by using lipid lowering drugs in the French Three-City cohort*, soumis (Annexe E)

Communications orales

M. Gassama, A.C.M. Thiébaut, J. Bénichou, "Use of the prevented fraction to estimate the proportion of stroke cases that can be avoided by using lipid lowering drugs in the French Three-City cohort" 37th Annual Conference of the International Society for Clinical Biostatistics, 21-25 août 2016, Birmingham (Royaume-Uni).

M. Gassama, J. Bénichou, L. Dartois, A.C.M. Thiébaut, "Comparison of methods for estimating the attributable risk in the context of survival analysis", Journées GDR & SFB, 27-28 juin 2016, Lyon.

Communication affichée

M. Gassama, J. Bénichou, A.C.M. Thiébaut, "Comparison of methods for estimating attributable risk in the context of survival analysis". 36th Annual Conference of the International Society for Clinical Biostatistics, 23-27 août 2015, Utrecht (Pays-Bas).

Activités d'enseignement

Contrat doctoral - mission d'enseignement à l'Université Versailles Saint Quentin en Yvelines (2013-2015) : Cours et travaux dirigés de Mathématiques Générales en première année de Licence, UFR des Sciences, Département de Mathématiques

Contrat doctoral - mission d'enseignement à l'Université Versailles Saint Quentin en Yvelines (2015-2016) : Cours et travaux pratiques de Biostatistique en troisième année de Licence, UFR des Sciences, Département de Biologie

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

AS : Survie attribuable (pour *attributable survival*)

AST : Temps de survie attribuable (pour *attributable survival time*)

ATC : Classification anatomique, thérapeutique et chimique

AVC : Accident vasculaire cérébral

CIM : Classification internationale des maladies

E3N : Étude Épidémiologique auprès de femmes de la MGEN (Mutuelle Générale de l'Éducation Nationale)

FP : Fraction préventive

HDL : Lipoprotéine de haute densité (pour *high density lipoprotein*)

HR : *Hazard ratio*

IC 95 % : Intervalle de confiance à 95 %

IMC : Indice de masse corporelle

KM : Kaplan-Meier

KMP : Kaplan-Meier pondéré

LDL : Lipoprotéine de basse densité (pour *low density lipoprotein*)

OMS : Organisation mondiale de la santé

OR : *Odds ratio*

PC : Probabilité de couverture

RA : Risque attribuable

RCM : Risque de base constant par morceaux

RR : Risque relatif

SEE : Écart-type estimé moyen (pour *sampling mean of standard error estimate*)

SSD : Écart-type empirique (pour *sampling standard deviation*)

THM : Traitements hormonaux de la ménopause

Chapitre 1

Introduction

1.1 Problématique générale

En épidémiologie, la force d'une association entre l'exposition à un facteur de risque et l'apparition d'une maladie est souvent estimée par un risque relatif (RR) ou un *odds ratio* (OR). Ces paramètres ne prennent pas en compte la prévalence de l'exposition. Le risque attribuable (RA) en revanche prend en compte non seulement la force du lien entre une exposition et la maladie mais aussi l'importance (la prévalence) de l'exposition dans la population dans le cas où cette dernière est non protectrice (RR $>$ 1). Le RA a été introduit par Levin en 1953 [1] et exprime la proportion de cas de maladie attribuables à une exposition, c'est-à-dire, sous certaines conditions, la proportion de cas potentiellement évitables si l'exposition était supprimée.

De nombreuses méthodes d'estimation du RA ont été développées, permettant l'ajustement sur des facteurs de confusion et la prise en compte d'interactions, principalement pour les études transversales et cas-témoins [2].

Dans les études de cohorte, lorsque l'on s'intéresse au délai de survenue d'un événement ou d'un décès, les participants sont suivis au cours du temps et l'information sur l'exposition est recueillie à l'inclusion et souvent au cours du suivi également. Au cours du suivi, de nouveaux cas de l'événement d'intérêt sont identifiés. L'estimation du RA n'a de sens que si l'exposition est un agent causal de l'événement d'intérêt. Par conséquent,

elle peut être considérée comme plus réaliste dans les études de cohorte et moins dans les études transversales. Cependant, le RA a été le plus souvent estimé dans les études transversales et cas-témoins et moins dans les études de cohorte, où la durée du suivi et la censure doivent être prises en compte en s'appuyant sur l'analyse de survie [3–7]. Or la littérature statistique et épidémiologique est récente et encore peu consensuelle concernant les définitions et méthodes d'estimation du RA dans le contexte de l'analyse de survie [3–8].

Dans le cas d'une exposition protectrice associée à un RR inférieur à 1, une alternative au RA est la fraction préventive (FP). Cette quantité a été introduite par Miettinen en 1974 [9] et mesure la proportion de cas de maladie potentiellement évités en présence d'une exposition protectrice. D'utilisation moins courante que le RA, la FP a cependant été utilisée dans le cadre d'études de cohorte, cas-témoins et transversales [9–12]. Dans le contexte de l'analyse de survie, aucune définition ni méthode d'estimation de la FP n'a été considérée.

1.2 Objectifs de la thèse

Les objectifs de la thèse sont de proposer et d'évaluer des définitions et méthodes d'estimation du RA et de la FP dans les études de cohorte. Pour ce faire, nous allons dans un premier temps faire une revue de la littérature sur les définitions et méthodes d'estimation du RA dans le cadre de la survie. Ensuite, nous comparerons ces méthodes dans une étude de simulation avec une application à des données réelles. Enfin, nous définirons la FP dans le contexte de l'analyse de survie et nous en présenterons une application à des données réelles.

Chapitre 2

Revue de la littérature

2.1 Risque attribuable dans le contexte général

2.1.1 Définition générale du risque attribuable

Pour une exposition binaire (exposé E vs non exposé $\bar{E}$), le RA est défini selon l'équation suivante [1] :

$$RA = \frac{\mathbf{P}(D) - \mathbf{P}(D|\bar{E})}{\mathbf{P}(D)} \tag{2.1}$$

où $\mathbf{P}(D)$ désigne la probabilité de la maladie (incidence) dans la population, qui comprend des exposés E et des non exposés $\bar{E}$, et $\mathbf{P}(D|\bar{E})$ est la probabilité hypothétique de la maladie dans la même population après élimination de l'exposition. Le RA mesure la proportion de cas de maladie potentiellement causés par l'exposition [10].

L'interprétation de l'estimation du RA nécessite la vérification de trois conditions [10]. L'estimation doit être non biaisée, l'exposition doit être de nature causale pour la maladie et l'élimination de l'exposition ne doit pas modifier la distribution des autres facteurs de risque (ou protecteurs) dans la population. Il est souvent difficile en épidémiologie d'affirmer la nature causale d'une relation entre une exposition et une maladie. Il est nécessaire de montrer l'existence de l'association entre l'exposition et la maladie mais cette dernière n'est pas suffisante pour conclure à l'existence d'une relation causale. Cette causalité entre l'exposition et la maladie a été discutée par Greenland et Robins [13],

Robins et Greenland [14] ainsi que Rothman et Greenland [15] et ces auteurs proposent d'estimer la fraction étiologique qui ne prend en compte que les cas de maladie pour lesquels l'exposition a joué un rôle dans la survenue de la maladie. Ils ont prouvé que la fraction étiologique est différente du RA et plus pertinente pour mesurer l'impact réel de l'exposition sur la survenue de la maladie. En pratique, il est impossible d'identifier, parmi les sujets exposés ceux pour lesquels l'exposition a joué un rôle étiologique. Ainsi, la fraction étiologique n'est pas calculable en pratique [14, 16–18] et par conséquent le RA est la mesure la plus adaptée pour estimer l'impact d'une exposition en santé publique.

On trouve de nombreux synonymes dans la littérature : risque attribuable dans la population [3], fraction étiologique [9], pourcentage de risque attribuable [19], fraction attribuable [4, 13, 20, 21], etc.

Le RA dépend à la fois de la force de l'association entre l'exposition et la maladie et de la prévalence de l'exposition dans la population; il peut s'écrire en effet [1, 19] :

$$RA = \frac{p_E(RR-1)}{1+p_E(RR-1)} \tag{2.2}$$

où $p_E = \mathbf{P}(E)$ est la prévalence de l'exposition dans la population considérée et $RR = \mathbf{P}(D|E)/\mathbf{P}(D|\bar{E})$ est le risque relatif.

On peut aussi exprimer le RA à partir de la prévalence de l'exposition dans la population malade, notée $p_{E|D}$ [9]:

$$RA = \frac{p_{E|D}(RR-1)}{RR}. \tag{2.3}$$

2.1.2 Définition générale de la fraction attribuable chez les exposés

Si on se restreint aux sujets exposés, la fraction attribuable (FA_E) mesure la proportion de cas de maladie attribuable à l'exposition chez les sujets exposés [1, 9, 19, 22]. Elle est définie comme :

$$FA_E = \frac{\mathbf{P}(D|E) - \mathbf{P}(D|\bar{E})}{\mathbf{P}(D|E)}.$$

Cette quantité est aussi appelée RA chez les exposés et est fonction du RR uniquement [1, 23] :

$$FA_E = \frac{RR - 1}{RR}.$$

2.1.3 Prise en compte d'une exposition à plusieurs niveaux ou d'un facteur d'ajustement

Une première généralisation de la définition du RA dans le cas d'une exposition à plusieurs niveaux a été proposée par Miettinen [9] :

$$RA = \sum_{s=1}^{S} \mathbf{P}(E_s|D)\frac{RR_s - 1}{RR_s} = 1 - \sum_{s=1}^{S} \frac{\mathbf{P}(E_s|D)}{RR_s} \tag{2.4}$$

où $s = 1, \ldots, S$ désigne le niveau d'exposition, $\mathbf{P}(E_s|D)$ est la prévalence de l'exposition de niveau s dans la population malade et RR_s est le RR associé par rapport à l'exposition de référence choisie.

Une généralisation de la définition du RA basée sur la définition proposée par Levin est souvent utilisée [10] :

$$RA = \frac{\sum_{s=1}^{S} \mathbf{P}(E_s)(RR_s - 1)}{1 + \sum_{s=1}^{S} \mathbf{P}(E_s)(RR_s - 1)} \tag{2.5}$$

où $\mathbf{P}(E_s)$ est la prévalence de l'exposition de niveau s.

En présence d'un facteur d'ajustement C à J niveaux, les définitions précédentes peuvent être généralisées [24, 25] :

$$RA = \frac{\mathbf{P}(D) - \sum_{j=1}^{J} \mathbf{P}(C_j)\mathbf{P}(D|C_j, \bar{E})}{\mathbf{P}(D)} \tag{2.6}$$

où $\mathbf{P}(C_j)$, $j = 1, \ldots, J$ est la proportion de sujets de niveau j pour le facteur d'ajustement C et $\mathbf{P}(D|C_j, \bar{E})$ l'incidence de la maladie chez les non exposés de niveau j.

2.2 Estimation du risque attribuable

2.2.1 Estimation du risque attribuable dans les études cas-témoins

Dans les études cas-témoins et avec une exposition binaire, un estimateur du RA est obtenu en utilisant les équations 2.2 ou 2.3 [2] :

$$\widehat{RA} = \frac{n_1 m_0 - n_0 m_1}{m_0 n}$$

où n_0 et n_1 représentent respectivement le nombre de sujets non exposés et exposés chez les cas ($n = n_0 + n_1$) et m_0 et m_1 le nombre de sujets non exposés et exposés chez les témoins. La variance peut être obtenue en utilisant la delta méthode [26] et en considérant des distributions appropriées des quantités n_1 et m_1. Par exemple, dans le cas d'un tirage au sort simple des témoins sans stratification ni assortissement par fréquence ou appariement, il s'agit de deux distributions binomiales indépendantes où les nombres totaux de cas et de témoins sont tous les deux fixes [10].

Différentes approches ont été proposées pour le calcul des intervalles de confiance (IC) du RA et peuvent s'appliquer à tous les schémas d'étude une fois les estimations ponctuelles du RA et de sa variance obtenues. Un IC standard du RA peut être construit en supposant une distribution asymptotique normale de l'estimateur du RA. Afin d'améliorer l'hypothèse de normalité et d'obtenir des probabilités de couverture plus conformes, Walter [27] suggère d'utiliser la transformation logarithmique, $\ln(1 - RA)$, Leung et Kupper [28] la transformation logistique, $\ln\left(RA/(1-RA)\right)$. Whittemore [29] a montré qu'une transformation logarithmique donne un IC plus large qu'un IC standard (sans transformation).

2.2.2 Estimation du risque attribuable dans les études transversales

Dans les études transversales, une estimation du RA peut être obtenue comme [2] :

$$\widehat{RA} = \frac{n_1 m_0 - n_0 m_1}{n(n_0 + m_0)}$$

où n_0 et n_1 représentent respectivement le nombre de sujets non exposés et exposés parmi les sujets malades ($n = n_0 + n_1$) et m_0 et m_1 le nombre de sujets non exposés et exposés parmi les sujets non malades ($m = m_0 + m_1$). Cet estimateur peut également être obtenu en remplaçant les quantités $\mathbf{P}(D)$ et $\mathbf{P}(D|\bar{E})$ par leurs estimateurs dans les équations 2.1, 2.2 et 2.3.

L'estimation de la variance peut être obtenue par delta méthode [26] en considérant une distribution multinomiale à quatre niveaux d'où proviennent les quantités n_0, n_1, m_0 et m_1 [10].

2.2.3 Estimation du risque attribuable dans les études de cohorte

Dans les études de cohorte, plusieurs approches ont été considérées. Avec un suivi fixe, les nombres n_0, n_1, m_0 et m_1 sont observés à la fin du suivi, ce qui est identique aux études transversales [2].

Lorsque les sujets sont suivis pendant une durée variable, le modèle multinomial ne s'applique pas et l'analyse doit reposer sur des modèles de survie comme le modèle de Cox [30] pour estimer les RR et en déduire une estimation unique du RA à l'échelle de la cohorte [31]. L'estimation d'un RA global est valable dans le cas où la maladie est rare. Dans le cas d'une maladie fréquente, l'estimation du RA global n'est plus adaptée. En effet, la probabilité de la maladie varie en fonction du temps et par conséquent le RA aussi. L'estimation du RA global peut donner des résultats biaisés dans le cas d'un suivi plus long. Des approches plus adaptées ont été proposées pour les études de cohorte et dans le contexte de l'analyse de survie. Plusieurs définitions et méthodes d'estimation ont été proposées, que nous détaillerons dans la suite, pour prendre en compte la censure et mieux estimer la probabilité de la maladie au cours du suivi. Ces approches sont plus adaptées et sont basées uniquement sur une interprétation de l'incidence de la maladie, dans la définition générale du RA, comme une fonction de répartition ou une fonction de risque instantané.

2.2.4 Estimation du risque attribuable ajusté

Plusieurs auteurs suggèrent d'utiliser l'équation 2.3 pour obtenir un RA ajusté [12, 32–34]. Le principe consiste à estimer les RR ajustés et la prévalence de l'exposition chez les personnes malades. L'approche de Mantel-Haenszel [23] pour le calcul des RA ajustés a beaucoup été utilisée dans les études transversales [15, 35–40] et dans les études cas-témoins [15, 35, 37, 39, 41–43] :

$$\widehat{RA} = \widehat{\mathbf{P}}(E|D)\frac{\widehat{OR}_{MH} - 1}{\widehat{OR}_{MH}} \tag{2.7}$$

où $\widehat{\mathbf{P}}(E|D)$ est la proportion estimée d'exposés chez les malades et $\widehat{OR}_{MH}$ l'OR estimé par l'approche de Mantel-Haenszel.

L'approche de la somme pondérée a aussi été proposée. Cette approche consiste à faire la somme pondérée de tous les RA calculés pour chaque niveau de facteur d'ajustement [10, 24, 29] :

$$\widehat{RA} = \sum_{j=1}^{J} \omega_j \widehat{RA}_j$$

où $\widehat{RA}_j$ et ω_j, $j = 1, \ldots, J$, représentent respectivement la valeur estimée du RA pour le niveau j et le poids correspondant. Le choix des pondérations est discuté dans la revue de Bénichou [2]. La définition des ω_j comme le nombre de personnes malades dans le niveau j donne un estimateur asymptotiquement sans biais du RA, qui correspond à l'estimateur du maximum de vraisemblance [29]. Une définition des pondérations utilisant l'inverse de la variance du RA dans le niveau j sur la somme des inverses des variances des RA de tous les niveaux a également été proposée [44].

Bruzzi *et al.* [25] proposent une définition du RA ajusté en utilisant un modèle de régression :

$$RA = 1 - \sum_i \sum_j \frac{\rho_{ij}}{RR_{i|j}}$$

où ρ_{ij} et $RR_{i|j}$ représentent respectivement la proportion de malades et le risque relatif de niveau d'exposition i, $i = 1, \ldots, I$ et de facteurs d'ajustement j, $j = 1, \ldots, J$. Un estimateur de cette quantité est obtenu en remplaçant $RR_{i|j}$ par son estimateur en utilisant une régression logistique, log-linéaire, Poisson ou Cox selon le type d'étude [2].

2.3 Fraction préventive dans le contexte général

2.3.1 Définition générale de la fraction préventive

La fraction préventive est définie selon l'équation suivante [9]:

$$FP = \frac{\mathbf{P}(D|\bar{E}) - \mathbf{P}(D)}{\mathbf{P}(D|\bar{E})}. \tag{2.8}$$

Elle mesure la proportion de cas de maladie potentiellement évités en présence d'une exposition protectrice [9]. La FP est à la fois fonction de la probabilité d'exposition et du RR et peut s'écrire :

$$FP = \mathbf{P}(E)(1 - RR). \tag{2.9}$$

La FP peut également être exprimée en fonction de la prévalence de l'exposition dans la population malade $\mathbf{P}(E|D)$ et du RR [12] :

$$FP = \frac{\mathbf{P}(E|D)(1 - RR)}{1 - (1 - RR)(1 - \mathbf{P}(E|D))}. \tag{2.10}$$

La FP et le RA sont mathématiquement interdépendants [10, 45]:

$$1 - FP = \frac{1}{1 - RA}. \tag{2.11}$$

2.3.2 Estimation de la fraction préventive

L'équation 2.11 nous montre que les questions d'estimation de la FP sont similaires à celles du RA. Un estimateur ajusté de la FP basé sur la méthode de Mantel-Haenszel [41] a été obtenu pour les études de cohorte, cas-témoins et transversales [12]. Un estimateur ajusté de la FP basé sur l'approche des sommes pondérées a été utilisé pour les études transversales [11]. Une expression de la variance a été proposée pour les études cas-témoins et transversales [11, 12, 45].

2.4 Contexte de l'analyse de survie

L'estimation du RA et de la FP nécessite l'utilisation de l'analyse de survie. Nous allons rappeler dans cette partie des notions de base de l'analyse qui nous seront utiles dans le cadre l'estimation du RA et de la FP dans les études de cohorte.

2.4.1 Définition des données de survie

Considérons la durée de survie ou le délai de survenue d'un événement T, variable aléatoire continue, positive ou nulle. Sa loi de probabilité peut être décrite par les fonctions de distribution usuelles :

— Densité de probabilité : Pour $t \geq 0$, la densité de probabilité représente la probabilité de présenter l'événement dans un petit intervalle de temps après l'instant t

$$f(t) = \lim_{\Delta \to 0} \frac{\mathbf{P}(t \leq T < t + \Delta)}{\Delta}.$$

— Fonction de répartition : La fonction de répartition représente la probabilité de présenter l'événement jusqu'au temps t, c'est-à-dire

$$F(t) = \mathbf{P}(T \leq t) = \int_0^t f(u)du.$$

D'autres fonctions de distribution spécifiques aux données de survie permettent également de décrire la loi de la variable aléatoire T :

— Fonction de survie : La fonction de survie est, pour $t \geq 0$, la probabilité de survivre jusqu'à l'instant t, c'est-à-dire

$$S(t) = P(T \geq t) = 1 - F(t).$$

— Fonction de risque instantané : Le risque instantané (ou taux de hasard) caractérise, pour t fixé, la probabilité de présenter l'événement dans un petit intervalle de temps après t, conditionnellement au fait d'avoir survécu jusqu'au temps t :

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

— Fonction de risque cumulé : La fonction de risque cumulé au temps t (ou taux cumulé) est la somme cumulée des risques instantanés jusqu'au temps t

$$\Lambda(t) = \int_0^t \lambda(u)du.$$

Pour un risque instantané constant λ au cours du temps, le risque cumulé est :

$$\Lambda(t) = \lambda t.$$

2.4.2 Censure et troncature

Une spécificité des données de survie est qu'elles sont souvent incomplètes. En effet, le délai d'événement peut ne pas être observé : nous sommes donc en présence de données dites censurées. La censure peut avoir lieu à gauche, à droite ou par intervalle. La censure à droite est la plus fréquente dans les études épidémiologique et nous allons nous limiter à cette dernière pour notre travail.

Considérons que la variable aléatoire durée de vie X est censurée par une variable aléatoire C. La censure à droite ($X \leq C$) implique que l'individu n'a pas subi l'événement à sa dernière observation ou a quitté l'étude en cours à une date à laquelle il n'a pas encore subi l'événement. C'est l'exemple le plus connu en analyse de survie dans les études de cohorte.

L'hypothèse d'une censure non informative est souvent faite en analyse de survie. Elle correspond à l'hypothèse d'indépendance entre le temps d'apparition de l'événement et la censure. Lorsqu'elle n'est pas vérifiée, les estimateurs classiques en survie peuvent être biaisés.

En plus de la censure à droite, les données de cohorte épidémiologique peuvent présenter une troncature à gauche. C'est le cas en situation d'entrée retardée, lorsque les individus sont observés à partir d'une date postérieure au début de la période où ils sont à risque de développer l'événement [46]. Ainsi, lorsque la variable X est l'âge, les données sont tronquées à gauche car seuls les sujets vivants à l'inclusion peuvent développer l'événement. La troncature est très différente de la censure. Dans le cas d'une troncature,

on perd l'information sur les observations en dehors de la période d'observation tandis que, dans le cas d'une censure, on sait qu'il existe une information mais on ne connaît pas sa valeur précise, simplement le fait qu'elle excède un seuil. Ainsi, une observation est dite tronquée si elle est conditionnelle à un autre événement.

2.4.3 Notations classiques et théorie des processus de comptage

Soit un échantillon aléatoire de n sujets. Pour tout $i = 1, ..., n$, considérons la variable aléatoire X_i qui représente une durée de vie et $Z_i = (Z_{1i}, ..., Z_{pi})^T$ un vecteur de p covariables. Supposons que X_i est censurée aléatoirement à droite par une variable aléatoire C_i indépendante.

Ainsi, pour chaque individu i, on observe (T_i, d_i, Z_i) où $T_i = \min(X_i, C_i)$ et $d_i = \mathbb{1}_{X_i \leqslant C_i}$ est l'indicatrice de censure.

Considérons les processus aléatoires :

$$N_i(t) = \mathbb{1}_{\{T_i \leqslant t, d_i = 1\}} = \begin{cases} 1 & \text{si le sujet } i \text{ a expérimenté l'événement à l'instant } t, \\ 0 & \text{sinon} \end{cases}$$

et

$$Y_i(t) = \mathbb{1}_{\{T_i \geqslant t\}} = \begin{cases} 1 & \text{si le sujet } i \text{ est à risque à l'instant } t, \\ 0 & \text{sinon.} \end{cases}$$

Considérons également le changement d'état $dN_i(u)$ du sujet i entre les instants $(u - du)$ et u. Cette variable ne prend que deux valeurs : 1 si le sujet i a expérimenté l'événement entre les instants $(u - du)$ et u, 0 sinon. La variable $dN_i(u)$ suit une loi de Bernoulli de paramètre $Y_i(u)d\Lambda(u)$.

Le nombre de sujets ayant expérimenté l'événement avant l'instant t $[N(t)]$ et le nombre d'individus à risque de subir l'événement avant l'instant t $[Y(t)]$ sont deux mesures importantes dans le domaine de la survie. Avec la théorie des processus de comptage, ils sont définis comme :

$$N(t) = \sum_{i=1}^{n} N_i(t)$$

et

$$Y(t) = \sum_{i=1}^{n} Y_i(t).$$

$N(t)$ et $Y(t)$ sont des processus de comptage.

2.4.4 Fonction de vraisemblance

La vraisemblance est une fonction des paramètres du modèle choisi et des données de l'échantillon étudié. Elle représente la probabilité d'observer l'échantillon d'après le modèle.

Considérons le cas d'une censure aléatoire à droite C indépendante de la durée d'intérêt X. Supposons que les variables aléatoires X et C ont pour densités respectives f et g et pour fonctions de survie S et G. La vraisemblance du modèle s'écrit :

$$L = \prod_{i=1}^{n} \left[f(t_i)G(t_i)\right]^{d_i} \times \left[g(t_i)S(t_i)\right]^{1-d_i}.$$

Dans le cas d'une censure non informative, les paramètres du modèle n'apparaissent pas dans la loi de la censure et par conséquent la vraisemblance peut s'écrire comme :

$$L = \prod_{i=1}^{n} \left[f(t_i)\right]^{d_i} \times \left[S(t_i)\right]^{1-d_i}.$$

2.4.5 Méthodes d'estimation non paramétriques

Soit m_i le nombre de sujets ayant développé l'événement d'intérêt en $T_{(i)}$ et r_i le nombre d'individus à risque en $T_{(i)}$ qui représente la statistique d'ordre de la variable aléatoire T_i.

2.4.5.1 Estimateur de Kaplan-Meier pour la survie

L'estimateur de Kaplan-Meier [47], aussi appelé estimateur produit-limite, est défini par :

$$\hat{S}(t) = \prod_{i:T_{(i)} \leq t} \left[1 - \frac{d_i}{r_i}\right] = \prod_{i:T_{(i)}} \left[1 - \frac{d_i}{n-(i-1)}\right] = \prod_{i:T_{(i)} \leq t} \left[\frac{n-i}{n-i+1}\right]^{d_i}.$$

Avec les processus de comptage, cet estimateur peut s'écrire comme [48] :

$$\hat{S}(t) = \prod_{u \leq t} \left[1 - \frac{\Delta N(u)}{Y(u)}\right]$$

où $\Delta N(u) = N(u) - N(u^-)$ est le nombre d'événements au temps u.

Sa variance est donnée par la formule de Greenwood [49] :

$$\text{Var}\left[\hat{S}(t)\right] \approx \hat{S}^2(t)\sigma(t)$$

où $\sigma(t) = \sum_{i:T_i \geq t} \frac{d_i}{r_i(r_i - d_i)}$,

ou encore avec les processus de comptage, $\sigma(t) = \int_0^t \frac{dN(u)}{Y(u)\left[Y(u) - \Delta N(u)\right]}$ qui représente la variance du risque cumulé $\Lambda(t)$.

2.4.5.2 Estimateur de Nelson-Aalen pour le risque cumulé

Pour $T_{(i)} \leq t < T_{(i+1)}$, on estime le risque instantané λ par

$$\hat{\lambda}(t) = \frac{m_i}{r_i}.$$

On estime alors $\Lambda(t) = \int_0^t \lambda(u)du$ par :

$$\hat{\Lambda}(t) = \sum_{\{i:T_i \leq t\}} \frac{m_i}{r_i}.$$

$\hat{\Lambda}$ est l'estimateur de Nelson-Aalen [50, 51].

C'est un estimateur sans biais de la fonction de risque cumulé et sa variance est donnée par :

$$\text{Var}\left[\hat{\Lambda}(t)\right] = \sum_{\{i:T_i \leq t\}} \frac{m_i}{r_i^2}.$$

Avec les processus de comptage, l'estimateur de Nelson-Aalen de la fonction de risque cumulé devient :

$$\hat{\Lambda}(t) = \int_0^t \frac{dN(u)}{Y(u)}$$

et sa variance peut également s'écrire comme :

$$\text{Var}\left[\hat{\Lambda}(t)\right] = \int_0^t \frac{dN(u)}{\left[Y(u)\right]^2}.$$

2.4.5.3 Estimateur de Breslow pour le risque cumulé

L'estimateur de Breslow [52, 53] pour le risque cumulé est obtenu à partir de l'estimateur de Kaplan-Meier [47] de la fonction de survie en utilisant la relation $\Lambda(t) = -\log[S(t)]$. Ainsi,

$$\hat{\Lambda}(t) = -\log\left[\hat{S}(t)\right] = -\sum_{i:T_{(i)}\leq t} \log\left(1 - \frac{m_i}{r_i}\right).$$

La variance de l'estimateur de Breslow pour le risque cumulé est donnée par :

$$\hat{\mathrm{Var}}\{\hat{\Lambda}(t)\} = \sum_{i:T_i\leq t} \frac{m_i}{r_i(r_i - m_i)}.$$

2.4.5.4 Test du log rank

Le test du log rank [54] est un test standard non paramétrique qui permet de comparer deux ou plusieurs courbes de survie avec comme hypothèse nulle que les deux courbes de survie sont identiques. Ce test compare le nombre d'événements observés au nombre d'événements attendus et permet d'utiliser toute l'information sur le suivi. Il attribue le même poids à chaque événement quel que soit le temps où il survient. La statistique du test suit asymptotiquement une distribution de Chi 2 à $k-1$ degrés de liberté avec k le nombre de groupes à comparer.

D'autres tests existent, comme le test de Gehan [55] où le poids correspond au nombre total de sujets à risque en T_i (appelé aussi test de Wilcoxon) ou le test de Peto et Prentice [56] où le poids est fonction du nombre d'événements et du nombre de sujets à risque en T_i.

2.4.6 Méthodes d'estimation paramétriques

Les modèles paramétriques supposent une forme explicite, une fonction paramétrique du temps, pour décrire la distribution de la fonction de survie, du risque instantané et de la densité de probabilité. Ces modèles permettent d'étudier de façon simple l'importance des facteurs de risque susceptibles d'être liés à la survie. Dans cette partie, nous allons en citer quelques uns.

2.4.6.1 Modèle exponentiel

Le modèle exponentiel suppose que le risque instantané $\lambda(t)$ est constant, $\lambda(t) = \theta$. En utilisant les relations d'équivalence, nous pouvons en déduire la forme de la densité et la fonction de survie :

$$f(t|\theta) = \theta \exp(-\theta t) \text{ et } S(t|\theta) = \exp(-\theta t)$$

où θ est le paramètre du modèle représentant le risque d'avoir l'événement et peut être estimé par maximum de vraisemblance :

$$\hat{\theta} = \frac{m}{\sum_{i=1}^{n} t_i}$$

avec m le nombre d'événements.

La variance de l'estimateur de θ est obtenue en utilisant la dérivée seconde de la log-vraisemblance par rapport à θ :

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \frac{m}{\left(\sum_{i=1}^{n} t_i\right)^2}.$$

Une généralisation de ce modèle a été proposée pour prendre en compte plusieurs covariables [57]. Dans ce cas, le risque instantané est indépendant du temps et est défini comme :

$$\lambda(Z_i) = \lambda(Z_{1i}, Z_{2i}, \ldots, Z_{pi})$$

La densité de probabilité et la fonction de survie sont données respectivement par :

$$f(t|Z_i) = \lambda(Z_i) \exp\left[-t\lambda(Z_i)\right] \text{ et } S(t|Z_i) = \exp\left[-t\lambda(Z_i)\right].$$

Plusieurs types de fonctions ont été proposés pour le risque instantané [58] et la plus simple est la suivante :

$$\lambda(Z_i) = \lambda_0 \exp(\beta^T Z_i)$$

où λ_0 est une constante et $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ est un vecteur de paramètres inconnus mesurant l'effet de chaque covariable sur la survie.

Le vecteur de paramètres (λ_0, β) maximisant la vraisemblance du modèle est obtenu en utilisant des algorithmes itératifs comme l'algorithme de Newton-Raphson.

2.4.6.2 Modèle de Weibull

Dans le cas d'une distribution de Weibull de paramètres de forme θ et d'échelle γ, la fonction de densité est donnée par :

$$f(t|\theta,\gamma) = \gamma \left(\frac{1}{\theta}\right)^{\gamma} t^{\gamma-1} \exp\left[-\left(\frac{t}{\theta}\right)^{\gamma}\right].$$

Les fonctions de survie et de risque instantané sont respectivement données par :

$$S(t|\theta,\gamma) = \exp\left[-\left(\frac{t}{\theta}\right)^{\gamma}\right] \quad \text{et} \quad \lambda(t|\theta,\gamma) = \gamma \left(\frac{1}{\theta}\right)^{\gamma} t^{\gamma-1}.$$

La fonction de risque instantané est strictement croissante (respectivement, décroissante) dans le cas où $\gamma > 1$ (respectivement, $\gamma < 1$). Dans le cas où $\gamma = 1$, elle est constante et nous retrouvons la loi exponentielle de paramètre $1/\theta$. Les paramètres du modèle sont obtenus par maximum de vraisemblance.

Il existe d'autres modèles paramétriques comme les distributions Gamma, log normale, log logistique, etc.

2.4.7 Méthodes d'estimation semi-paramétriques

2.4.7.1 Modèles à risques proportionnels et modèle de Cox

Ces modèles sont adaptés au cas d'une variable à expliquer qui peut être censurée. Ils permettent de relier le risque instantané de développer l'événement à une ou plusieurs variables explicatives. Comparativement aux modèles paramétriques qui considèrent que le risque instantané a une forme paramétrique, les modèles semi-paramétriques sont plus adaptés pour évaluer l'effet des covariables sur la survie. Considérons un vecteur Z de covariables de dimension p. Les modèles à risques proportionnels supposent que le risque instantané est de la forme :

$$\lambda(t|Z) = \lambda_0(t)\lambda_\theta(Z)$$

où

— λ_0 est le risque de base qui est commun à tous les sujets et ne dépend pas de Z,

— λ_θ est une fonction connue au paramètre θ près. Elle est appelée risque relatif et ne dépend pas du temps t.

θ est le paramètre de régression et il est inconnu. Ce modèle permet d'étudier l'influence des covariables sur la survenue d'événements. Il est dit à risques proportionnels car si on considère deux individus de profils de covariables z et z^* respectifs alors :

$$\frac{\lambda(t|Z=z)}{\lambda(t|Z=z^*)} = \frac{\lambda_0(t)\lambda_\theta(z)}{\lambda_0(t)\lambda_\theta(z^*)} = \frac{\lambda_\theta(z)}{\lambda_\theta(z^*)} \text{ ne dépend pas du temps.}$$

Ainsi, pour deux individus, leurs risques instantanés de décès restent dans un rapport constant au cours du temps. Ce modèle est un modèle semi-paramétrique car il fait intervenir une partie paramétrique (λ_θ) et une partie non paramétrique (λ_0).

Le modèle de Cox est beaucoup utilisé dans les études épidémiologiques et cliniques. C'est un modèle à risques proportionnels où le risque de base n'est pas spécifié et le risque relatif a une forme exponentielle [30]:

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta^T Z)$$

où β un vecteur de coefficients de régression de dimension p.

Nous pouvons aussi définir le modèle de Cox avec la fonction de survie conditionnellement aux covariables :

$$S(t|Z) = \exp\left[-\int_0^t \lambda_0(u)\exp(\beta^T Z)du\right] = \exp\left[-\int_0^t \lambda_0(u)du\right]^{\exp(\beta^T Z)} = [S_0(t)]^{\exp(\beta^T Z)}$$

où $S_0(t) = S(t|Z=0) = \exp\left[-\int_0^t \lambda_0(u)du\right]$ désigne la fonction de survie en l'absence de covariables.

En prenant le logarithme de la fonction de risque instantané, on obtient une fonction linéaire de Z :

$$\ln\{\lambda(t|Z)\} - \ln\{\lambda_0(t)\} = \beta^T Z$$

C'est l'hypothèse de log-linéarité faite par le modèle de Cox à risques proportionnels.

2.4.7.2 Estimation dans le modèle de Cox

Les paramètres du modèle de Cox sont estimés en utilisant la vraisemblance partielle de Cox [59] :

$$L_n(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta^T Z_i)}{\sum_{j \in R(T_i)} \exp(\beta^T Z_j)} \right]^{d_i}.$$

Avec la théorie des processus de comptage, la vraisemblance partielle de Cox s'écrit [48, 60] :

$$L_n(\beta) = \prod_{i=1}^{n} \prod_{t \leq \tau} \left[\frac{Y_i(t) \exp(\beta^T Z_i)}{\sum_j Y_j(t) \exp(\beta^T Z_j)} \right]^{dN_i(t)} = \prod_{i=1}^{n} \prod_{t \leq \tau} \left[\frac{Y_i(t) \exp(\beta^T Z_i)}{nS^{(0)}(t;\beta)} \right]^{dN_i(t)}$$

où τ est le maximum des temps observés et

$$S^{(0)}(t;\beta) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) \exp(\beta^T Z_i).$$

La fonction de log-vraisemblance peut alors s'écrire comme :

$$l_n(\beta) = \log[L_n(\beta)] = \sum_{i=1}^{n} \int_0^{\tau} \left[\beta^T Z_i - \ln \left(\sum_{j=1}^{n} nS^{(0)}(t;\beta) \right) \right] dN_i(t) + constante.$$

À partir de la fonction de vraisemblance partielle, nous pouvons en déduire le vecteur score :

$$U(\beta) = \frac{\partial \ln [L_n(\beta)]}{\partial \beta} = \sum_{i=1}^{n} \int_0^{\infty} [Z_i - E(u,\beta)] \, dN_i(u)$$

où

$$E(t,\beta) = \frac{S^{(1)}(t;\beta)}{S^{(0)}(t;\beta)} \text{ et } S^{(1)}(t;\beta) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) \exp(\beta^T Z_i) Z_i.$$

À partir du vecteur score, nous pouvons estimer le paramètre β par maximum de vraisemblance partielle. L'estimateur $\hat{\beta}$ est tel que :

$$U(\hat{\beta}) = 0.$$

La matrice d'information de Fisher est :

$$\mathcal{I}(\beta) = \sum_{i=1}^{n} \int_0^{\infty} V(\beta,u) dN_i(u)$$

où V est la variance du vecteur Z à l'instant u,

$$V(\beta, u) = \frac{\sum_{i=1}^{n} Y_i(u) \exp(\beta^T Z_i) \left[Z_i - E(u;\beta)\right]^T \left[Z_i - E(u;\beta)\right]}{\sum_{i=1}^{n} Y_i(u) \exp(\beta^T Z_i)}.$$

2.4.7.3 Estimation du risque de base cumulé

Un estimateur de Breslow [52] de la fonction de risque de base cumulé est :

$$\hat{\Lambda}_0(t) = \sum_{i:T_i \geq t} \frac{m_i}{\sum_{j \in R(T_i)} \exp(\hat{\beta}^T Z_j)}.$$

Avec les processus de comptage, l'estimateur de Breslow peut s'écrire comme :

$$\hat{\Lambda}_0(t) = \int_0^t \frac{dN(u)}{nS^{(0)}(u;\hat{\beta})} = \int_0^t \frac{dN(u)}{\sum_{i=1}^{n} Y_i(u) \exp(\hat{\beta}^T Z_i)}.$$

2.4.7.4 Définition des résidus du modèle de Cox

L'étude des résidus est une partie importante pour la validation du modèle de Cox dans l'analyse des données. Elle nous permet de vérifier certaines hypothèses sous-jacentes au modèle comme l'hypothèse de log-linéarité du risque instantané et l'hypothèse des risques proportionnels. Nous rappelons dans cette partie les résidus du modèle de Cox que nous avons utilisés :

— **Résidus martingales**

Théoriquement, les résidus martingales sont définis comme :

$$M_i(t) = N_i(t) - \int_0^t Y_i(u) \exp(\beta^T Z_i) d\Lambda_0(u).$$

Ils représentent la différence entre la nature observée de l'événement relatif au sujet i et la nature que l'on attendait en théorie compte tenu de la durée de suivi, des covariables et du modèle ajusté. Ils sont de moyenne nulle, compris entre $-\infty$ et 1, négatifs lorsque les durées sont inférieures à celles attendues en théorie, asymptotiquement non corrélés et présentent une forte tendance à l'asymétrie. Leur tracé en fonction des covariables incluses dans le modèle permet de détecter

la non linéarité, c'est-à-dire une forme fonctionnelle mal spécifiée dans la partie paramétrique du modèle. Leur tracé en fonction des durées de vie ou des rangs des durées de vie permet de détecter une influence du temps.

— **Résidus du score**

Les résidus du score pour le i-ème sujet sont définis par :

$$U_i(\beta, t) = \int_0^t [Z_i(s) - E(s, \beta)]\, dM_i(s).$$

Ils représentent chaque contribution individuelle au vecteur score et permettent d'identifier les observations qui contribuent fortement à la détermination des paramètres du modèle.

— **Résidus de Schoenfeld**

Les résidus de Schoenfeld [61] mesurent la distance entre le vecteur des covariables des sujets et la moyenne pondérée des vecteurs des covariables des sujets à risque. Les résidus de Schoenfeld servent à tester l'hypothèse des risques proportionnels. Le résidu correspondant à la covariable $j = 1, \ldots, p$ et à la i-ème durée non censurée est donné par :

$$\hat{S}_i^{(j)} = Z_i^{(j)} - E(T_i, \hat{\beta}).$$

2.4.7.5 Quelques extensions du modèle de Cox

Des extensions du modèle de Cox sont utilisées. Dans le cas où l'hypothèse des risques proportionnels n'est pas vérifiée pour certaines covariables, l'utilisation du modèle stratifié est proposée :

$$\lambda_s(t|Z) = \lambda_s(t) \exp(\beta^T Z),$$

avec $\lambda_s(t|Z)$ et $\lambda_s(t)$ respectivement la fonction de risque instantané et la fonction de risque instantané de base dans la strate s.

Le modèle de Cox s'adapte facilement aux cas de covariables dépendantes du temps. Dans ce cas, le risque instantané s'écrit comme :

$$\lambda\left[t|Z(t)\right] = \lambda_0(t) \exp\left[\beta^T Z(t)\right]$$

Les paramètres du modèle sont facilement estimés mais l'interprétation des coefficients devient délicate.

Enfin, une des hypothèses fortes du modèle de Cox est l'indépendance entres les observations conditionnellement aux covariables. Le modèle de fragilité a été introduit par Vaupel *et al.* [62] et permet de prendre en compte l'hétérogénéité entre les sujets en introduisant, dans le modèle de Cox, un effet aléatoire :

$$\lambda(t|Z,\omega) = \lambda_0(t)\omega \exp(\beta^T Z)$$

où ω est appelé fragilité.

2.4.7.6 Risques compétitifs

Au cours du suivi des participants d'une étude, plusieurs événements pouvant empêcher l'événement d'intérêt de se produire peuvent apparaître : on parle d'événements concurrents ou de compétition [63,64]. Dans cette situation, les méthodes classiques ont tendance à surestimer les fonctions de risque et de survie et ne sont donc pas adaptées. La figure 2.1 représente le cas d'un modèle à K événements compétitifs où $\lambda_k (k = 1, \ldots, K)$ représente le risque spécifique à chaque cause et est défini par :

$$\lambda_k(t) = \lim_{\Delta \to 0} \frac{\mathbf{P}(t \leq T \leq t + \Delta; \epsilon = k | T \geq t)}{\Delta}$$

$\lambda_k(t)$ représente la probabilité de développer l'événement k dans un petit intervalle conditionnellement au fait d'avoir survécu à tous les événements jusqu'au temps t.

Plusieurs approches ont été développées pour estimer le risque spécifique de l'événement d'intérêt et des événements concurrents [64–68], en particulier le modèle semi-paramétrique à risques proportionnels de Cox [30] :

$$\lambda_k(t|Z) = \lambda_{0k} \exp(\beta^T Z)$$

où λ_{0k} est le risque de base non spécifié de l'événement k. La fonction de survie marginale dépend de toutes les fonctions de risque spécifiques à chacun des K événements :

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

Une représentation plus simple est souvent considérée, elle consiste à regrouper tous les événements concurrents pour avoir un modèle avec l'événement d'intérêt et les autres causes ($\varepsilon = 1, 2$).

D'autres approches, permettant d'estimer les probabilités conditionnelles et marginales de l'événement d'intérêt et des événements concurrents, ont également été proposées [69–72]. Ces modèles peuvent être généralisés en utilisant les modèles multi-états [68].

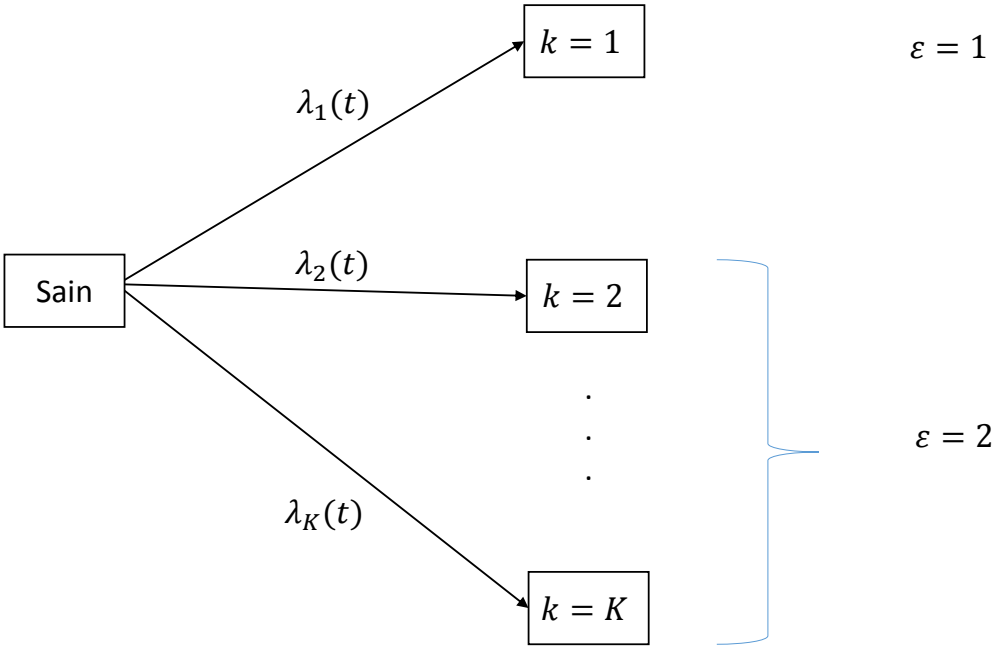


FIGURE 2.1: Risques concurrents

2.5 Risque attribuable et fraction préventive dans le contexte de l'analyse de survie

L'analyse de survie est une partie importante pour l'estimation du RA et de la FP. En effet, les définitions du RA et de la FP reposent essentiellement sur les fonctions de risque instantané, de répartition ou de survie. Dans le contexte de la survie et des études de cohorte, deux types d'approches, que nous détaillerons dans cette partie, sont proposés pour l'estimation du RA. Spiegelman *et al.* [31] proposent d'estimer un RA global ou partiel à l'échelle de la cohorte tandis que d'autres auteurs proposent d'estimer le RA

comme fonction du temps de suivi [4–8, 73]. Les deux types d'approches sont différents. En effet, l'approche de Spiegelman *et al.* [31] utilise une généralisation de la définition du RA, proposée pour les études cas-témoins, dans le contexte de la survie et des études de cohorte tandis que les autres approches, qui donnent un RA fonction du temps, partent de la définition principale proposée par Levin [1] et interprètent la probabilité de la maladie comme une fonction de risque instantané [4, 5] ou une fonction de répartition [3, 6–8].

2.5.1 Risque attribuable global et partiel

Pour les études de cohorte, Spiegelman *et al.* [31] ont proposé d'estimer un RA ajusté global ou partiel en adaptant les estimateurs proposés pour les études cas-témoins [25]. Ces estimateurs prennent en compte la prévalence de l'exposition tout au long du suivi et sont basés sur les personnes-années dans la cohorte et l'estimation du rapport des risques instantanés (HR pour *hazard ratio*) de l'événement d'intérêt associé à l'exposition par un modèle de Cox (ou un modèle de régression logistique groupé) [31]. Le premier est le RA global :

$$RA_G = \frac{\sum_{s=1}^{S} p_s(HR_s - 1)}{1 + \sum_{s=1}^{S} p_s(HR_s - 1)} = 1 - \frac{1}{\sum_{s=1}^{S} p_s HR_s}$$

où HR_s et p_s représentent respectivement le rapport des risques instantanés et la prévalence de l'exposition dans la population pour la s-ième combinaison des facteurs de risque.

Dans les études de cohorte, certains facteurs de risque ne sont pas modifiables, même après élimination de l'exposition ; c'est le cas de l'âge d'entrée dans l'étude et les antécédents familiaux d'une pathologie. Spiegelman *et al.* proposent d'estimer un RA partiel dans ce contexte avec la formule suivante :

$$RA_p = \frac{\sum_{s=1}^{S}\sum_{t=1}^{T} p_{st} HR_{1s} HR_{2t} - \sum_{s=1}^{S}\sum_{t=1}^{T} p_{st} HR_{2t}}{\sum_{s=1}^{S}\sum_{t=1}^{T} p_{st} HR_{1s} HR_{2t}} = 1 - \frac{\sum_{t=1}^{T} p_{.t} HR_{2t}}{\sum_{s=1}^{S}\sum_{t=1}^{T} p_{st} HR_{1s} HR_{2t}}$$

où t désigne un profil unique parmi les niveaux combinés de tous les facteurs de risque d'ajustement (facteurs non modifiables, autres que les facteurs d'intérêt pour l'étude), $t = 1, \ldots, T$ et HR_{2t} est le HR associé à la combinaison t relativement au niveau de risque

le plus bas pour lequel $HR_{2,1} = 1$. Comme dans la définition du RA global plus haut, s représente un profil d'exposition unique parmi les niveaux combinés des facteurs de risque d'intérêt (facteurs modifiables auxquels le RA partiel s'applique), $s = 1, \ldots, S$ et HR_{1s} est le HR associé à la combinaison s relativement au niveau de risque le plus bas pour lequel $HR_{1,1} = 1$. La prévalence conjointe des profils d'exposition s et de facteurs non modifiables t est notée p_{st} et $p_{.t} = \sum_{s=1}^{S} p_{st}$ représente la prévalence marginale dans la strate t.

Le RA partiel représente la différence entre le nombre de cas attendus dans la cohorte initiale et le nombre de cas attendus si tous les sujets des sous-ensembles de la cohorte qui avaient été à l'origine exposés aux facteurs de risque modifiables avaient éliminé leur exposition de telle sorte que le HR par rapport aux non exposés soit égal à 1, divisé par le nombre de cas attendus dans la cohorte initiale.

Spiegelman *et al.* [31] proposent d'estimer les RA global et partiel en estimant les prévalences des niveaux d'exposition avec les personnes-années dans les strates correspondantes et les HR de l'événement d'intérêt et des facteurs de confusion en utilisant une régression de Poisson, une régression logistique groupée ou un modèle de Cox à risques proportionnels. Une expression de la variance est également proposée en utilisant la delta méthode [31].

2.5.2 Risque attribuable comme fonction du temps

Contrairement à Spiegelman *et al.* [31], des auteurs définissent le RA comme fonction du temps [3–8]. Ces auteurs interprètent différemment les probabilités $\mathbf{P}(D)$ et $\mathbf{P}(D|\bar{E})$ dans le contexte de l'analyse de survie. Par conséquent, il n'existe pas de définition consensuelle du RA.

2.5.3 Définition et méthodes d'estimation avec la fonction de survie

Plusieurs auteurs [3, 4, 6–8] interprètent $\mathbf{P}(D)$ comme la probabilité d'événement jusqu'à un temps t, soit $F(t) = \mathbf{P}(T \leq t)$ la fonction de répartition. Le RA est alors défini comme :

$$A(t) = \frac{\mathbf{P}(T \leq t) - \mathbf{P}(T \leq t | Z = z^*)}{\mathbf{P}(T \leq t)} \tag{2.12}$$

où z^* représente les valeurs cibles prises par Z afin de quantifier leur impact potentiel. Lorsqu'il y a un seul facteur de risque Z, nous considérons comme valeurs cibles $Z = 0$, qui représente les non exposés, à la place de $Z = z^*$. En utilisant les fonctions de survie $S(t) = \mathbf{P}(T \geq t)$ et $S_0(t) = \mathbf{P}(T \geq t | Z = 0)$, le RA peut s'écrire comme :

$$A(t) = \frac{\mathbf{P}(T \leq t) - \mathbf{P}(T \leq t | Z = 0)}{\mathbf{P}(T \leq t)} = \frac{F(t) - F(t|Z = 0)}{F(t)} = \frac{S_0(t) - S(t)}{1 - S(t)}. \tag{2.13}$$

Plusieurs méthodes d'estimation du RA ont été proposées par des auteurs différents, voire par les auteurs d'une même publication.

Un estimateur naturel consiste à remplacer $S(t)$ et $S_0(t)$ par leurs estimateurs respectifs $\hat{S}(t)$ et $\hat{S}_0(t)$ obtenus selon différentes approches. Ainsi,

$$\hat{A}(t) = \frac{\hat{S}_0(t) - \hat{S}(t)}{1 - \hat{S}(t)}.$$

Chen *et al.* [6] ont montré que $\sqrt{n}\left[\hat{A}(t) - A(t)\right]$ converge faiblement vers un processus gaussien de moyenne nulle et de fonction de variance-covariance $\mathbf{E}\left[\xi(t)\xi(s)^T\right]$ entre les temps t et s, où

$$\xi(t) = \frac{1}{1 - S(t)} \left\{ \eta_1(t) - \frac{1 - S_0(t)}{1 - S(t)} \eta_2(t) \right\}$$

et $\eta_1(t)$ et $\eta_2(t)$ dépendent des méthodes d'estimation de $S_0(t)$ et $S(t)$ respectivement [6]. Un estimateur consistant de la fonction de variance-covariance est alors

$$\hat{\sigma}_A(s,t) = \frac{1}{n} \sum_{i=1}^{n} \hat{\xi}_i(t) \hat{\xi}_i(s)^T$$

où

$$\hat{\xi}_i(t) = \frac{1}{1 - \hat{S}(t)} \left\{ \hat{\eta}_{1i}(t) - \frac{1 - \hat{S}_0(t)}{1 - \hat{S}(t)} \hat{\eta}_{2i}(t) \right\}$$

et $\hat{\eta}_{1i}(t)$ et $\hat{\eta}_{2i}(t)$ sont les estimations de $\eta_1(t)$ et $\eta_2(t)$ respectivement pour le i-ème sujet, $i = 1, ..., n$ [6]. La variance de $\hat{A}(t)$ est alors approchée par [6]:

$$\hat{\sigma}_A(t)^2 = \hat{\sigma}_A(t,t) = \frac{1}{n}\sum_{i=1}^{n}\hat{\xi}_i(t)^2.$$

En présence d'autres facteurs d'ajustement W, Chen *et al.* [6] ont proposé une définition du RA ajusté :

$$A_{adj}(t) = \frac{\mathbf{E}\left[S\{t|(Z=0,W)^T\}\right] - S(t)}{1-S(t)} = \frac{\int S\{t|(Z=0,w)\}^T dF_W(w) - S(t)}{1-S(t)}$$

où $F_W(.)$ est la distribution marginale de W. Ils proposent également un estimateur du RA ajusté [6] :

$$\hat{A}_{adj}(t) = \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{S}\{t|(Z=0,W_i)^T\} - \hat{S}(t)}{1-\hat{S}(t)}$$

où W_i est la valeur observée de W pour le i-ème sujet. Chen *et al.* [6] ont également montré que $\sqrt{n}\left[\hat{A}_{adj}(t) - A_{adj}(t)\right]$ converge faiblement vers un processus gaussien de moyenne nulle et de fonction de variance-covariance $\mathbf{E}\{\xi(t)\xi(s)^T\}$ entre les temps t et s avec

$$\xi(t) = \frac{1}{1-S(t)}\left\{\eta_1(t) - \frac{1-\mathbf{E}\left[S\{t|(Z=0,W)^T\}\right]}{1-S(t)}\eta_2(t)\right\}.$$

L'utilisation de la transformation logarithmique est proposée pour le calcul des IC du RA [3, 6]:

$$IC_{95\%}\left[A(t)\right] = 1 - \{1-A(t)\}\exp\left[\pm 1{,}96\sqrt{\widehat{Var}\left\{\ln\left[1-\hat{A}(t)\right]\right\}}\right].$$

En appliquant la delta méthode, cet intervalle devient :

$$IC_{95\%}\left[A(t)\right] = 1 - \{1-A(t)\}\exp\left[\pm\frac{1{,}96\sqrt{\widehat{Var}\left\{\hat{A}(t)\right\}}}{1-A(t)}\right].$$

La définition de la FP n'a pas été considérée dans le contexte de l'analyse de survie mais nous pouvons aisément la déduire de la définition du RA. Dans la suite, nous allons détailler les différentes méthodes d'estimation des fonctions de survie, par conséquent du RA.

2.5.3.1 Méthodes non paramétriques

Lorsque les covariables sont catégorielles et non dépendantes du temps et sous l'hypothèse d'une censure indépendante des covariables, Chen *et al.* [6] proposent d'estimer $S_0(.)$ et $S(.)$ selon la méthode de Kaplan-Meier [47]. Dans ce cas,

$$\eta_1(t) = -S_0(t) \int_0^t \frac{dM_0(u)}{\mathbf{E}\left[\mathbb{1}_{Z=0} Y(u)\right]}$$

et

$$\eta_2(t) = -S(t) \int_0^t \frac{dM(u)}{\mathbf{E}\left[Y(u)\right]}$$

où $M(t) = N(t) - \int_0^t Y(u) d\Lambda(u)$ est le résidu martingale, avec $\Lambda(t) = -\ln\{S(t)\}$ et $M_0(t) = \mathbb{1}_{Z=0} \left\{ N(t) - \int_0^t Y(u) d\Lambda_0(u) \right\}$ son équivalent chez les non exposés avec $\Lambda_0(t) = -\log\{S_0(t)\}$.

Chen *et al.* notent que l'estimateur de Kaplan-Meier pour $S_0(.)$ peut être instable et inefficace si le nombre de sujets non exposés est faible [6].

Dans le cas où l'hypothèse de la censure indépendante des covariables n'est pas vérifiée, Chen *et al.* [6] proposent d'estimer $S_0(.)$ par la méthode de Kaplan-Meier et $S(.)$ par la méthode de Kaplan-Meier pondérée, soit l'estimateur [74] :

$$\hat{S}(t) = n^{-1} \sum_{k=0}^{K} n_k \hat{S}_k(t),$$

où $\hat{S}_k(t)$ est l'estimateur non paramétrique de Kaplan-Meier pour chaque profil de covariable $k = 0, 1, 2, \ldots, K$ et n_k le nombre de sujets avec comme profil de covariable k tel que $\sum_{k=0}^{K} n_k = n$ le nombre total de sujets. Dans le cas d'une covariable dichotomique,

$$\hat{S}(t) = (1 - \hat{p})\hat{S}_0(t) + \hat{p}\hat{S}_1(t)$$

avec $\hat{p} = \dfrac{n_1}{n}$ la proportion de sujets exposés à $t = 0$, n_1 étant le nombre total d'exposés. Pour le calcul de la variance,

$$\eta_2(t) = S(t|Z) - S(t) + \psi(t; Z)$$

avec

$$\psi(t;z) = -\mathbf{P}(Z=z)S(t|z)\int_0^t \frac{dN(u) - Y(u)d\Lambda(u|z)}{\mathbf{E}\left[\mathbb{1}_{Z=z}Y(u)\right]}.$$

Si $z = 0$

$$\begin{aligned}\psi(t;0) &= -\mathbf{P}(Z=0)S(t|z=0)\int_0^t \frac{dN(u) - Y(u)d\Lambda(u|z=0)}{\mathbf{E}\left[\mathbb{1}_{Z=0}Y(u)\right]} \\ &= -(1-p)S_0(t)\int_0^t \frac{dN(u) - Y(u)d\Lambda(u|z=0)}{\mathbf{E}\left[\mathbb{1}_{Z=0}Y(u)\right]}\end{aligned}$$

et si $z = 1$

$$\begin{aligned}\psi(t;1) &= -\mathbf{P}(Z=1)S(t|z=1)\int_0^t \frac{dN(u) - Y(u)d\Lambda(u|z=1)}{\mathbf{E}(\mathbb{1}_{Z=1}Y(u)} \\ &= -pS_1(t)\int_0^t \frac{dN(u) - Y(u)d\Lambda(u|z=1)}{\mathbf{E}\left[\mathbb{1}_{Z=1}Y(u)\right]}.\end{aligned}$$

Nous avons développé un programme sous le logiciel `R` permettant d'estimer le RA et sa variance pour les approches non paramétriques KM et KMP (cf. annexe A).

Cette approche peut être adaptée pour estimer le RA ajusté sur un ensemble de covariables W et sa variance en utilisant $\mathbf{E}\left[S\{t|(Z=0,W)\}\right]$ et le vecteur $(Z=0,W)$ au lieu de $S_0(t)$ et $Z=0$ respectivement [6].

2.5.3.2 Méthodes semi-paramétriques

D'une manière générale, lorsque les covariables sont continues et/ou dépendantes du temps, Chen *et al.* [6] suggèrent d'utiliser les modèles semi-paramétriques pour estimer les fonctions de survie $S_0(.)$ et $S(.)$. Dans le cas particulier du modèle de Cox,

$$\hat{S}_0(t) = \exp\Big[-\hat{\Lambda}_0(t)\Big] \text{ et } \hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n}\exp\left[-\int_0^t \exp\{\hat{\beta}^T z_i(u)\}d\hat{\Lambda}_0(u)\right]$$

où $\hat{\Lambda}_0(.)$ est l'estimateur de Breslow [52] du risque de base cumulé $\Lambda_0(t) = \int_0^t \lambda_0(u)du$.

Dans ce cas,

$$\eta_1(t) = -S_0(t)\left\{\int_0^t \frac{dM(u,\beta)}{\mathbf{E}\left[Y(u)\exp\left\{\beta^T Z(u)\right\}\right]} - \int_0^t e^T(u,\beta)d\Lambda_0(u)\mathcal{I}^{-1}(\beta)U(\beta)\right\}$$

où

$$M(t,\beta) = N(t) - \int_0^t Y(u)\exp\left[\beta^T Z(u)\right] d\Lambda_0(u) \text{ est le résidu martingale,}$$

$$U(\beta) = \int_0^\tau \left[Z(u) - e(u,\beta)\right] dM(u,\beta) \text{ est le résidu du score,}$$

$e(t,\beta) = \dfrac{\mathbf{E}\left[Y(t)\exp\left\{\beta^T Z(u)\right\} Z(u)\right]}{\mathbf{E}\left[Y(t)\exp\left\{\beta^T Z(u)\right\}\right]}$ représente la moyenne pondérée des covariables Z sur les observations encore à risque à l'instant t, $\mathcal{I}(\beta)$ est la matrice d'information de Fisher de β et τ est la durée d'étude d'une part ; d'autre part,

$$\eta_2(t) = S(t|Z) - S(t) + (\mathcal{S}_\beta, \mathcal{S}_{\Lambda_0}) I^{-1}_{\beta,\Lambda_0} \left[\mathbf{E}\{h_1(t,Z)\}, \mathbf{E}\{(h_2(.;t,Z)\}\right]$$

où $(\mathcal{S}_\beta, \mathcal{S}_{\Lambda_0})$ est un vecteur score de β et Λ_0,

$$h_1(t,z) = -S(t|z) \int_0^t \exp\left[\beta^T z(u)\right] z(u) d\Lambda_0(u)$$

et

$$h_2(v;t,z) = -S(t|z)\exp\left[\beta^T z(u)\right] \mathbb{1}_{v\leq t}.$$

Comme pour l'approche non paramétrique, un estimateur du RA ajusté est également proposé dans le cas d'une approche semi-paramétrique [6]. Chen [75] a publié un *package* `R` permettant de calculer le RA et sa variance pour l'approche semi-paramétrique reposant sur le modèle de Cox.

Sjölander et Vansteelandt [8] ont récemment proposé deux approches semi-paramétriques robustes pour estimer les fonctions de survie $S(.)$ et $S_0(.)$: l'estimateur IPW (pour *Inverse Probability Weighted*) et l'estimateur DR (pour *Doubly Robust*) lorsque les modèles utilisés pour estimer les fonctions de survie sont mal spécifiés (c'est-à-dire, par exemple, lorsque les temps d'événements sont générés en utilisant un modèle de Cox à risques proportionnels avec une interaction entre les covariables du modèle et que les estimations sont obtenues avec le même modèle mais sans interaction). Lorsque le nombre de sujets non exposés est faible, l'estimateur de la fonction de survie chez les non exposés en utilisant l'approche semi-paramétrique proposée par Chen *et al.* [6] peut être biaisé.

Une alternative serait donc d'utiliser les estimateurs IPW et DR, que nous détaillerons dans la suite, pour estimer le RA.

— **Estimateur IPW**

Dans le cas d'une censure non informative, Sjölander et Vansteelandt [8] proposent d'estimer la fonction de survie marginale en résolvant l'équation suivante d'inconnue $S(t)$ [76] :

$$\sum_{i=1}^{n} S(t) - \frac{\mathbb{1}_{\{T_i>t\}}}{S_C(t|Z_i, X_i)} = 0$$

Dans le cas où tous les facteurs de confusion sont mesurés, ces auteurs proposent d'estimer la fonction de survie chez les non exposés en résolvant l'équation suivante d'inconnue $S_0(t)$:

$$\sum_{i=1}^{n} S_0(t) - \frac{\mathbb{1}_{\{Z_i=0\}}\mathbb{1}_{\{T_i>t\}}}{\pi(X_i)S_C(t|Z_i, X_i)} = 0$$

où X représente un ensemble de facteurs à l'inclusion, $S_C(t|Z, X) = \mathbf{P}(C > t|Z, X)$ et $\pi(X) = \mathbf{P}(Z = 0|X)$. En pratique, $S_C(t|Z, X)$ et $\pi(X)$ ne sont pas connus. Sjölander et Vansteelandt [8] proposent donc d'utiliser un modèle semi-paramétrique à risques proportionnels de Cox pour $S_C(t|Z, X)$ et une régression logistique pour $\pi(X)$. L'estimateur IPW du RA résultant est un estimateur consistant et asymptotiquement normal. Sa variance est obtenue par delta méthode :

$$\frac{\partial A(t)}{\partial\{S(t), S_0(t)\}}\Sigma\frac{\partial A(t)}{\partial\{S(t), S_0(t)\}^T} = \frac{1}{\{1 - S(t)\}^2}\left\{-\frac{1 - S_0(t)}{1 - S(t)}, 1\right\}\Sigma\left\{-\frac{1 - S_0(t)}{1 - S(t)}, 1\right\}^T$$

où Σ est la matrice de variance-covariance du vecteur $\{\hat{S}(t), \hat{S}_0(t)\}^T$

— **Estimateur DR**

Dans les mêmes conditions, Bai *et al.* [76] proposent un estimateur sans biais des fonctions de survie marginale et conditionnelle chez les non exposés en résolvant les équations d'inconnues $S(t)$ et $S_0(t)$ suivantes :

$$\sum_{i=1}^{n} S(t) - \frac{\mathbb{1}_{\{T_i>t\}}}{S_C(t|Z_i, X_i)} - S(t|Z_i, X_i)\int_0^t \frac{dM_C(u, Z_i, X_i, T_i, \delta_i)}{S_C(t|Z_i, X_i)S(t|Z_i, X_i)} = 0$$

et

$$\sum_{i=1}^{n}\left[S_0(t)-\frac{\mathbb{1}_{\{Z_i=0\}}\mathbb{1}_{\{T_i>t\}}}{\pi(X_i)S_C(t|Z_i,X_i)}+\frac{\{\mathbb{1}_{\{Z_i=0\}}-\pi(X_i)\}}{\pi(X_i)}S(t|X_i,Z_i=0)\right.$$
$$\left.-S(t|X_i,Z_i=0)\frac{\mathbb{1}_{\{Z_i=0\}}}{\pi(X_i)}\int_0^t\frac{dM_C(u,Z_i,X_i,T_i,\delta_i)}{S_C(t|Z_i,X_i)S(t|Z_i,X_i)}\right]=0$$

Un estimateur de la variance basé sur la delta méthode a également été proposé [8].

En pratique, il est souvent difficile de vérifier si les modèles utilisés sont mal spécifiés, ce qui peut être une limite d'utilisation des deux approches proposées par Sjölander et Vansteelandt [8].

2.5.3.3 Modèles de transformation

Dans le cas où l'hypothèse des risques proportionnels n'est pas vérifiée, Chen *et al.* [6] suggèrent d'utiliser des modèles de transformation :

$$\Lambda(t|Z)=G\left\{\int_0^t\exp\left\{\beta^TZ(u)\right\}d\Lambda_0(u)\right\}$$

où G est une fonction strictement croissante. Deux classes de transformation ont été proposées par Chen *et al.* [6]. La première est une transformation Box-Cox soit

$$G(x)=\begin{cases}\dfrac{(1+x)^\rho-1}{\rho} & \text{si } \rho>0,\\ \ln(1+x) & \text{si } \rho=0.\end{cases}$$

La deuxième est une transformation logarithmique soit

$$G(x)=\begin{cases}\dfrac{\ln(1+rx)}{r} & \text{si } r>0,\\ x & \text{si } r=0.\end{cases}$$

Les cas $\rho=1$ et $r=0$ correspondent au modèle de Cox à risques proportionnels. Chen *et al.* [6] proposent d'estimer $S_0(.)$ comme suit :

$$\hat{S}_0(t)=\exp\left[-G\left\{\hat{\Lambda}_0(t)\right\}\right]$$

où $\hat{\Lambda}_0(t)$ est un estimateur non paramétrique du risque de base cumulé obtenu par maximum de vraisemblance [77]. Dans ce cas,

$$\eta_1(t)=-S_0(t)G'\left\{\Lambda_0(t)\right\}\left(\mathcal{S}_\beta,\mathcal{S}_{\Lambda_0}\right)I^{-1}_{\beta,\Lambda_0}\left\{0,h(.;t)\right\}$$

où $(\mathcal{S}_\beta, \mathcal{S}_{\Lambda_0})$ est le vecteur score en β et Λ_0, I_{β,Λ_0} est la matrice d'information en β et Λ_0 et $h(v;t) = \mathbb{1}_{\{v \leqslant t\}}$.

Chen *et al.* [6] proposent d'estimer $S(.)$ par

$$\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-G\left[\int_0^t \exp\{\hat{\beta}^T z_i(u)\} d\hat{\Lambda}_0(u)\right]\right)$$

où $\hat{\beta}$ et $\hat{\Lambda}_0$ sont des estimateurs non paramétriques de β et Λ_0 obtenus par maximum de vraisemblance. Ainsi,

$$\eta_2(t) = S(t|Z) - S(t) + (\mathcal{S}_\beta, \mathcal{S}_{\Lambda_0})\, I^{-1}_{\beta,\Lambda_0}\left[\mathbf{E}\{h_1(t,Z)\}, \mathbf{E}\{h_2(.;t,Z)\}\right]$$

où

$$h_1(t,z) = -S(t|z)G'\left[\int_0^t \exp\{\beta^T z(u)\} d\Lambda_0(u)\right]\int_0^t \exp\{\beta^T z(u)\} d\Lambda_0(u)$$

et

$$h_2(v;t;z) = -S(t|z)G'\left[\int_0^t \exp\{\beta^T z(u)\} d\Lambda_0(u)\right]\exp\{\beta^T z(v)\}\mathbb{1}_{\{v \leqslant t\}}.$$

2.5.3.4 Méthode paramétrique

Laaksonen *et al.* [3] ont proposé un estimateur du RA encore appelé fraction attribuable dans la population (PAF, pour *population attributable fraction*) basé sur un modèle à risques proportionnels avec un risque de base constant par morceaux :

$$PAF(t; t+\Delta t) = 1 - \frac{S(t|Z^*) - S(t+\Delta t|Z^*)}{S(t) - S(t+\Delta t)}.$$

En particulier, pour $t = 0$ et $\Delta t = t$, $PAF(0;t)$ correspond à la définition du RA à partir des fonctions de survie [6–8] :

$$PAF(0;t) = 1 - \frac{S(t=0|Z^*) - S(t|Z^*)}{S(0) - S(t)} = 1 - \frac{1 - S_0(t)}{1 - S(t)} = \frac{S_0(t) - S(t)}{1 - S(t)} = A(t).$$

Dans l'approche proposée par Laaksonen *et al.* [3], le temps de suivi est divisé en J intervalles $]0 = a_0, a_1],]a_1, a_2], \ldots,]a_{j-1}, a_j], \ldots,]a_{J-1}, a_J]$ et le risque instantané est fonction du temps avec un risque de base qui change d'un intervalle à l'autre [78]. L'effet de l'âge peut être pris en compte dans le modèle en divisant l'étendue des dates de naissance

en C cohortes de naissance $]v_0, v_1],]v_1, v_2], \ldots,]v_{C-1}, v_C]$ et en stratifiant le modèle sur cette variable [79].

Le risque instantané au temps t pour le i-ème sujet conditionnellement à la cohorte de naissance c_i et au facteur de risque Z_i est donné par :

$$\lambda(t|c_i, Z_i) = \sum_{j=1}^{J} \mathbb{1}_{\{a_{j-1} \leq t \leq a_j\}} \lambda_{ij}(t|c_i, Z_i) = \sum_{j=1}^{J} \mathbb{1}_{\{a_{j-1} \leq t \leq a_j\}} \exp(\alpha_{0jc_i} + \beta^T Z_i)$$

où $\lambda_{ij}(t|c_i, Z_i) = \exp(\alpha_{jc_i} + \beta^T Z_i)$ et $\alpha_{jc_i} = \ln(\lambda_{0jc_i})$, avec λ_{0jc_i} le risque de base dans le j-ème intervalle et pour la cohorte de naissance c_i.

Le risque cumulé peut être calculé comme :

$$\begin{aligned}
\Lambda(t|c_i, Z_i) &= \int_0^t \sum_{j=1}^{J} \mathbb{1}_{\{a_{j-1} \leq u \leq a_j\}} \exp(\alpha_{0jc_i} + \beta^T Z_i) du \\
&= \sum_{j=1}^{J} \exp(\alpha_{0jc_i} + \beta^T Z_i) \int_0^t \mathbb{1}_{\{a_{j-1} \leq u \leq a_j\}} du \\
&= \sum_{j=1}^{J} \exp(\alpha_{0jc_i} + \beta^T Z_i) \delta_j(t)
\end{aligned}$$

où $\delta_j(t)$ est la longueur du j-ème intervalle :

$$\delta_j(t) = \begin{cases} 0 & \text{si } t \leq a_{j-1}, \\ t - a_{j-1} & \text{si } a_{j-1} < t \leq a_j, \\ a_j - a_{j-1} & \text{si } t > a_j. \end{cases}$$

La fonction de survie $S(t)$ est estimée par :

$$\hat{S}(t) = \frac{1}{n} \sum_{i=1}^{n} \exp\left[- \sum_{j=1}^{J} \exp(\hat{\alpha}_{jc_i} + \hat{\beta}^T Z_i) \delta_j(t) \right].$$

L'estimation des paramètres du modèle $\mu = (\{\alpha_{jc_i}\}, \beta)$ découle de la maximisation de la vraisemblance du modèle :

$$\begin{aligned}
L(\mu) &= \prod_{i=1}^{n} \prod_{j=1}^{J} (\lambda_{ij})^{d_{ij}} \exp\left[-\lambda_{ij} \delta_j(t_i)\right] \\
&= \prod_{i=1}^{n} \prod_{j=1}^{J} \exp(\alpha_{jc_i} + \beta^T Z_i)^{d_{ij}} \exp\left[- \exp(\alpha_{jc_i} + \beta^T Z_i) \right] \\
&= \exp\left[\sum_{i=1}^{n} \sum_{j=1}^{J} d_{ij} (\alpha_{jc_i} + \beta^T Z_i) \right] \exp\left[- \sum_{i=1}^{n} \sum_{j=1}^{J} \exp(\alpha_{jc_i} + \beta^T Z_i) \delta_j(t_i) \right]
\end{aligned}$$

ou de la log vraisemblance qui s'écrit :

$$l(\mu) = \ln\left[L(\mu)\right] = \sum_{i=1}^{n}\sum_{j=1}^{J} d_{ij}(\alpha_{jc_i} + \beta^T Z_i) - \sum_{i=1}^{n}\sum_{j=1}^{J} \exp(\alpha_{jc_i} + \beta^T Z_i)\delta_j(t_i).$$

Dans ces équations, d_{ij} indique si l'individu i a eu l'événement dans le j-ème intervalle :

$$d_{ij} = \begin{cases} 1 & \text{si } 0 < \delta_j(t_i) < \min(t_i, a_j) - a_{j-1}, \\ 0 & \text{sinon} \end{cases}$$

L'estimation des paramètres du modèle est obtenue en posant l'équation : $U(\mu) = 0$ où $U(\mu)$ est le vecteur score obtenu en dérivant la log vraisemblance par rapport aux paramètres du modèle.

La variance de l'estimateur du RA est calculée en utilisant la delta méthode [73] :

$$\widehat{\text{Var}}\{\hat{A}(t)\} = \frac{\partial A(t)}{\partial \mu}\Sigma\frac{\partial A(t)}{\partial \mu}^T$$

où Σ est la matrice de variance-covariance associée au vecteur des paramètres du modèle μ.

Une généralisation de cet estimateur permettant de prendre en compte des événements concurrents a également été proposée [73]. Dans le cas où le résultat d'intérêt est l'incidence de la maladie, Laaksonen *et al.* [73] définissent le RA pour l'incidence de la maladie encore appelé *population attributable risk for incidence of disease* qui mesure la proportion de cas de maladie théoriquement évités au cours du suivi dans l'intervalle $(0, t]$ lorsque les facteurs de risque ont été modifiés en considérant la mortalité et les différentes types de censure comme des événements concurrents [73] :

$$PAF(T^M \leq T^D) = \frac{\sum_{i=1}^{n} \mathbf{P}(T_i^M < T_i^D | Z_i) - \sum_{i=1}^{n} \mathbf{P}(T_i^M < T_i^D | Z_i^*)}{\sum_{i=1}^{n} \mathbf{P}(T_i^M < T_i^D | Z_i)}$$

où T_M et T_D représentent respectivement le délai jusqu'à l'apparition de la maladie et le délai jusqu'à l'apparition d'un décès. En pratique, le temps d'événement n'étant pas toujours observé, nous considérons la probabilité $\mathbf{P}\left[T^M \leq \min(T^D, t)\right]$ où t est la durée de suivi [73]. La figure 2.2 représente le modèle associé où $\lambda_M(t)$ représente le risque associé à l'incidence de la maladie, $\lambda_{D_1}(t)$ le risque de décès avant de contracter la maladie et

$\lambda_{D_2}(t)$ le risque de décès après avoir contracté la maladie. Un estimateur paramétrique du RA de l'incidence de la maladie, basé sur un modèle à risques proportionnels avec un risque de base constant par morceaux, a été proposé par Laaksonen *et al.* [73] :

$$PAF(T^M \leq T^D) = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{J} \frac{\lambda_{ij}^{*M}}{\lambda_{ij}^{*M} + \lambda_{ij}^{*D_1}}(S_{i,j-1}^{*} - S_{ij}^{*})}{\sum_{i=1}^{n}\sum_{j=1}^{J} \frac{\lambda_{ij}^{M}}{\lambda_{ij}^{M} + \lambda_{ij}^{D_1}}(S_{i,j-1} - S_{ij})}$$

où

$$\lambda_{ij}^{M} = \lambda_{0jc_i}\exp(\beta_M^T Z_i), \quad \lambda_{ij}^{*M} = \lambda_{0jc_i}\exp(\beta_M^T Z_i^*)$$

$$\lambda_{ij}^{D_1} = \lambda_{0jc_i}\exp(\beta_{D_1}^T Z_i), \quad \lambda_{ij}^{*D_1} = \lambda_{0jc_i}\exp(\beta_{D_1}^T Z_i^*),$$

$$S_{ij} = S_{ij}^{M} S_{ij}^{D} = \exp\left[-\sum_{k=1}^{j}\left(\lambda_{ik}^{M} + \lambda_{ik}^{D}\right)(a_k - a_{k-1})\right]$$

et

$$S_{ij}^{*} = S_{ij}^{*M} S_{ij}^{*D} = \exp\left[-\sum_{k=1}^{j}\left(\lambda_{ik}^{*M} + \lambda_{ik}^{*D}\right)(a_k - a_{k-1})\right].$$

Lorsque l'événement d'intérêt est le décès, la censure due à la fin de suivi ou aux perdus de vue doit être considérée. Lorsque l'événement d'intérêt est l'incidence de la maladie, la censure due aux événements concurrents de l'événement d'intérêt, en particulier les décès d'autres causes que l'événement d'intérêt, doit être prise en compte dans l'estimation du RA. Si les facteurs de risque qui sont liés à l'incidence de la maladie sont également liés à ces événements concurrents, la modification de ces facteurs de risque est susceptible d'influer différemment sur le risque de maladie et le risque de décès. Ainsi, dans l'estimation du RA pour l'incidence de la maladie, en plus de la censure due à la fin de suivi, la censure due aux décès doit être prise en compte. Ignorer la censure due aux décès lors de l'estimation du RA pour l'incidence de la maladie signifie que les estimations obtenues ne sont applicables que dans l'hypothèse que personne ne meurt pendant le suivi au cours duquel l'incidence de la maladie est estimée. Ainsi, nous avons besoin de différents estimateurs du RA selon l'événement d'intérêt afin d'obtenir des résultats précis et sans biais.

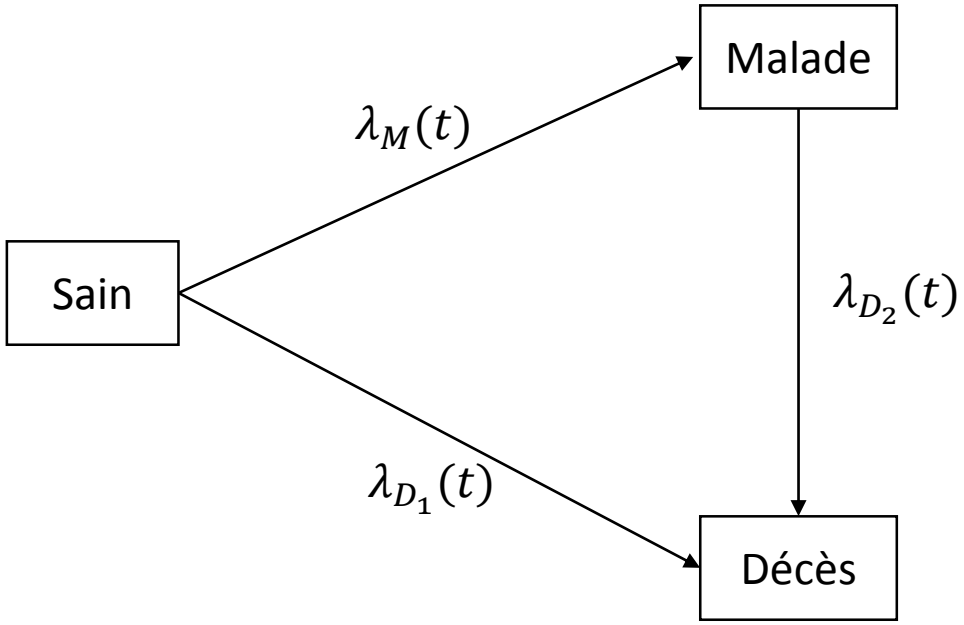


FIGURE 2.2: Modèle maladie-décès de la maladie d'intérêt et risques correspondants

2.5.3.5 Études comparatives

Les auteurs qui ont proposé ces définitions et méthodes d'estimation du RA ont souvent réalisé des études de simulation pour évaluer la performance des estimateurs proposés. Chen *et al.* [6] ont réalisé une étude de simulation pour les approches non paramétriques KM et KMP, l'approche semi-paramétrique et les modèles de transformation dans le cas où les temps d'événements sont générés en utilisant un modèle de transformation avec une censure indépendante et dépendante des covariables en supposant un risque de base constant. Dans le cas d'une censure indépendante, Chen *et al.* [6] ont trouvé que les estimateurs du RA sont sans biais avec de bons taux de recouvrement sauf pour les approches non paramétriques KM et KMP en fin d'étude. Dans le cas d'une censure dépendant des covariables, les résultats obtenus étaient satisfaisants sauf lorsque la fonction de survie marginale était estimée en utilisant l'approche non paramétrique KM. Cependant, une des limites de leur étude de simulation est que ces auteurs n'ont pas évalué la performance des estimateurs proposés pour l'approche semi-paramétrique par modèle de Cox dans le cas où l'hypothèse des risques proportionnels n'est pas vérifiée.

Sjölander et Vansteelandt [8] ont également réalisé une étude de simulation en comparant les trois approches semi-paramétriques à savoir l'approche semi-paramétrique proposée par Chen *et al.* [6] et les deux approches qu'ils ont proposées, les estimateurs IPW et DR, dans trois situations : lorsque les temps d'événement, la censure ou l'exposition sont mal spécifiés. Sjölander et Vansteelandt ont trouvé que l'estimateur proposé par Chen *et al.* était biaisé lorsque le modèle utilisé pour estimer la fonction de survie $S(t|Z,X)$ était mal spécifié. L'estimateur IPW était également biaisé lorsque les modèles utilisés pour estimer $S_C(t|Z,X)$ et $\pi(X)$ étaient mal spécifiés tandis que l'estimateur DR était sans biais dans tous les cas considérés [8]. L'estimateur proposé par Chen *et al.* avait un écart-type plus faible que ceux obtenus par IPW et DR. Lorsque les estimateurs du RA étaient sans biais, les taux de recouvrement obtenus étaient satisfaisants.

Enfin, pour l'approche paramétrique proposée par Laaksonen *et al.* [3], nous n'avons pas retrouvé d'étude de simulation pour évaluer la performance des estimateurs proposés. Ces auteurs présentent cependant un exemple numérique pour illustrer le biais quand on omet de prendre en compte les risques concurrents.

2.5.4 Définition et méthodes d'estimation avec la fonction de risque instantané

Plutôt que les fonctions de répartition, des auteurs [4, 5] interprètent les probabilités $\mathbf{P}(D)$ et $\mathbf{P}(D|\bar{E})$ comme la fonction de risque instantané au temps t et la fonction de risque instantané chez les non exposés, respectivement. Ces auteurs définissent ainsi la fonction de risque instantané attribuable :

$$\varphi(t) = \frac{\lambda(t) - \lambda(t|Z=0)}{\lambda(t)} = 1 - \frac{\lambda_0(t)}{\lambda(t)} \text{ où } \lambda_0(t) = \lambda(t|Z=0).$$

Comme avec les fonctions de survie, un estimateur naturel de φ consiste à remplacer $\lambda(t)$ et $\lambda_0(t)$ par leurs estimateurs respectifs $\hat{\lambda}(t)$ et $\hat{\lambda}_0(t)$:

$$\hat{\varphi}(t) = 1 - \frac{\hat{\lambda}_0(t)}{\hat{\lambda}(t)}.$$

Chen *et al.* [5] proposent d'estimer le risque de base $\lambda_0(t)$ par un estimateur de Breslow :

$$\hat{\lambda}_0(t) = \frac{\sum\limits_{i|t_i \leq t} dN_i(t)}{S^{(0)}(t;\hat{\beta})}$$

et le risque instantané marginal $\lambda(t)$ à partir de l'estimateur de Nelson-Aalen, c'est-à-dire

$$\hat{\lambda}(t) = \frac{\sum\limits_{i|t_i \leq t} dN_i(t)}{S^{*}(t)}$$

où $S^*(t) = \dfrac{1}{n}\sum\limits_{i=1}^{n} Y_i(t)$.

Un estimateur de $\varphi(t)$ est donc [5] :

$$\hat{\varphi}(t) = 1 - \frac{\hat{\lambda}_0(t)}{\hat{\lambda}(t)} = 1 - \frac{S^*(t)}{S^{(0)}(t;\hat{\beta})}.$$

Sous certaines conditions de régularité, Chen *et al.* [5] ont montré que $\sqrt{n}\,[\hat{\varphi}(t) - \varphi(t)]$ converge faiblement vers un processus gaussien de moyenne nulle. Un estimateur consistant de la fonction de variance-covariance entre les temps t et s est alors

$$\hat{\sigma}_\varphi(s,t) = n^{-1}\sum_{i=1}^{n} \hat{v}_i(s)\hat{v}_i(t),$$

où

$$\hat{v}_i(t) = \frac{S^*(t)\exp(\hat{\beta}^T Z_i)Y_i(t)}{S^{(0)}(t,\hat{\beta})^2} - \frac{Y_i(t)}{S^{(0)}(t,\hat{\beta})} + \frac{S^*(t)S^{(1)}(t,\hat{\beta})\hat{\Sigma}^{-1}(\hat{\beta})}{S^{(0)}(t,\hat{\beta})^2}\int_0^\tau \left[Z_i - \bar{Z}(u,\hat{\beta})\right] d\hat{M}_i(u),$$

$\hat{M}_i(t) = N_i(t) - \int\limits_0^t Y_i(u)\exp(\hat{\beta}^T Z_i)d\hat{\Lambda}_0(u)$ est le résidu martingale du modèle de Cox et $\hat{\Sigma}(\hat{\beta})$ est l'estimateur de la variance de $\hat{\beta}$.

Dans l'expression de $\hat{v}_i(t)$, on reconnaît que

$$U(\hat{\beta}) = \int_0^\tau \left[Z_i - \bar{Z}(u,\hat{\beta})\right] d\hat{M}_i(u)$$

est le vecteur du score. Ainsi,

$$\hat{v}_i(t) = \frac{S^*(t)\exp(\hat{\beta}^T Z_i)Y_i(t)}{S^{(0)}(t,\hat{\beta})^2} - \frac{Y_i(t)}{S^{(0)}(t,\hat{\beta})} + \frac{S^*(t)S^{(1)}(t,\hat{\beta})\hat{\Sigma}^{-1}(\hat{\beta})}{S^{(0)}(t,\hat{\beta})^2} \times U(\hat{\beta})$$

La variance de $\sqrt{n}\,[\hat{\varphi}(t) - \varphi(t)]$ est approximativement égale [5] à

$$\hat{\sigma}_\varphi(t)^2 = \hat{\sigma}_\varphi(t,t) = n^{-1}\sum_{i=1}^{n} \hat{v}_i(t)^2.$$

Chen *et al.* [5] ont proposé une étude de comparaison numérique des RA $A(t)$ et $\varphi(t)$ définis à partir des fonctions de répartition et des fonctions de risque instantané respectivement dans le cas d'un modèle à risques proportionnels pour différentes valeurs de risque de base et de probabilité d'exposition et un paramètre β fixé à $\ln(2)$ (Figure 2.3). Comme d'autres auteurs [3, 4, 6], pour un RR constant, Chen *et al.* [6] ont montré que le RA n'est pas constant mais décroissant dans le temps, même lorsque le risque de base et la probabilité d'exposition sont constants. Dans le cas d'un risque de base important, les fonctions de risque attribuable décroissent rapidement au cours du premier quart de suivi tandis que, dans le cas où le risque de base est faible, elles décroissent plus régulièrement lorsque la maladie est plus (respectivement moins) fréquente chez les sujets non exposés, le risque attribuable à l'exposition a tendance à changer plus (respectivement moins) rapidement dans le temps. En comparant $\varphi(t)$ avec $A(t)$, on constate que $\varphi(t)$ se rapproche d'autant mieux de $A(t)$ que la maladie est moins fréquente et que la prévalence de l'exposition est faible [5].

Chen *et al.* [5] ont aussi réalisé une étude de simulation pour évaluer les performances de leur estimateur. Nous l'avons reproduite en annexe B.

2.5.5 Autres définitions

D'autres définitions dans le contexte de l'analyse de survie ont également été proposées.

La survie attribuable (AS pour *attributable survival*) a été proposée par Cox *et al.* [7] et mesure la proportion de sujets n'ayant pas développé l'événement jusqu'à t (entre 0 et t) si l'exposition avait été éliminée et qui auraient développé l'événement dans ce même intervalle de temps s'ils avaient été exposés,

$$AS(t) = \frac{S_0(t) - S(t)}{S_0(t)}.$$

Cox *et al.* [7] ne proposent pas de formule explicite pour l'écart-type associé à leur estimateur $\widehat{AS}(t)$. Ils suggèrent en discussion de leur article de s'appuyer sur la delta méthode dans le cas d'estimateurs paramétriques et sur la formule de Greenwood dans le

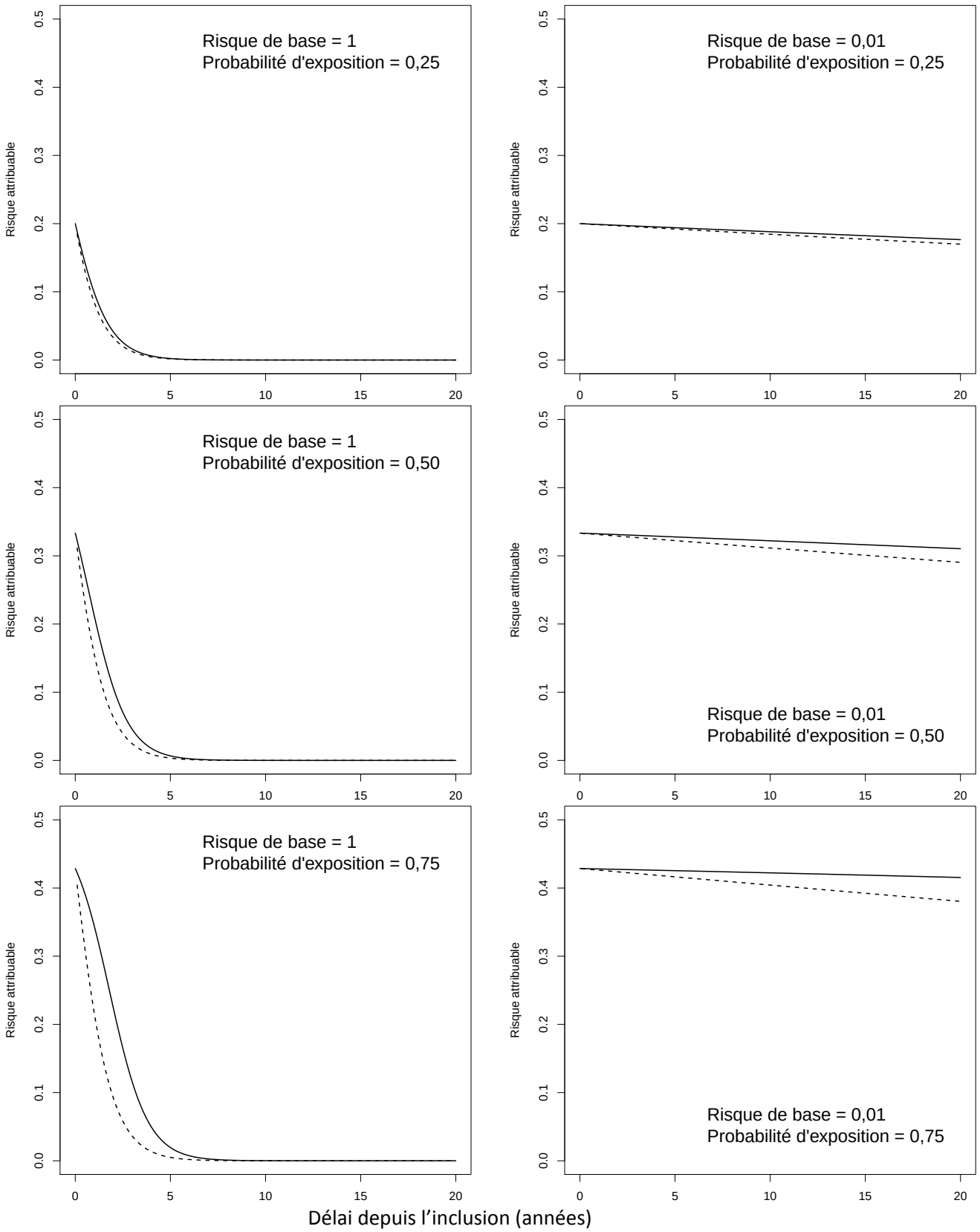


FIGURE 2.3: Comparaison entre risque attribuable A (en pointillé) et φ (en trait plein) dans le cas d'un modèle à risques proportionnels avec un risque de base constant égal à 1 (à gauche) ou 0,01 (à droite) et β fixé à ln(2) pour différentes valeurs de probabilité d'exposition 0,25, 0,50 et 0,75

cas d'estimateurs non paramétriques, ou encore de procéder par *bootstrap*. Nous proposons une expression de la variance d'AS (cf. annexe C) en nous référant aux travaux de Chen *et al.* [6].

Le temps de survie attribuable (AST pour *attributable survival time*) a également été proposé par Cox *et al.* [7] et mesure la proportion de temps gagné grâce à l'élimination de l'exposition sur le temps attendu en l'absence d'exposition :

$$AST(t) = \frac{E_0(t) - E(t)}{E_0(t)} \tag{2.14}$$

où $E_k(t) = \int_0^t S_k(u)du$ pour $k = 0, 1$, $E(t) = pE_1(t) + (1-p)E_0(t)$ et $S_1(t) = S(t|Z=1)$.

Ces auteurs ont également proposé une généralisation des mesures AR, AS et AST dans le cas où l'élimination de l'exposition est partielle et a lieu en un temps $t > 0$ [7].

Samuelsen et Eide [4] quant à eux définissent la fraction de risque attribuable (AHF pour *attributable hazard fraction*) qui s'interprète comme la limite de la proportion d'événements évités dans l'intervalle $[t, t+\Delta)$ quand la longueur de l'intervalle tend vers zéro. Ils utilisent pour cela la moyenne du risque instantané conditionnel :

$$AHF(t) = \frac{\mathbf{E}\left[\lambda(t|Z)\right] - \mathbf{E}\left[\lambda(t|Z^*)\right]}{\mathbf{E}\left[\lambda(t|Z)\right]}$$

où $\mathbf{E}\left[\lambda(t|Z)\right] = \int \lambda(t|z)dP(z)$ et $\mathbf{E}\left[\lambda(t|Z^*)\right] = \int \lambda(t|z^*)dP^*(z)$ avec $P(z)$ la distribution initiale de Z et $P^*(z)$ sa distribution modifiée après intervention, souvent appelée distribution contrefactuelle.

Dans le cas d'un modèle à risques proportionnels, cette mesure ne dépend plus du temps et peut s'écrire comme suit [4] :

$$AHF = \frac{\mathbf{E}\left[\exp(\beta^T Z)\right] - \mathbf{E}\left[\exp(\beta^T Z^*)\right]}{\mathbf{E}\left[\exp(\beta^T Z)\right]}.$$

Dans le cas d'une exposition binaire avec une probabilité d'exposition égale à p, elle s'écrit:

$$AHF = \frac{p\left[\exp(\beta) - 1\right]}{p\left[\exp(\beta) - 1\right] + 1}.$$

Enfin, Laaksonen *et al.* [73] proposent une variante du PAF pour la prévalence de la maladie :

$$PAF(PM_t) = \frac{PM_t(Z) - PM_t(Z^*)}{PM_t(Z)}$$

où $PM_t(Z) = \dfrac{\sum_{i=1}^{n} \mathbf{P}\{T_i^M < t < T_i^D | Z_i\}}{\sum_{i=1}^{n} \mathbf{P}\{\min(T_i^D, T_i^M) > t | Z_i\} + \mathbf{P}\{T_i^M < t < T_i^D | Z_i\}}$.

Un estimateur de cette mesure est proposé dans le cas d'un modèle paramétrique à risque de base constant par morceaux : [73] :

$$PM_t(Z) = \frac{\sum_{i=1}^{n} S_{ik}^{D_2} \sum_{j=1}^{k} S_{i,j-1}^{M} \frac{S_{i,j-1}^{D_1}}{S_{i,j-1}^{D_2}} \frac{\lambda_{ij}^{M}}{\lambda_{ij}^{M}+\lambda_{ij}^{D_1}-\lambda_{ij}^{D_2}} \left\{1 - \exp\left[-(\lambda_{ij}^{M} + \lambda_{ij}^{D_1} - \lambda_{ij}^{D_2})(a_j - a_{j-1})\right]\right\}}{\sum_{i=1}^{n} S_{ik}^{M} S_{ik}^{D_1} + S_{ik}^{D_2} \sum_{j=1}^{k} S_{i,j-1}^{M} \frac{S_{i,j-1}^{D_1}}{S_{i,j-1}^{D_2}} \frac{\lambda_{ij}^{M}}{\lambda_{ij}^{M}+\lambda_{ij}^{D_1}-\lambda_{ij}^{D_2}} \left\{1 - \exp\left[-(\lambda_{ij}^{M} + \lambda_{ij}^{D_1} - \lambda_{ij}^{d_2})(a_j - a_{j-1})\right]\right\}}$$

2.5.6 Logiciels disponibles

Dans la dernière décennie, plusieurs programmes utilisant les logiciels statistiques ont été proposés afin de faciliter et d'encourager l'estimation du RA dans les études de cohorte.

En ce qui concerne le logiciel SAS, Laaksonen *et al.* [80] ont proposé un ensemble de macros permettant d'estimer le RA dans le cas d'un modèle à risques proportionnels avec un risque de base constant par morceaux [80] avec la possibilité de stratifier sur la cohorte de naissance et de tenir compte des événements concurrents.

Spiegelman *et al.* [31] ont aussi proposé une macro qui permet de calculer le RA.

Sous le logiciel `R`, les packages `epiR` [81], `Attribrisk` [82], `paf` [75] et `AF` [83, 84] ont été développés pour estimer le RA.

Le package `epiR` utilise la fonction epi.2by2 pour estimer le RA mais ne permet pas l'ajustement sur des facteurs de confusion.

Le package `Attribrisk` permet l'ajustement sur les facteurs de confusion par une régression logistique et est essentiellement limité aux études cas-témoins.

Les packages `paf` et `AF` développés respectivement par Chen [75] et Dahlqwist *et al.* [83] permettent l'ajustement sur les facteurs de confusion en utilisant un modèle de Cox à risques proportionnels [6, 8].

Pour l'estimation de la FP, aucun logiciel n'est disponible à ce jour.

Chapitre 3

Comparaison de méthodes pour l'estimation du risque attribuable dans le contexte de l'analyse de survie

3.1 Problématique et objectifs

Dans cette étude, nous avons retenu la première définition du RA basée sur les fonctions de répartition (équation 2.12). Cette définition est plus conforme à la définition standard du RA et est la plus utilisée dans la littérature. Comme nous l'avons vu, plusieurs méthodes d'estimation du RA ont été proposées dans ce cas [3, 6, 8]. Quelques études de simulation ont été réalisées pour les approches non paramétriques et semi-paramétriques mais les performances de ces différentes approches n'ont pas été systématiquement comparées.

Dans un premier temps, nous avons cherché à reproduire les résultats publiés de façon à valider l'implémentation des méthodes (cf. annexe A). Dans un deuxième temps, nous nous sommes inspirés des plans de simulation des publications antérieures pour réaliser notre propre étude de simulation.

L'objectif de ce travail était de comparer les méthodes d'estimation du RA défini à partir des fonctions de répartition. Nous avons effectué une comparaison systématique à l'aide d'une étude de simulation que nous détaillons pour commencer. Les méthodes ont ensuite été appliquées aux données de la cohorte E3N (Étude Épidémiologique auprès de Femmes de la Mutuelle Générale de l'Éducation Nationale) [85] dans le but d'estimer la proportion de cas de cancer du sein attribuable à l'utilisation de traitements hormonaux de la ménopause (THM). Une publication issue de la cohorte E3N a estimé qu'une proportion de 14,5 % des cas de cancer du sein est attribuable à une utilisation récente des THM après 15 ans de suivi [86].

3.2 Méthodes de simulation

3.2.1 Génération de l'exposition

Nous considérons une variable d'exposition Z unique, binaire et fixe dans le temps. Elle est simulée suivant une loi de Bernoulli de paramètre p. Le paramètre de régression β correspondant est donc unidimensionnel.

3.2.2 Génération des temps d'événement

Nous avons utilisé la méthode de la fonction de répartition inverse (dite aussi méthode de l'anamorphose) pour générer des temps d'événement. Cette méthode s'appuie sur le résultat suivant: *Soit X une variable aléatoire réelle de fonction de répartition F_X; on définit sa fonction de répartition inverse par* $(\forall y \in [0,1])\ F_X^{-1}(y) = \inf\{x | F_X(x) \geq y\}$. *Alors, si U suit une loi uniforme sur* $[0,1]$, *on montre que* $V = F_X^{-1}(U)$ *a la même fonction de répartition que X.*

En effet, pour tout $x \in \mathbb{R}$, la fonction de répartition de V est définie par:
$F_V(x) = \mathbf{P}(V \leq x) = \mathbf{P}[F_X^{-1}(U) \leq x] = \mathbf{P}\left[U \leq F_X(x)\right] = F_X(x)$. Dans un premier temps,

nous avons considéré un modèle à risques proportionnels $\lambda(t|Z) = \lambda_0(t)\exp(\beta Z)$ où $\lambda_0(t)$

est le risque de base et β le paramètre de régression. Nous avons considéré un risque de base issu d'une loi de Weibull de paramètres d'échelle θ et de forme γ : $\lambda_0(t) = \gamma\theta^{-\gamma}t^{\gamma-1}$. Dans ce cas, la variable aléatoire T est simulée conditionnellement à la covariable Z selon la loi :

$$T|Z \sim \theta\sqrt[\gamma]{\frac{-\ln(U)}{\exp(\beta Z)}}$$

où U suit une loi uniforme sur $[0, 1]$.

Dans un deuxième temps, nous avons considéré un modèle à risque non proportionnels $\Lambda(t|Z) = G\left[\lambda_0 t \exp(\beta Z)\right]$, où λ_0 est constant et G est la transformation logarithmique : $G(t) = \ln(1 + 2t)/2$ [6]. Dans ce cas, la variable aléatoire T est simulée suivant la loi :

$$T|Z \sim \frac{G^{-1}\left[-\ln(U)\right]}{\lambda_0 \exp(\beta Z)} \text{ où } G^{-1}(t) = \frac{\exp(2t) - 1}{2}.$$

3.2.3 Génération des temps de censure

Nous avons généré une censure C indépendante de la covariable Z selon une loi uniforme sur $[0, \tau]$ où τ est la durée maximale de l'étude.

3.2.4 Critères de comparaison

Soit m le nombre d'échantillons simulés. Pour chaque estimateur du RA, $\hat{A}(t)$, en des temps donnés t préalablement choisis, nous avons calculé :

— la moyenne empirique des estimations du RA

$$\bar{\hat{A}}(t) = \frac{1}{m}\sum_{k=1}^{m}\hat{A}_k(t),$$

— le biais moyen par rapport à la valeur théorique $A(t)$

$$b(t) = \frac{1}{m}\sum_{k=1}^{m}\left[\hat{A}_k(t) - A(t)\right] = \bar{\hat{A}}(t) - A(t),$$

— l'écart-type empirique SSD (pour *Sampling Standard Deviation*)

$$\sqrt{\hat{\text{Var}}\left[\hat{A}(t)\right]} = \sqrt{\frac{1}{m-1}\sum_{k=1}^{m}\left[\hat{A}_k(t) - \bar{\hat{A}}(t)\right]^2},$$

— l'écart-type estimé moyen ou SEE (pour *Standard Error Estimator*) de l'estimateur du RA obtenu comme la moyenne empirique, sur m jeux de données simulés, des écarts-types estimés $\hat{\sigma}_k(t)$, $k = 1, ..., m$:

$$\bar{\hat{\sigma}}_k(t) = \frac{1}{m} \sum_{k=1}^{m} \hat{\sigma}_k(t),$$

— la probabilité de couverture (PC ou taux de recouvrement) : proportion des itérations pour lesquelles la valeur théorique appartient à l'intervalle de confiance du RA :

$$IC_{1-\alpha}\left[A(t)\right] = \left[\hat{A}_k(t) \pm z_{1-\frac{\alpha}{2}} \times \hat{\sigma}_k(t)\right]$$

où $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite. Pour $\alpha = 5\,\%$ et un intervalle de confiance à 95 %, le quantile d'ordre $1 - \frac{\alpha}{2}$ est égal à 1,96. Dans ce cas, pour $m = 1\,000$ simulations, on s'attend à un taux de couverture compris entre 0,936 et 0,963 $\left(0{,}95 \pm 1{,}96\sqrt{0{,}95 \times 0{,}05/1\,000}\right)$.

Nous avons estimé le RA et son intervalle de confiance en utilisant quatre méthodes : deux méthodes non paramétriques, l'une où $S_0(.)$ et $S(.)$ sont estimées en utilisant la méthode de Kaplan-Meier (KM), l'autre où $S_0(.)$ et $S(.)$ sont estimées par la méthode de Kaplan-Meier et la méthode de Kaplan-Meier pondérée respectivement (KMP) [6] ; une méthode semi-paramétrique basée sur le modèle de Cox à risques proportionnels (COX) [6] et une méthode paramétrique utilisant un modèle à risque de base constant par morceaux [3] en considérant quatre intervalles de 5 ans (RCM). Lorsque nous n'avons eu aucun événement dans le dernier intervalle, le jeu de données simulé a été éliminé et remplacé par un nouveau.

3.2.5 Choix des paramètres

Nous avons généré 1 000 jeux de données de 1 000 ou 10 000 observations indépendantes chacun. Nous avons choisi le nombre de 1 000 simulations après avoir vérifié la vitesse de convergence des écarts-types estimés moyens SEE pour les quatre méthodes à chaque

temps considéré (résultats non montrés). Dans notre étude de simulation, nous avons fait varier la probabilité d'exposition en considérant les valeurs $p = 0{,}25$, 0,50 et 0,75.

Dans le cas du modèle à risques proportionnels, nous avons exploré trois situations. Nous avons considéré un risque de base constant ($\gamma = 1$), puis dépendant du temps, croissant ($\gamma = 4/3$) ou décroissant ($\gamma = 3/4$). Le paramètre d'échelle θ a été choisi pour avoir une médiane de survie $\theta \ln(2)^{1/\gamma}$ de 15 ans la durée maximale de suivi, τ, étant fixée à 20 ans. Le paramètre β du modèle de Cox a été fixé à $\ln(2)$ ou 0, soit un RR instantané de 2 ou 1 respectivement. Nous avons aussi exploré le cas où $\gamma = 1/2$ mais avons rencontré des problèmes de convergence pour l'approche paramétrique. En effet la plupart des événements se produisaient dans le premier intervalle et par conséquent le choix des intervalles de pas régulier de 5 ans n'était pas adapté pour ce cas.

Le RA théorique peut être calculé en fonction des paramètres choisis et du temps de suivi. Ainsi, pour les trois modèles à risques proportionnels considérés,

$$\begin{aligned} A(t) &= \frac{S_0(t) - S(t)}{1 - S(t)} \\ &= \frac{\exp\left[-\left(\frac{t}{\theta}\right)^{\gamma}\right] - (1-p)\exp\left[-\left(\frac{t}{\theta}\right)^{\gamma}\right] - p\exp\left[-\exp(\beta)\left(\frac{t}{\theta}\right)^{\gamma}\right]}{1 - (1-p)\exp\left[-\left(\frac{t}{\theta}\right)^{\gamma}\right] - p\exp\left[-\exp(\beta)\left(\frac{t}{\theta}\right)^{\gamma}\right]} \\ &= \frac{p\left\{\exp\left[-\left(\frac{t}{\theta}\right)^{\gamma}\right] - \exp\left[-\exp(\beta)\left(\frac{t}{\theta}\right)^{\gamma}\right]\right\}}{1 - \exp\left[-\left(\frac{t}{\theta}\right)^{\gamma}\right] + p\left\{\exp\left[-\left(\frac{t}{\theta}\right)^{\gamma}\right] - \exp\left[-\exp(\beta)\left(\frac{t}{\theta}\right)^{\gamma}\right]\right\}}. \end{aligned}$$

Dans le cas d'un modèle à risques non proportionnels, le paramètre λ_0 a été fixé à 0,1 et choisi pour avoir également une médiane de survie $\dfrac{3}{2\lambda_0}$ à 15 ans. Le paramètre β a été

fixé à ln(2). Le RA théorique devient :

$$\begin{aligned}
A(t) &= \frac{S_0(t) - S(t)}{1 - S(t)} \\
&= \frac{S_0(t) - (1-p)S_0(t) - pS_1(t)}{1 - (1-p)S_0(t) - pS_1(t)} \\
&= \frac{p(S_0(t) - S_1(t))}{1 - (1-p)S_0(t) - pS_1(t)} \\
&= \frac{p\left\{\exp\left\{-\dfrac{\ln\left[1+2\lambda_0 t\right]}{2}\right\} - \exp\left\{-\dfrac{\ln\left[1+2\lambda_0 t\exp(\beta)\right]}{2}\right\}\right\}}{1-(1-p)\exp\left\{-\dfrac{\ln\left[1+2\lambda_0 t\right]}{2}\right\} - p\exp\left\{-\dfrac{\ln\left[1+2\lambda_0 t\exp(\beta)\right]}{2}\right\}}.
\end{aligned}$$

Pour une probabilité d'exposition égale à 0,50, les figures 3.1 et 3.2 représentent les fonctions de survie et le RA théoriques respectivement pour un modèle à risques proportionnels (paramètre de forme $\gamma = 3/4, 1$ et $4/3$) et un modèle à risques non proportionnels ($\lambda_0 = 0{,}1$).

Comme l'ont montré d'autres auteurs [5, 6], le RA, dans le contexte de l'analyse de survie avec un RR constant, est une fonction décroissante du temps et varie entre 33,3 % et 9,3 % pour une probabilité d'exposition à 0,50. Avec le modèle à risques non proportionnels, il décroît plus vite dans le premier quart de suivi et reste inférieur au RA théorique avec un modèle à risques proportionnels.

La figure 3.3 montre que le rapport des risques instantanés chez les exposés et chez les non exposés, pour le modèle à risques non proportionnels considéré, est décroissant de 2 à 1 au cours du suivi.

Les tableaux 3.1 et 3.2 présentent les pourcentages de censure moyens dans nos données simulées sous les modèles à risques proportionnels et non proportionnels, respectivement. Les pourcentages de censure moyens sont compris entre 47,1 % et 67,6 % (étendue, 42,0 % à 72,9 %) dans nos simulations.

3.2.6 Analyse des résultats

Les résultats sont présentés aux temps $t = \tau/4$, $\tau/2$, $3\tau/4$ et τ (respectivement 5, 10, 15 et 20 ans). Pour les approches non paramétriques et semi-paramétrique, les estimations

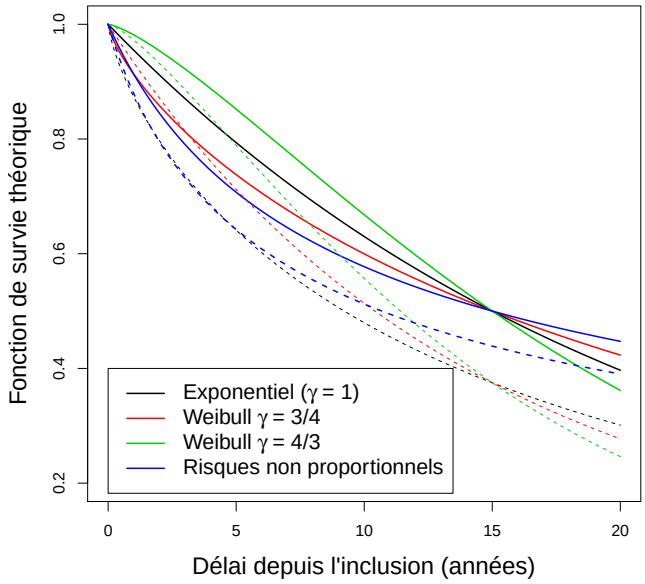


FIGURE 3.1: Fonction de survie théorique marginale (trait pointillé) et chez les non exposés (trait plein) pour différents modèles de génération des données à risques proportionnels et non proportionnels pour $\beta = \ln(2)$ et une probabilité d'exposition de 0,50

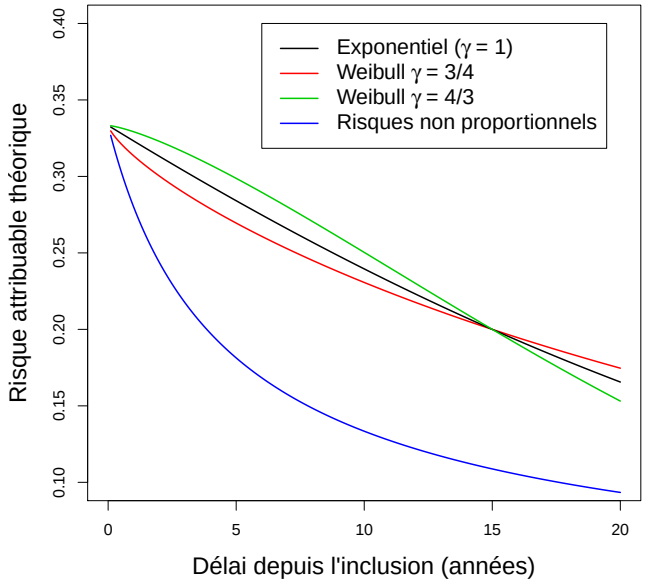


FIGURE 3.2: Risque attribuable théorique pour $\beta = \ln(2)$, une probabilité d'exposition de 0,50 et pour différents modèles de génération des données

Taille d'échantillon n	γ	β	Probabilité d'exposition p	% de censure moyen (étendue)
1 000	3/4	ln(2)	0,25	57,7 (52,8 – 63,1)
1 000	1	ln(2)	0,25	60,4 (55,6 – 65,5)
1 000	4/3	ln(2)	0,25	63,1 (585 – 68,6)
1 000	3/4	ln(2)	0,50	52,3 (47,6 – 57,7)
1 000	1	ln(2)	0,50	55,5 (50,8 – 61,4)
1 000	4/3	ln(2)	0,50	58,6 (54,1 – 64,2)
1 000	3/4	ln(2)	0,75	47,1 (42,0 – 52,8)
1 000	1	ln(2)	0,75	50,5 (44,8 – 56,6)
1 000	4/3	ln(2)	0,75	54,0 (48,1 – 59,7)
1 000	3/4	0	0,25	62,9 (58,0 – 69,4)
1 000	1	0	0,25	65,3 (60,0 – 71,2)
1 000	4/3	0	0,25	67,6 (63,1 – 72,9)
1 000	3/4	0	0,50	62,9 (58,0 – 69,4)
1 000	1	0	0,50	65,3 (60,0 – 71,2)
1 000	4/3	0	0,50	67,6 (63,1 – 72,9)
1 000	3/4	0	0,75	62,9 (58,0 – 69,4)
1 000	1	0	0,75	65,3 (60,0 – 71,2)
1 000	4/3	0	0,75	67,6 (63,1 – 72,9)
10 000	3/4	ln(2)	0,25	57,6 (56,1 – 59,1)
10 000	1	ln(2)	0,25	60,3 (58,9 – 61,9)
10 000	4/3	ln(2)	0,25	63,1 (61,7 – 64,5)
10 000	3/4	ln(2)	0,50	52,3 (50,7 – 53,7)
10 000	1	ln(2)	0,50	55,4 (53,8 – 56,9)
10 000	4/3	ln(2)	0,50	58,6 (57,1 – 60,0)
10 000	3/4	ln(2)	0,75	47,1 (45,3 – 48,8)
10 000	1	ln(2)	0,75	50,5 (49,0 – 52,3)
10 000	4/3	ln(2)	0,75	54,1 (52,5 – 55,9)
10 000	3/4	0	0,25	62,9 (61,5 – 64,6)
10 000	1	0	0,25	65,3 (63,9 – 66,8)
10 000	4/3	0	0,25	67,6 (66,1 – 69,1)
10 000	3/4	0	0,50	62,9 (61,5 – 64,6)
10 000	1	0	0,50	65,3 (63,9 – 66,8)
10 000	4/3	0	0,50	67,6 (66,1 – 69,1)
10 000	3/4	0	0,75	62,9 (61,5 – 64,6)
10 000	1	0	0,75	65,3 (63,9 – 66,8)
10 000	4/3	0	0,75	67,6 (66,1 – 69,1)

Tableau 3.1: Pourcentage de censure dans nos données simulées pour des modèles à risques proportionnels

Taille d'échantillon n	λ_0	Probabilité d'exposition p	% de censure moyen (étendue)
1 000	0,1	0,50	55,9 (51,1 – 60,9)
10 000	0,1	0,50	55,9 (54,0 – 57,2)

Tableau 3.2: Pourcentage de censure dans nos données simulées pour des modèles à risques non proportionnels

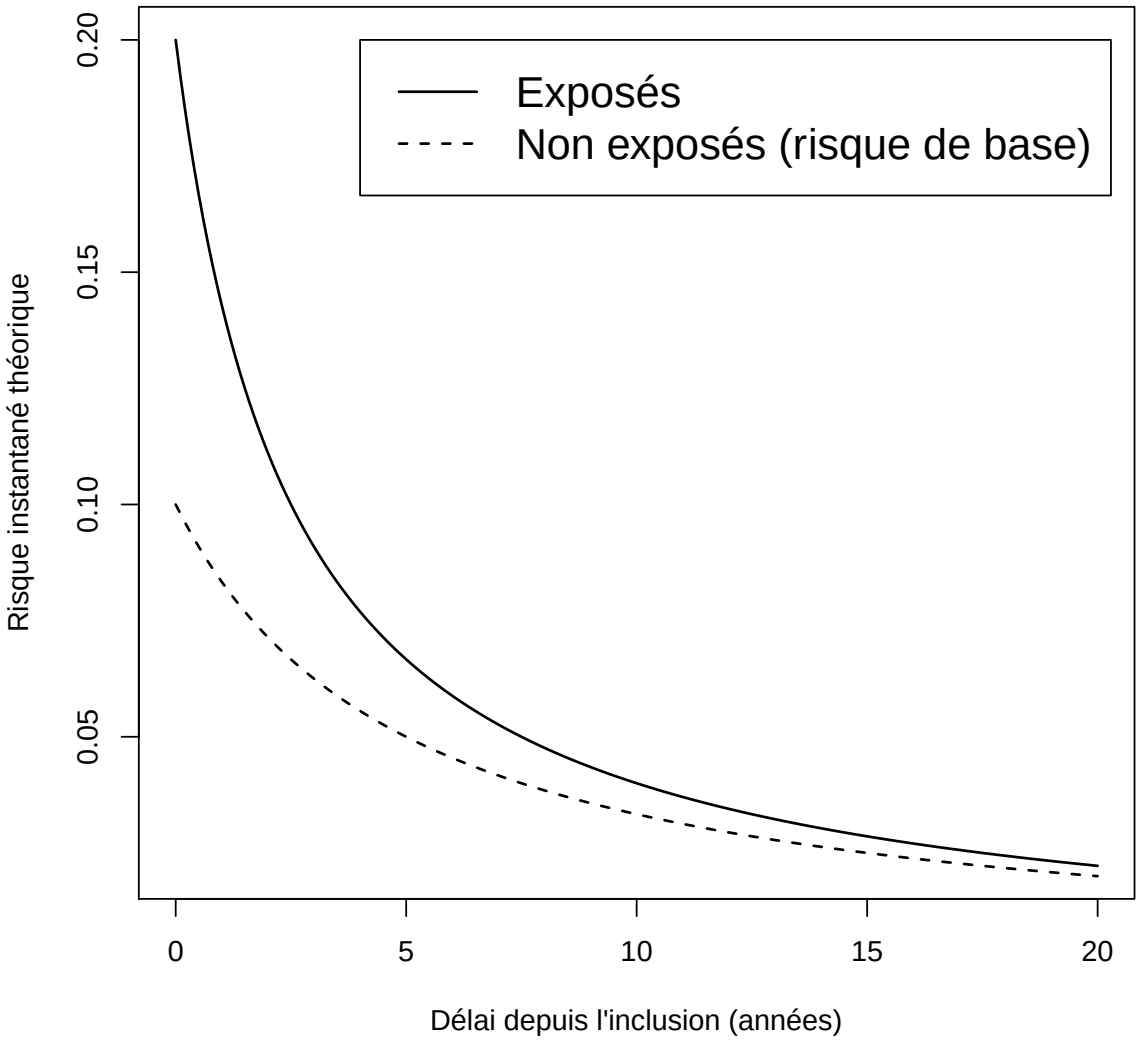


FIGURE 3.3: Risque instantané théorique chez les exposés et non exposés pour un modèle à risques non proportionnels pour $\beta = \ln(2)$, $p = 0{,}50$ et $\lambda_0 = 0{,}1$

ont été obtenues à des temps réellement observés ; nous avons donc considéré les valeurs prises au temps antérieur le plus proche des temps $t = \tau/4, \tau/2, 3\tau/4$ et τ.

Bien que des auteurs [3, 6] aient suggéré d'utiliser la transformation logarithmique $\ln\{1 - A(t)\}$ pour améliorer les probabilités de couverture dans le cas l'échantillon est de petite taille, cela n'a pas amélioré la probabilité de couverture dans nos résultats. Les résultats sont donc présentés sans avoir appliqué de transformation.

Les estimations par les méthodes non paramétriques se basent sur les informations disponibles jusqu'au temps d'intérêt tandis que les approches semi-paramétrique et paramétrique utilisent toute l'information disponible dans le suivi complet (jusqu'à τ) pour estimer le paramètre β. Pour permettre une comparaison plus juste sous l'hypothèse des risques proportionnels, nous avons également calculé les estimateurs du RA par les méthodes semi-paramétrique et paramétrique après avoir censuré à la moitié et au quart de la durée de suivi.

Les données ont été simulées avec le logiciel R version 3.0.1. Nous avons codé les méthodes non paramétriques en utilisant le logiciel R et testé la validité de notre code en comparant nos résultats de simulation avec ceux de Chen *et al.* [6] qui ont proposé ces approches en utilisant les mêmes paramètres (cf. annexe A). Pour la méthode semi-paramétrique, nous avons utilisé le package `paf` développé par Chen [75]. Pour la méthode paramétrique, nous avons utilisé le logiciel SAS version 9.3 et un ensemble de macros développé par Laaksonen *et al.* [80] pour le calcul du RA.

3.3 Résultats de simulation

3.3.1 Risques proportionnels avec un risque de base constant

Nous considérons pour commencer le cas des risques proportionnels entre les exposés et les non exposés avec $\beta = \ln(2)$, une probabilité d'exposition à 0,50 et un risque de base constant. Avec une taille d'échantillon de 1 000 observations (Tableau 3.3, partie gauche), nous avons un biais légèrement plus important à la fin du suivi en τ qu'aux

temps précédents, particulièrement pour les méthodes non paramétriques KM et KMP (dans une moindre mesure), mais les estimateurs du RA sont pratiquement sans biais (biais relatif $< 2{,}5\,\%$) pour toutes les méthodes d'estimation. Les variances estimées reflètent fidèlement la vraie variation (écart-type estimé moyen proche de l'écart-type empirique) et les intervalles de confiance à 95 % ont des probabilités de couverture appropriées sauf en τ pour les méthodes non paramétriques KM et KMP où la variance est quelque peu sous-estimée par rapport à la variance empirique, ce qui conduit à un taux de recouvrement plus bas comparé au taux de recouvrement attendu en τ. Les estimations des méthodes paramétrique RCM et semi-paramétrique COX sont plus précises que des méthodes non paramétriques, particulièrement aux temps $\tau/4$ et τ. Les estimations du paramètre β par les méthodes paramétrique et semi-paramétrique sont sans biais (biais relatif $< 0{,}7\,\%$, Tableau 3.4), tout comme l'estimation du risque de base par la méthode paramétrique RCM semble être satisfaisante (Figure 3.4).

Lorsque nous considérons une taille d'échantillon de 10 000 observations (Tableau 3.3, partie droite), les biais des estimateurs du RA deviennent plus faibles comparés à ceux obtenus avec une taille d'échantillon plus petite pour toutes les méthodes d'estimation (biais relatif $< 0{,}7\,\%$). Comme attendu, la précision augmente nettement pour toutes les méthodes d'estimation du RA d'un facteur d'environ $\sqrt{10}$. De plus, les écarts-types estimés moyens SEE et empiriques SSD sont très proches, même en τ pour les méthodes non paramétriques, avec de bons taux de recouvrement compris entre 0,94 et 0,96. L'estimation du risque de base par l'approche paramétrique RCM est satisfaisante et le paramètre β estimé par les méthodes semi-paramétrique COX et paramétrique RCM est également sans biais (bais relatif $< 0{,}04\,\%$, Tableau 3.5).

3.3.2 Risques proportionnels avec un risque de base non constant

Des résultats similaires sont observés avec un risque de base décroissant (Tableau 3.6). Lorsque $\gamma = 3/4$, les biais sont en effet proches de ceux obtenus pour un risque de base constant à l'exception d'une augmentation modérée pour l'approche paramétrique RCM

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	0,001584	0,052440	0,052591	0,949	−0,000011	0,016622	0,016349	0,944
	$\tau/2$	0,001496	0,039210	0,039099	0,948	0,000235	0,012434	0,012420	0,944
	$3\tau/4$	0,001100	0,035666	0,035948	0,946	−0,000333	0,011353	0,011354	0,949
	τ	0,004047	0,043238	0,053015	0,912	0,001025	0,017251	0,019598	0,943
KMP	$\tau/4$	0,001594	0,052516	0,052483	0,949	0,000003	0,016613	0,016357	0,946
	$\tau/2$	0,001541	0,039144	0,038926	0,950	0,000285	0,012401	0,012398	0,946
	$3\tau/4$	0,001093	0,035402	0,035479	0,953	−0,000286	0,011283	0,011297	0,952
	τ	0,002922	0,040635	0,048602	0,902	0,000497	0,016646	0,018245	0,942
COX	$\tau/4$	0,000977	0,038843	0,038208	0,958	−0,000136	0,012292	0,012206	0,956
	$\tau/2$	0,001108	0,033847	0,033524	0,951	0,000006	0,010700	0,010616	0,958
	$3\tau/4$	0,001031	0,029264	0,028893	0,958	−0,000081	0,009237	0,009253	0,954
	τ	0,002577	0,027146	0,027753	0,946	0,000148	0,008965	0,009087	0,950
RCM	$\tau/4$	0,001356	0,038338	0,038248	0,952	−0,000086	0,012120	0,012209	0,953
	$\tau/2$	0,001372	0,033380	0,033529	0,948	0,000034	0,010543	0,010608	0,952
	$3\tau/4$	0,001113	0,028804	0,028870	0,957	−0,000081	0,009088	0,009263	0,952
	τ	0,001564	0,025811	0,025420	0,961	−0,000154	0,008105	0,008153	0,952

Tableau 3.3: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base constant ($\gamma = 1$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,5. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}284$, $A(\tau/2) = 0{,}240$, $A(3\tau/4) = 0{,}200$ et $A(\tau) = 0{,}166$.

Méthode	Valeur théorique	Minimum	Médiane	Moyenne	Maximum	SEE	SSD	IC empirique
Paramétrique	ln(2)	0,378	0,697	0,698	1,018	0,098	0,096	0,515 – 0,886
Semi-paramétrique	ln(2)	0,379	0,696	0,697	1,026	0,098	0,096	0,515 – 0,883

Tableau 3.4: Paramètre β estimé par les méthodes semi-paramétrique COX et paramétrique RCM avec les paramètres $n = 1\,000$, $\gamma = 1$ et $p = 0{,}50$

Méthode	Valeur théorique	Minimum	Médiane	Moyenne	Maximum	SEE	SSD	IC empirique
Paramétrique	ln(2)	0,608	0,693	0,693	0,791	0,031	0,030	0,635 – 0,752
Semi-paramétrique	ln(2)	0,608	0,693	0,693	0,791	0,031	0,030	0,635 – 0,751

Tableau 3.5: Paramètre β estimé par les méthodes semi-paramétrique COX et paramétrique RCM avec les paramètres $n = 10\,000$, $\gamma = 1$ et $p = 0{,}50$

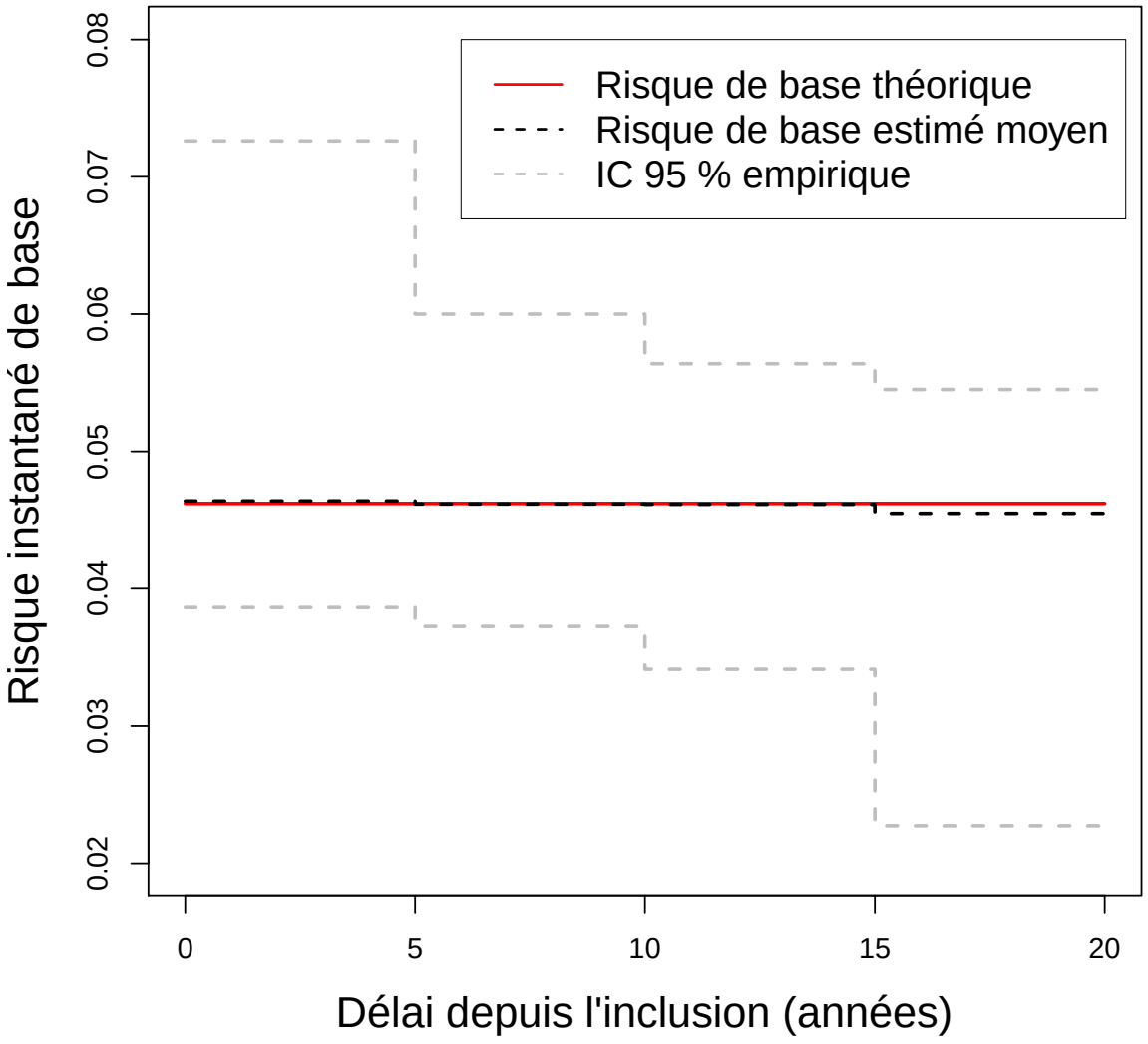


FIGURE 3.4: Estimation moyenne sur les 1 000 simulations du risque instantané de base par la méthode paramétrique RCM pour une durée de suivi divisée en quatre intervalles réguliers pour $\gamma = 1$, $n = 1\,000$ et $p = 0{,}50$

avec les deux tailles d'échantillon considérées (biais relatif $< 2{,}4\,\%$). Cette augmentation est peut-être due au fait que le modèle paramétrique a du mal à estimer un risque de base constant dans le premier intervalle alors que la fonction théorique n'est pas constante et décroît très vite dans cet intervalle (Figure 3.5). Malgré cela, les taux de recouvrement restent satisfaisants pour la méthode paramétrique comme pour les autres méthodes, sauf encore en τ pour les deux approches non paramétriques KM et KMP et une taille d'échantillon de 1 000 observations (0,915 et 0,906 pour KM et KMP respectivement).

Dans le cas d'un risque de base croissant ($\gamma = 4/3$, Tableau 3.7), les taux de recouvrement en τ pour les deux méthodes non paramétriques et $n = 1\,000$ sont encore plus mauvais (0,891 et 0,898 pour KM et KMP respectivement) avec un écart important entre les écarts-types estimés moyens SEE et empiriques SSD. Ce faible recouvrement peut s'expliquer par l'augmentation des biais en τ pour les deux approches non paramétriques KM et KMP avec un risque de base croissant comparé à un risque de base constant ou décroissant. Les résultats sont satisfaisants autrement pour l'approche paramétrique RCM et les biais des estimateurs du RA, de β et du risque de base (Figure 3.6) sont comparables à ceux obtenus avec un risque de base constant malgré une mauvaise spécification du modèle.

3.3.3 Influence de la probabilité d'exposition

Avec un risque de base constant ($\gamma = 1$) et une plus faible probabilité d'exposition à 0,25 (Tableau 3.8), les taux de recouvrement s'améliorent en τ pour les approches non paramétriques KM et KMP mais restent inférieurs à 93 % pour les échantillons de taille plus petite $n = 1\,000$ (Tableau 3.8, partie gauche). Les écarts-types estimés moyens sont inférieurs à ceux estimés lorsque la probabilité d'exposition est égale à 0,50 avec un risque de base constant et une taille d'échantillon de 1 000 observations. Pour $n = 10\,000$ (Tableau 3.8, partie droite), les biais restent faibles avec de bons taux de recouvrement sauf en τ pour la méthode non paramétrique KMP où la probabilité de couverture est égale à 0,926.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	0,001799	0,044659	0,045486	0,940	0,000129	0,014162	0,014200	0,946
	$\tau/2$	0,001217	0,036054	0,036037	0,943	0,000351	0,011437	0,011547	0,946
	$3\tau/4$	0,001164	0,034218	0,034637	0,948	−0,000204	0,010895	0,010746	0,956
	τ	0,003532	0,041550	0,047835	0,915	0,000299	0,016351	0,019086	0,948
KMP	$\tau/4$	0,001832	0,044713	0,045359	0,942	0,000131	0,014153	0,014197	0,946
	$\tau/2$	0,001283	0,035999	0,035858	0,947	0,000368	0,011408	0,011509	0,947
	$3\tau/4$	0,001132	0,034004	0,034272	0,950	−0,000193	0,010838	0,010716	0,956
	τ	0,002628	0,039647	0,045615	0,906	0,000116	0,015851	0,017720	0,947
COX	$\tau/4$	0,000957	0,036029	0,035611	0,955	0,000107	0,011401	0,011229	0,955
	$\tau/2$	0,001067	0,031741	0,031499	0,954	0,000129	0,010031	0,009949	0,953
	$3\tau/4$	0,000972	0,028300	0,028071	0,962	0,000060	0,008937	0,008899	0,949
	τ	0,002177	0,026818	0,027274	0,955	0,000168	0,008790	0,008771	0,956
RCM	$\tau/4$	0,003717	0,035027	0,035896	0,940	0,002630	0,011076	0,011300	0,939
	$\tau/2$	0,002926	0,030819	0,031734	0,945	0,001853	0,009736	0,009995	0,936
	$3\tau/4$	0,002124	0,027440	0,028260	0,949	0,001247	0,008666	0,008949	0,940
	τ	0,001883	0,025457	0,025679	0,958	0,000621	0,008014	0,008240	0,946

Tableau 3.6: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base décroissant ($\gamma = 3/4$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,5. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}269$, $A(\tau/2) = 0{,}231$, $A(3\tau/4) = 0{,}200$ et $A(\tau) = 0{,}176$.

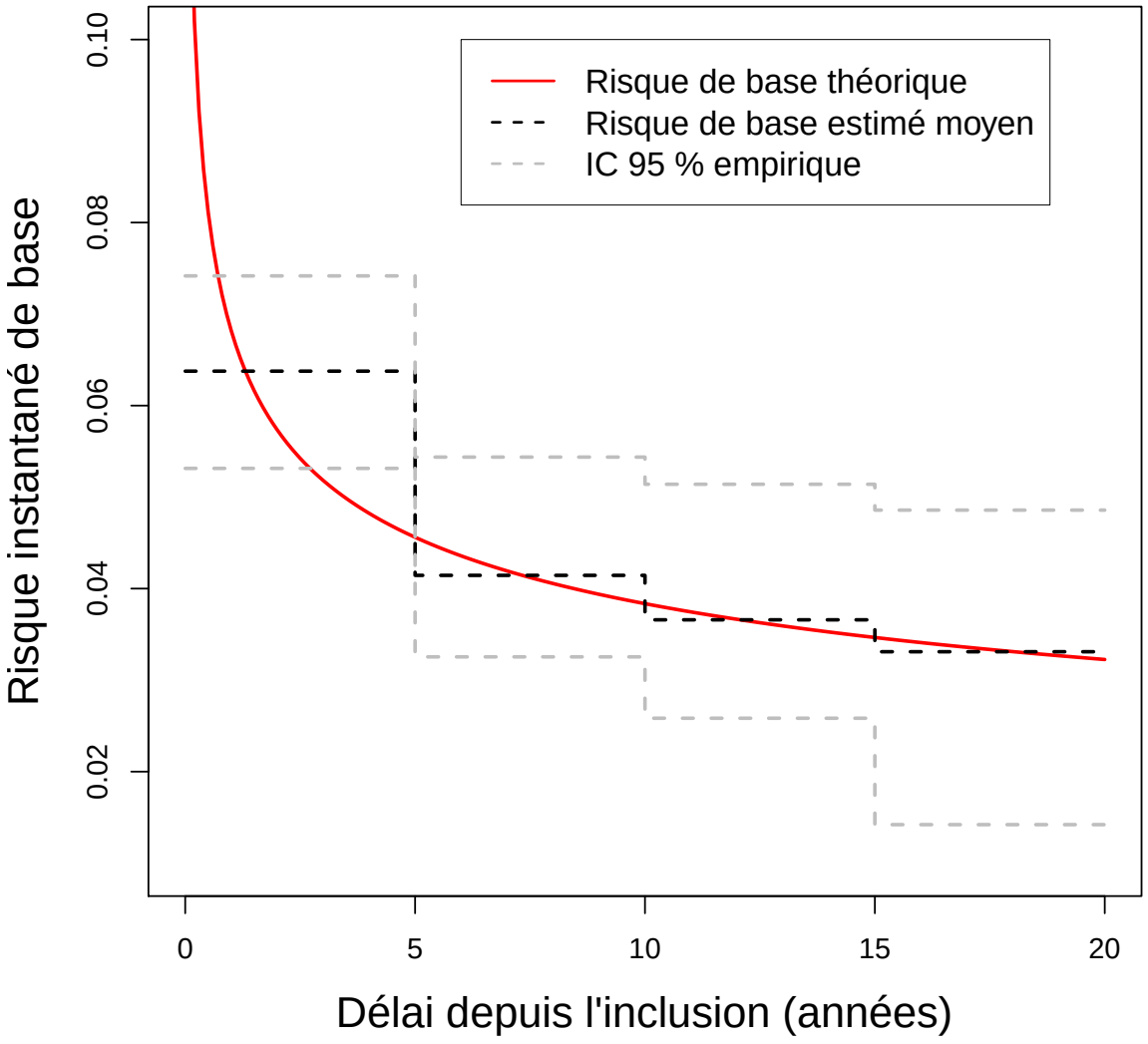


FIGURE 3.5: Estimation moyenne sur les 1 000 simulations du risque instantané de base par la méthode paramétrique RCM pour une durée de suivi divisée en quatre intervalles réguliers pour $\gamma = 3/4$, $n = 1\,000$ et $p = 0{,}50$

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	0,000814	0,064311	0,064377	0,947	−0,000024	0,020388	0,020204	0,956
	$\tau/2$	0,002020	0,043388	0,043169	0,952	0,000210	0,013761	0,013651	0,944
	$3\tau/4$	0,001174	0,037152	0,037027	0,955	−0,000469	0,011824	0,011798	0,960
	τ	0,007382	0,043968	0,054032	0,891	0,000554	0,018140	0,021081	0,939
KMP	$\tau/4$	0,000805	0,064427	0,064296	0,950	−0,000010	0,020380	0,020196	0,954
	$\tau/2$	0,002055	0,043322	0,042973	0,949	0,000272	0,013722	0,013643	0,947
	$3\tau/4$	0,001193	0,036838	0,036463	0,962	−0,000410	0,011739	0,011741	0,958
	τ	0,005596	0,040652	0,048586	0,898	0,000055	0,017280	0,019095	0,935
COX	$\tau/4$	0,001207	0,041863	0,040891	0,960	−0,000209	0,013250	0,013076	0,962
	$\tau/2$	0,001321	0,036377	0,035580	0,954	−0,000062	0,011499	0,011341	0,958
	$3\tau/4$	0,001300	0,030350	0,029672	0,956	−0,000121	0,009572	0,009502	0,965
	τ	0,002791	0,027165	0,028199	0,945	−0,000309	0,009206	0,010402	0,945
RCM	$\tau/4$	−0,000084	0,041594	0,040674	0,961	−0,001831	0,013151	0,013022	0,957
	$\tau/2$	0,000876	0,036176	0,035464	0,956	−0,000759	0,011424	0,011313	0,958
	$3\tau/4$	0,001462	0,030163	0,029655	0,959	−0,000051	0,009509	0,009485	0,961
	τ	0,002572	0,025716	0,024704	0,961	0,000622	0,008058	0,007962	0,945

Tableau 3.7: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base croissant ($\gamma = 4/3$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,5. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}299$, $A(\tau/2) = 0{,}250$, $A(3\tau/4) = 0{,}200$ et $A(\tau) = 0{,}153$.

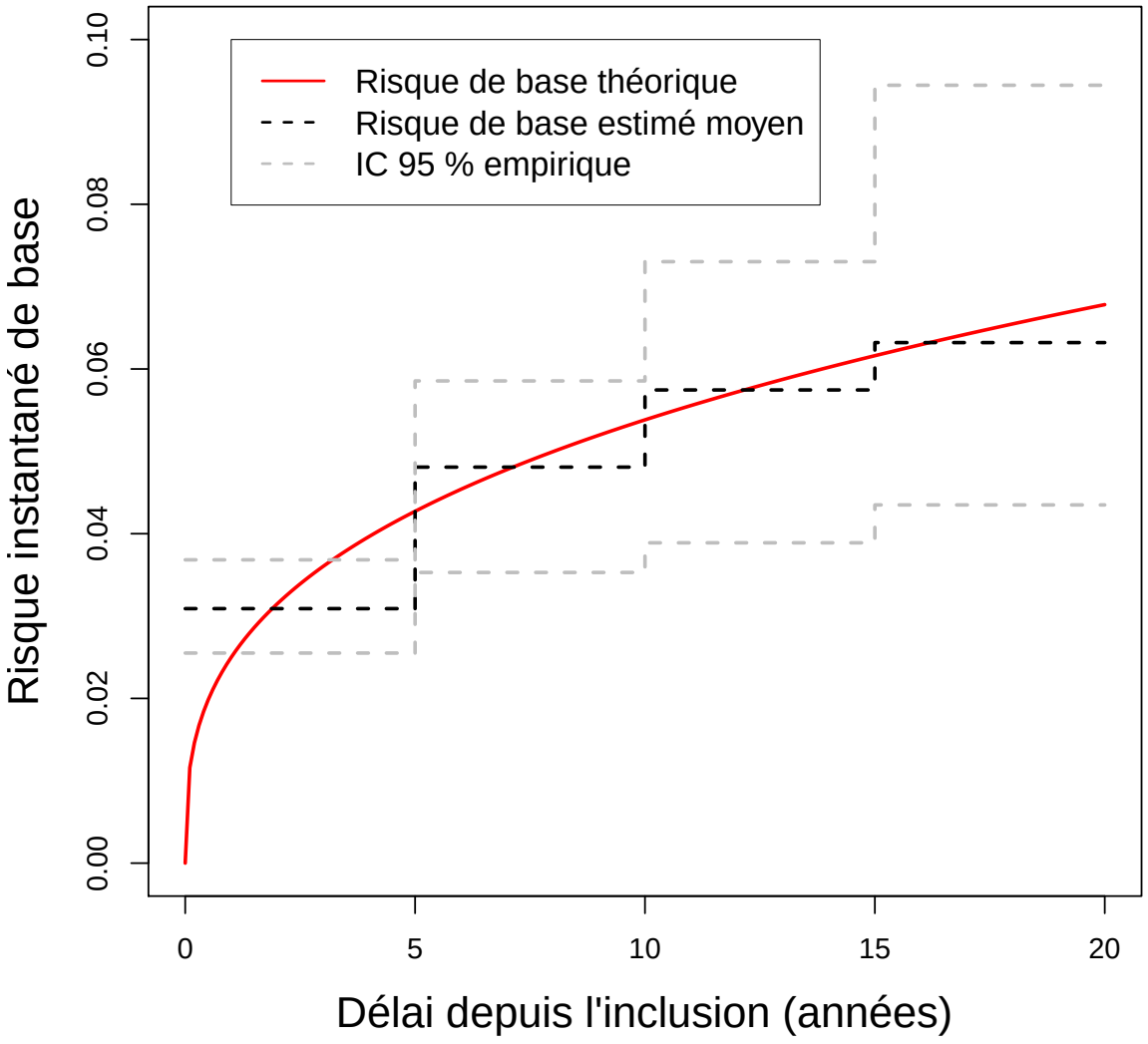


FIGURE 3.6: Estimation moyenne sur les 1 000 simulations du risque instantané de base par la méthode paramétrique RCM pour une durée de suivi divisée en quatre intervalles réguliers pour $\gamma = 4/3$, $n = 1\,000$ et $p = 0{,}50$

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	0,000767	0,037566	0,037107	0,957	−0,000344	0,011889	0,011883	0,954
	$\tau/2$	0,000318	0,026562	0,026090	0,954	−0,000062	0,008412	0,008505	0,948
	$3\tau/4$	0,000345	0,022928	0,022308	0,958	−0,000162	0,007283	0,007317	0,951
	τ	0,000608	0,027222	0,033360	0,921	0,000311	0,010786	0,012074	0,951
KMP	$\tau/4$	0,000772	0,037668	0,036910	0,959	−0,000311	0,011871	0,011871	0,954
	$\tau/2$	0,000310	0,026405	0,025910	0,961	−0,000018	0,008350	0,008425	0,947
	$3\tau/4$	0,000490	0,022436	0,022317	0,954	−0,000149	0,007161	0,007210	0,955
	τ	0,000011	0,023394	0,027320	0,908	0,000035	0,009801	0,010984	0,926
COX	$\tau/4$	0,000446	0,028454	0,027789	0,962	−0,000238	0,009000	0,009034	0,950
	$\tau/2$	0,000327	0,022835	0,022332	0,957	−0,000132	0,007217	0,007226	0,952
	$3\tau/4$	0,000112	0,018279	0,017625	0,962	−0,000202	0,005767	0,005767	0,953
	τ	0,000947	0,015938	0,016030	0,947	−0,000031	0,005241	0,005310	0,950
RCM	$\tau/4$	0,000672	0,027455	0,027806	0,956	−0,000210	0,008677	0,009038	0,938
	$\tau/2$	0,000474	0,021919	0,022349	0,950	−0,000116	0,006924	0,007227	0,940
	$3\tau/4$	0,000145	0,017434	0,017630	0,953	−0,000208	0,005502	0,005761	0,944
	τ	0,000344	0,014600	0,014513	0,952	−0,000189	0,004577	0,004667	0,949

Tableau 3.8: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base constant ($\gamma = 1$), un paramétre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,25. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}166$, $A(\tau/2) = 0{,}136$, $A(3\tau/4) = 0{,}111$ et $A(\tau) = 0{,}090$.

Pour une plus forte probabilité d'exposition à 0,75 (Tableau 3.9), les biais moyens restent faibles avec de bons taux de recouvrement sauf en τ pour les approches non paramétriques quelle que soit la taille d'échantillon considérée (0,880 et 0,879 avec $n =$ 1 000 observations et 0,923 et 0,919 avec $n = 10\,000$ observations pour KM et KMP respectivement). Les écarts-types estimés moyens sont plus grands que ceux obtenus avec des probabilités d'exposition inférieures 0,50 ou 0,25.

Lorsque le risque de base est décroissant avec une probabilité d'exposition à 0,25, nous retrouvons les mêmes résultats qu'avec un risque de base constant, une plus faible probabilité d'exposition et une taille d'échantillon de 1 000 observations (Tableau 3.10, partie gauche). Avec une taille d'échantillon plus importante $n = 10\,000$ (Tableau 3.10, partie droite), les biais restent faibles avec de bons taux de recouvrement sauf pour l'approche paramétrique RCM en des temps inférieurs à τ (0,931, 0,930 et 0,925 en $\tau/4$, $\tau/2$ et $3\tau/4$ respectivement). Les résultats sont moins meilleurs pour une probabilité d'exposition égale à 0,25 pour l'approche paramétrique en comparaison aux résultats obtenus pour $p = 0{,}50$ et similaires pour les taux de recouvrement en τ pour les approches non paramétriques.

Dans le cas d'une probabilité d'exposition plus importante ($p = 0{,}75$, Tableau 3.11), les biais moyens sont faibles avec de bons taux de recouvrement sauf pour les approches non paramétriques en des temps supérieurs à $\tau/4$ (0,93, 0,928 et 0,895 en $\tau/2$, $3\tau/4$ et τ pour KM et 0,932, 0,928 et 0,896 en $\tau/2$, $3\tau/4$ et τ pour KMP, Tableau 3.11, partie gauche). Avec $n = 10\,000$ observations, les résultats sont satisfaisants avec de bons taux de recouvrement sauf en $3\tau/4$ pour l'approche paramétrique RCM (Tableau 3.11, partie droite). Les résultats obtenus dans ce cas sont plus mauvais pour les approches non paramétriques comparés à ceux obtenus pour une probabilité d'exposition inférieure avec un risque de base décroissant.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	CP	Biais	SEE	SSD	CP
KM	$\tau/4$	−0,002134	0,075473	0,075167	0,956	0,000578	0,023935	0,024026	0,953
	$\tau/2$	−0,001817	0,058878	0,058752	0,944	0,000143	0,018721	0,018706	0,960
	$3\tau/4$	−0,001380	0,055677	0,058463	0,933	0,000299	0,017818	0,017566	0,950
	τ	0,002988	0,064332	0,079263	0,880	−0,000864	0,026996	0,033048	0,923
WKM	$\tau/4$	−0,002145	0,075537	0,075176	0,956	0,000561	0,023933	0,024012	0,953
	$\tau/2$	−0,001797	0,058865	0,058784	0,943	0,000131	0,018709	0,018715	0,959
	$3\tau/4$	−0,001410	0,055508	0,058293	0,933	0,000328	0,017795	0,017551	0,948
	τ	0,002194	0,063226	0,077751	0,879	−0,000705	0,026799	0,032880	0,919
COX	$\tau/4$	−0,002641	0,054982	0,055028	0,948	0,000035	0,017371	0,017306	0,955
	$\tau/2$	−0,001861	0,050909	0,050975	0,947	0,000099	0,016102	0,016095	0,952
	$3\tau/4$	−0,001044	0,046584	0,046373	0,952	0,000180	0,014736	0,014743	0,949
	τ	0,000817	0,043942	0,045675	0,944	−0,000077	0,014623	0,015048	0,951
RCM	$\tau/4$	−0,002187	0,054875	0,055066	0,949	0,000104	0,017318	0,017314	0,952
	$\tau/2$	−0,001530	0,050815	0,050969	0,946	0,000130	0,016052	0,016088	0,951
	$3\tau/4$	−0,000930	0,046465	0,046375	0,951	0,000192	0,014684	0,014709	0,948
	τ	−0,000619	0,042921	0,043412	0,944	0,000102	0,013576	0,013704	0,948

Tableau 3.9: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base constant ($\gamma = 1$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,75. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}373$, $A(\tau/2) = 0{,}321$, $A(3\tau/4) = 0{,}273$ et $A(\tau) = 0{,}229$.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	0,000576	0,031436	0,031512	0,947	−0,000156	0,009951	0,009821	0,949
	$\tau/2$	−0,000018	0,024138	0,024207	0,948	0,000132	0,007650	0,007527	0,948
	$3\tau/4$	0,000282	0,021997	0,021864	0,954	−0,000028	0,006992	0,006834	0,962
	τ	0,000929	0,026094	0,030775	0,926	−0,000096	0,010183	0,011707	0,959
KMP	$\tau/4$	0,000641	0,031509	0,031370	0,948	−0,000145	0,009934	0,009798	0,951
	$\tau/2$	0,000087	0,024014	0,024173	0,950	0,000141	0,007598	0,007467	0,949
	$3\tau/4$	0,000434	0,021602	0,021875	0,947	−0,000042	0,006892	0,006768	0,956
	τ	0,000353	0,023234	0,026842	0,909	−0,000277	0,009454	0,010385	0,942
COX	$\tau/4$	0,000282	0,025741	0,025522	0,957	−0,000003	0,008145	0,007991	0,953
	$\tau/2$	0,000167	0,021086	0,020887	0,955	0,000009	0,006665	0,006539	0,953
	$3\tau/4$	0,000031	0,017711	0,017354	0,961	−0,000053	0,005591	0,005492	0,948
	τ	0,000791	0,016048	0,016095	0,952	0,000038	0,005229	0,005080	0,956
RCM	$\tau/4$	0,002277	0,024392	0,025857	0,936	0,001856	0,007712	0,008081	0,931
	$\tau/2$	0,001393	0,019816	0,021115	0,938	0,001172	0,006262	0,006590	0,930
	$3\tau/4$	0,000714	0,016529	0,017509	0,940	0,000656	0,005220	0,005519	0,925
	τ	0,000516	0,014581	0,015170	0,948	0,000246	0,004581	0,004757	0,937

Tableau 3.10: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base décroissant ($\gamma = 3/4$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,25. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}156$, $A(\tau/2) = 0{,}130$, $A(3\tau/4) = 0{,}111$ et $A(\tau) = 0{,}096$.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	−0,000871	0,065201	0,065421	0,947	0,000248	0,020686	0,021064	0,945
	$\tau/2$	−0,000767	0,054678	0,056619	0,930	0,000137	0,017381	0,017500	0,948
	$3\tau/4$	−0,000572	0,053503	0,056763	0,928	0,000277	0,017098	0,016872	0,952
	τ	0,001196	0,062055	0,075090	0,895	0,000022	0,025266	0,028301	0,939
KMP	$\tau/4$	−0,000883	0,065244	0,065428	0,947	0,000240	0,020684	0,021052	0,946
	$\tau/2$	−0,000764	0,054661	0,056654	0,932	0,000133	0,017371	0,017503	0,948
	$3\tau/4$	−0,000613	0,053364	0,056605	0,928	0,000302	0,017079	0,016849	0,954
	τ	0,001006	0,061298	0,074694	0,896	0,000161	0,025076	0,028034	0,937
COX	$\tau/4$	−0,001577	0,051995	0,052297	0,943	−0,000026	0,016433	0,016484	0,948
	$\tau/2$	−0,000866	0,048360	0,048908	0,941	0,000030	0,015291	0,015337	0,951
	$3\tau/4$	−0,000287	0,045083	0,045280	0,943	0,000118	0,014250	0,014345	0,944
	τ	0,001123	0,043349	0,044765	0,947	0,000066	0,014142	0,014571	0,948
RCM	$\tau/4$	0,001271	0,051193	0,052540	0,943	0,002648	0,016163	0,016546	0,937
	$\tau/2$	0,001207	0,047658	0,049158	0,935	0,001984	0,015054	0,015398	0,939
	$3\tau/4$	0,001149	0,044445	0,045465	0,940	0,001617	0,014039	0,014405	0,930
	τ	0,000797	0,042143	0,043353	0,936	0,000989	0,013315	0,013726	0,944

Tableau 3.11: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base décroissant ($\gamma = 3/4$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,75. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}356$, $A(\tau/2) = 0{,}310$, $A(3\tau/4) = 0{,}273$ et $A(\tau) = 0{,}241$.

Lorsque le risque de base est croissant ($\gamma = 4/3$) et la probabilité d'exposition plus faible ($p = 0{,}25$, nous retrouvons également de bons taux de recouvrement avec une taille d'échantillon de 1 000 observations sauf en τ pour les approches non paramétriques où ces taux sont trop faibles (0,908 et 0,910 pour KM et KMP respectivement, Tableau 3.12, partie gauche) mais meilleurs que ceux obtenus avec un risque de base croissant et une probabilité d'exposition à 0,50. Les écarts-type estimés moyens restent inférieurs à ceux obtenus avec une probabilité d'exposition à 0,50. Pour $n = 10\,000$, les résultats sont satisfaisants pour toutes les méthodes (Tableau 3.12, partie droite).

Pour une probabilité d'exposition ($p = 0{,}75$, Tableau 3.13), les biais restent faibles pour toutes les méthodes d'estimation avec de bons taux de recouvrement sauf pour les méthodes non paramétriques en τ pour toutes les tailles d'échantillon considérées.

En résumé, nos résultats montrent que, pour une plus faible probabilité d'exposition ($p = 0{,}25$ comparé à 0,50), les taux de recouvrement s'améliorent en τ pour les approches non paramétriques KM et KMP par rapport à une probabilité d'exposition de 0,50. Dans le cas d'une probabilité d'exposition plus importante ($p = 0{,}75$) avec un risque de base décroissant, les taux de recouvrement obtenus pour les approches non paramétriques en des temps supérieurs à $\tau/4$ sont inférieurs à la valeur attendue. Les résultats obtenus pour l'approche semi-paramétrique sont satisfaisants pour les trois probabilités d'exposition considérées ($p = 0{,}25$, 0,50, 0,75). On obtient la même conclusion pour l'approche paramétrique sauf pour $p = 0{,}25$ avec un risque de base décroissant où les taux de recouvrement sont inférieurs à la valeur attendue en des temps inférieurs à τ.

3.3.4 Hypothèse nulle

Lorsque nous considérons un risque de base constant ($\gamma = 1$) avec une probabilité d'exposition à 0,50 et $\beta = 0$ (Tableau 3.14), les résultats sont similaires à ceux obtenus avec $\beta = \ln(2)$ sauf pour les écarts-types estimés moyens qui sont plus importants.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	0,000661	0,046833	0,045817	0,957	−0,000317	0,014849	0,014767	0,960
	$\tau/2$	0,000773	0,029813	0,029627	0,946	−0,000092	0,009441	0,009468	0,953
	$3\tau/4$	0,000440	0,023881	0,023482	0,955	−0,000257	0,007585	0,007642	0,951
	τ	0,003221	0,027486	0,034493	0,908	0,000153	0,010948	0,012184	0,962
KMP	$\tau/4$	0,000643	0,046998	0,045644	0,959	−0,000285	0,014833	0,014746	0,958
	$\tau/2$	0,000709	0,029640	0,029408	0,948	−0,000041	0,009367	0,009383	0,951
	$3\tau/4$	0,000558	0,023299	0,023301	0,954	−0,000235	0,007437	0,007501	0,955
	τ	0,001232	0,023006	0,026720	0,910	0,000099	0,009749	0,010679	0,942
COX	$\tau/4$	0,000846	0,031443	0,030831	0,959	−0,000290	0,009952	0,009901	0,952
	$\tau/2$	0,000662	0,025014	0,024554	0,963	−0,000176	0,007907	0,007862	0,952
	$3\tau/4$	0,000337	0,018922	0,018316	0,959	−0,000229	0,005967	0,005913	0,953
	τ	0,001204	0,015569	0,016130	0,937	−0,000222	0,005239	0,005524	0,952
RCM	$\tau/4$	−0,000170	0,030598	0,030618	0,952	−0,001482	0,009673	0,009838	0,949
	$\tau/2$	0,000306	0,024318	0,024478	0,954	−0,000674	0,007679	0,007841	0,946
	$3\tau/4$	0,000414	0,018327	0,018333	0,947	−0,000192	0,005780	0,005906	0,945
	τ	0,001036	0,014274	0,013874	0,953	0,000320	0,004462	0,004419	0,961

Tableau 3.12: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base croissant ($\gamma = 4/3$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,25. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}176$, $A(\tau/2) = 0{,}143$, $A(3\tau/4) = 0{,}111$ et $A(\tau) = 0{,}083$.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	−0,001309	0,091279	0,089801	0,946	0,000485	0,028931	0,029514	0,951
	$\tau/2$	−0,000412	0,064396	0,064549	0,960	−0,000176	0,020489	0,020258	0,956
	$3\tau/4$	−0,001473	0,057973	0,061039	0,936	0,000221	0,018559	0,018382	0,939
	τ	0,006152	0,065252	0,080308	0,870	0,000409	0,028079	0,034680	0,917
KMP	$\tau/4$	−0,001317	0,091372	0,089783	0,946	0,000468	0,028929	0,029504	0,951
	$\tau/2$	−0,000374	0,064390	0,064602	0,960	−0,000189	0,020475	0,020265	0,956
	$3\tau/4$	−0,001491	0,057773	0,060894	0,937	0,000255	0,018530	0,018366	0,944
	τ	0,004745	0,063764	0,078790	0,864	0,000267	0,027712	0,034499	0,915
COX	$\tau/4$	−0,003044	0,058159	0,058034	0,949	0,000011	0,018360	0,018412	0,943
	$\tau/2$	−0,002172	0,053883	0,053788	0,946	0,000090	0,017036	0,017101	0,946
	$3\tau/4$	−0,000959	0,048296	0,048116	0,949	0,000212	0,015272	0,015391	0,941
	τ	0,001886	0,044411	0,045302	0,939	0,000449	0,014880	0,015904	0,944
RCM	$\tau/4$	−0,004390	0,058419	0,057896	0,948	−0,001709	0,018416	0,018355	0,951
	$\tau/2$	−0,002620	0,054150	0,053711	0,948	−0,000627	0,017094	0,017064	0,950
	$3\tau/4$	−0,000804	0,048516	0,048154	0,954	0,000330	0,015324	0,015369	0,944
	τ	0,000336	0,043175	0,043502	0,946	0,001345	0,013664	0,013993	0,940

Tableau 3.13: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base croissant ($\gamma = 4/3$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,75. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}390$, $A(\tau/2) = 0{,}334$, $A(3\tau/4) = 0{,}273$ et $A(\tau) = 0{,}213$.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	0,001588	0,066720	0,066661	0,951	0,000298	0,021095	0,020829	0,959
	$\tau/2$	0,001831	0,049184	0,049848	0,945	0,000730	0,015565	0,015568	0,940
	$3\tau/4$	0,000775	0,044459	0,044906	0,948	−0,000044	0,014117	0,014006	0,950
	τ	0,003632	0,054707	0,063368	0,921	0,001753	0,021781	0,024252	0,955
WKM	$\tau/4$	0,001584	0,066767	0,066675	0,951	0,000300	0,021097	0,020840	0,959
	$\tau/2$	0,001849	0,049203	0,049846	0,946	0,000731	0,015568	0,015582	0,940
	$3\tau/4$	0,000670	0,044428	0,044884	0,948	−0,000039	0,014115	0,014006	0,951
	τ	0,002958	0,052653	0,060895	0,905	0,001605	0,021354	0,024770	0,948
COX	$\tau/4$	0,001693	0,047709	0,047293	0,961	0,000334	0,015079	0,014993	0,958
	$\tau/2$	0,001521	0,042265	0,041903	0,963	0,000294	0,013349	0,013276	0,958
	$3\tau/4$	0,001314	0,037280	0,036941	0,963	0,000261	0,011758	0,011698	0,958
	τ	0,001276	0,032920	0,032690	0,962	0,000222	0,010319	0,010275	0,958
RCM	$\tau/4$	0,001755	0,047745	0,047311	0,961	0,000335	0,015080	0,014992	0,959
	$\tau/2$	0,001581	0,042294	0,041918	0,962	0,000295	0,013349	0,013274	0,959
	$3\tau/4$	0,001378	0,037290	0,036950	0,962	0,000262	0,011757	0,011694	0,959
	τ	0,001248	0,032792	0,032573	0,963	0,000221	0,010312	0,010254	0,959

Tableau 3.14: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base constant ($\gamma = 1$), un paramètre de régression $\beta = 0$ et une probabilité d'exposition égale à 0,5. Les valeurs théoriques du risque attribuable aux temps $\tau/4$, $\tau/2$, $3\tau/4$ et τ sont toutes nulles.

Avec une probabilité d'exposition plus faible ($p = 0{,}25$, Tableau 3.15) les résultats sont similaires à ceux obtenus avec $\beta = \ln(2)$ avec en particulier un bon taux de recouvrement en τ pour l'approche non paramétrique KMP.

Pour une probabilité d'exposition plus importante ($p = 0{,}75$, Tableau 3.16), les écarts-types estimés moyens sont plus importants que ceux obtenus pour $\beta = \ln(2)$. Les taux de recouvrement en τ sont meilleurs pour les deux approches non paramétriques KM et KMP quand la taille d'échantillon est égale à 10 000 observations.

Avec une risque de base décroissant ($\gamma = 3/4$) et une probabilité d'exposition à 0,50 (Tableau 3.17), les biais restent faibles. Le taux de recouvrement s'améliore pour l'approche non paramétrique KM en τ avec $n = 1\,000$ comparé au taux de recouvrement obtenu pour $\beta = \ln(2)$ avec les mêmes paramètres. Pour l'approche non paramétrique KMP, le taux de recouvrement en τ s'améliore également mais reste inférieur à la valeur nominale attendue. Avec $n = 10\,000$, les résultats sont satisfaisants pour toutes les méthodes sauf peut-être pour les deux approches non paramétriques KM et KMP en $\tau/2$ avec un taux de recouvrement égal à 0,935.

Avec une probabilité d'exposition plus faible ($p = 0{,}25$, Tableau 3.18), les résultats obtenus sont meilleurs pour l'approche paramétrique RCM par rapport à ceux obtenus avec $\beta = \ln(2)$ et une taille d'échantillon de 10 000 observations. Les taux de recouvrement s'améliorent pour les approches non paramétriques KM et KMP, restent légèrement inférieurs à 93 % pour $n = 1\,000$ et n'atteignent pas le taux de recouvrement attendu en τ pour $n = 10\,000$ avec l'approche KMP.

Lorsque la probabilité d'exposition est plus importante ($p = 0{,}75$, Tableau 3.19), les résultats s'améliorent nettement pour les approches non paramétriques, surtout en $\tau/2$ et $3\tau/4$, mais la probabilité de recouvrement n'atteint pas la valeur nominale en τ pour $n = 1\,000$ et se détériore pour $n = 10\,000$.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	0,001425	0,038511	0,037543	0,958	−0,000626	0,012177	0,012270	0,954
	$\tau/2$	0,001205	0,028406	0,028036	0,956	−0,000006	0,008991	0,009019	0,955
	$3\tau/4$	0,000660	0,025674	0,025729	0,953	−0,000393	0,008154	0,008028	0,950
	τ	0,000445	0,032276	0,038851	0,931	0,000701	0,012779	0,015050	0,952
KMP	$\tau/4$	0,001422	0,038518	0,037502	0,958	−0,000629	0,012180	0,012273	0,954
	$\tau/2$	0,001212	0,028346	0,027943	0,954	−0,000006	0,008993	0,009026	0,955
	$3\tau/4$	0,000697	0,025518	0,025619	0,950	−0,000392	0,008148	0,008022	0,949
	τ	0,000346	0,028872	0,034740	0,887	0,000417	0,011999	0,013944	0,938
COX	$\tau/4$	0,001402	0,027538	0,026947	0,962	−0,000232	0,008708	0,008698	0,951
	$\tau/2$	0,001068	0,024368	0,023838	0,961	−0,000223	0,007710	0,007703	0,953
	$3\tau/4$	0,000771	0,021471	0,020982	0,960	−0,000212	0,006792	0,006785	0,956
	τ	0,000560	0,018954	0,018553	0,958	−0,000203	0,005961	0,005956	0,956
RCM	$\tau/4$	0,001409	0,027544	0,026951	0,960	−0,000232	0,008708	0,008699	0,952
	$\tau/2$	0,001076	0,024372	0,023842	0,960	−0,000223	0,007709	0,007703	0,954
	$3\tau/4$	0,000783	0,021466	0,020975	0,958	−0,000212	0,006791	0,006784	0,954
	τ	0,000541	0,018861	0,018485	0,959	−0,000204	0,005956	0,005952	0,953

Tableau 3.15: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base constant ($\gamma = 1$), un paramètre de régression $\beta = 0$ et une probabilité d'exposition égale à 0,25. Les valeurs théoriques du risque attribuable aux temps $\tau/4$, $\tau/2$, $3\tau/4$ et τ sont toutes nulles.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	−0,004259	0,115661	0,113081	0,957	0,001867	0,036511	0,036788	0,954
	$\tau/2$	−0,003684	0,085124	0,084204	0,954	0,000005	0,026959	0,027088	0,955
	$3\tau/4$	−0,002343	0,076706	0,077100	0,950	0,001178	0,024421	0,024068	0,949
	τ	−0,001418	0,087108	0,105760	0,892	−0,001280	0,036039	0,042169	0,939
KMP	$\tau/4$	−0,004266	0,115772	0,113108	0,957	0,001864	0,036511	0,036784	0,954
	$\tau/2$	−0,003685	0,085204	0,084263	0,954	0,000006	0,026958	0,027080	0,955
	$3\tau/4$	−0,002307	0,076708	0,077079	0,950	0,001178	0,024423	0,024072	0,949
	τ	−0,001516	0,086734	0,104647	0,886	−0,001286	0,035975	0,041855	0,937
COX	$\tau/4$	−0,004361	0,082776	0,081227	0,958	0,000685	0,026104	0,026087	0,951
	$\tau/2$	−0,003334	0,073247	0,071849	0,960	0,000659	0,023111	0,023102	0,953
	$3\tau/4$	−0,002423	0,064537	0,063232	0,959	0,000627	0,020359	0,020349	0,955
	τ	−0,001777	0,056972	0,055921	0,958	0,000602	0,017869	0,017863	0,956
RCM	$\tau/4$	−0,004385	0,082903	0,081234	0,960	0,000684	0,026108	0,026088	0,952
	$\tau/2$	−0,003362	0,073354	0,071858	0,960	0,000659	0,023112	0,023102	0,954
	$3\tau/4$	−0,002463	0,064608	0,063212	0,958	0,000627	0,020359	0,020346	0,954
	τ	−0,001730	0,056772	0,055712	0,959	0,000604	0,017858	0,017851	0,953

Tableau 3.16: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base constant ($\gamma = 1$), un paramètre de régression $\beta = 0$ et une probabilité d'exposition égale à 0,75. Les valeurs théoriques du risque attribuable aux temps $\tau/4$, $\tau/2$, $3\tau/4$ et τ sont toutes nulles.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	0,000910	0,056562	0,057461	0,945	0,000391	0,017883	0,017828	0,953
	$\tau/2$	0,001376	0,045152	0,045351	0,952	0,000441	0,014296	0,014733	0,935
	$3\tau/4$	0,000467	0,042654	0,042845	0,954	−0,000198	0,013549	0,013642	0,950
	τ	0,002500	0,052082	0,059759	0,940	0,000620	0,020352	0,022852	0,954
KMP	$\tau/4$	0,000917	0,056613	0,057468	0,944	0,000390	0,017885	0,017837	0,953
	$\tau/2$	0,001372	0,045184	0,045358	0,953	0,000438	0,014299	0,014738	0,935
	$3\tau/4$	0,000387	0,042642	0,042828	0,955	−0,000197	0,013548	0,013647	0,950
	τ	0,001736	0,050807	0,057491	0,930	0,000518	0,020118	0,022282	0,950
COX	$\tau/4$	0,000936	0,044443	0,043998	0,960	0,000221	0,014045	0,014137	0,948
	$\tau/2$	0,000839	0,039816	0,039430	0,960	0,000196	0,012575	0,012659	0,948
	$3\tau/4$	0,000744	0,036064	0,035704	0,960	0,000178	0,011378	0,011459	0,948
	τ	0,000698	0,032970	0,032703	0,960	0,000149	0,010358	0,010419	0,948
RCM	$\tau/4$	0,000945	0,044158	0,044439	0,952	0,000231	0,013947	0,014264	0,939
	$\tau/2$	0,000854	0,039508	0,039776	0,953	0,000205	0,012474	0,012758	0,939
	$3\tau/4$	0,000770	0,035739	0,035983	0,953	0,000186	0,011274	0,011536	0,939
	τ	0,000709	0,032545	0,032831	0,953	0,000163	0,010244	0,010480	0,939

Tableau 3.17: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base décroissant ($\gamma = 3/4$), un paramètre de régression $\beta = 0$ et une probabilité d'exposition égale à 0,5. Les valeurs théoriques du risque attribuable aux temps $\tau/4$, $\tau/2$, $3\tau/4$ et τ sont toutes nulles.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	0,001108	0,032643	0,032315	0,957	−0,000186	0,010328	0,010369	0,950
	$\tau/2$	0,000583	0,026065	0,026649	0,949	−0,000009	0,008258	0,008282	0,943
	$3\tau/4$	0,000324	0,024626	0,024929	0,949	−0,000306	0,007827	0,007652	0,951
	τ	0,001439	0,030560	0,036175	0,959	0,000195	0,011802	0,013115	0,938
KMP	$\tau/4$	0,001112	0,032657	0,032280	0,957	−0,000187	0,010329	0,010374	0,950
	$\tau/2$	0,000616	0,026025	0,026543	0,949	−0,000009	0,008258	0,008284	0,943
	$3\tau/4$	0,000343	0,024510	0,024899	0,951	−0,000302	0,007820	0,007654	0,952
	τ	0,000546	0,028037	0,032266	0,926	0,000092	0,011313	0,012669	0,934
COX	$\tau/4$	0,000837	0,025639	0,025377	0,956	−0,000099	0,008112	0,008105	0,947
	$\tau/2$	0,000600	0,022952	0,022706	0,955	−0,000103	0,007263	0,007258	0,947
	$3\tau/4$	0,000415	0,020777	0,020543	0,955	−0,000104	0,006572	0,006567	0,948
	τ	0,000270	0,018988	0,018776	0,956	−0,000111	0,005983	0,005979	0,950
RCM	$\tau/4$	0,000856	0,025458	0,025616	0,952	−0,000094	0,008055	0,008173	0,943
	$\tau/2$	0,000615	0,022760	0,022891	0,952	−0,000099	0,007204	0,007311	0,943
	$3\tau/4$	0,000432	0,020577	0,020681	0,952	−0,000101	0,006511	0,006608	0,945
	τ	0,000287	0,018730	0,018838	0,950	−0,000104	0,005916	0,006006	0,944

Tableau 3.18: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base décroissant ($\gamma = 3/4$), un paramètre de régression $\beta = 0$ et une probabilité d'exposition égale à 0,25. Les valeurs théoriques du risque attribuable aux temps $\tau/4$, $\tau/2$, $3\tau/4$ et τ sont toutes nulles.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Bias	SEE	SSD	PC
KM	$\tau/4$	−0,003399	0,098078	0,097150	0,956	0,000548	0,030962	0,031127	0,950
	$\tau/2$	−0,001819	0,078169	0,079771	0,948	0,000005	0,024756	0,024835	0,943
	$3\tau/4$	−0,001227	0,073688	0,075031	0,949	0,000903	0,023437	0,022978	0,952
	τ	−0,001173	0,084725	0,097406	0,930	−0,000184	0,033880	0,037768	0,934
KMP	$\tau/4$	−0,003398	0,098156	0,097179	0,955	0,000547	0,030964	0,031122	0,950
	$\tau/2$	−0,001794	0,078220	0,079838	0,948	0,000005	0,024756	0,024832	0,943
	$3\tau/4$	−0,001202	0,073674	0,074952	0,947	0,000907	0,023440	0,022975	0,952
	τ	−0,002239	0,084271	0,097886	0,924	−0,000297	0,033912	0,037922	0,934
COX	$\tau/4$	−0,002606	0,077064	0,076445	0,955	0,000289	0,024317	0,024312	0,946
	$\tau/2$	−0,001877	0,068988	0,068406	0,953	0,000303	0,021773	0,021773	0,947
	$3\tau/4$	−0,001314	0,062449	0,061884	0,955	0,000307	0,019700	0,019699	0,948
	τ	−0,000876	0,057078	0,056557	0,955	0,000328	0,017934	0,017937	0,948
RCM	$\tau/4$	−0,002661	0,076623	0,077157	0,952	0,000273	0,024149	0,024518	0,943
	$\tau/2$	−0,001920	0,068503	0,068954	0,952	0,000289	0,021597	0,021931	0,943
	$3\tau/4$	−0,001362	0,061931	0,062294	0,952	0,000295	0,019522	0,019822	0,945
	τ	−0,000927	0,056376	0,056750	0,950	0,000305	0,017738	0,018018	0,944

Tableau 3.19: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base décroissant ($\gamma = 3/4$), un paramètre de régression $\beta = 0$ et une probabilité d'exposition égale à 0,75. Les valeurs théoriques du risque attribuable aux temps $\tau/4$, $\tau/2$, $3\tau/4$ et τ sont toutes nulles.

Enfin, avec un risque de base croissant ($\gamma = 4/3$) et une probabilité d'exposition à 0,50 (Tableau 3.20), les résultats sont similaires à ceux obtenus pour un risque de base constant avec les mêmes paramètres.

Avec une probabilité d'exposition plus faible ($p = 0{,}25$, Tableau 3.21), les résultats s'améliorent pour l'approche non paramétrique KM en τ avec $n = 1\,000$ (Tableau 3.21, partie gauche). Les taux de recouvrement sont satisfaisants sauf pour l'approche non paramétrique KMP en τ avec $n = 1\,000$ comme $10\,000$.

Lorsque la prévalence de l'exposition est plus forte ($p = 0{,}75$, Tableau 3.22), les taux de recouvrement s'améliorent pour les approches non paramétriques en τ mais n'atteignent pas la valeur attendue. Les écarts-types estimés moyens sont plus importants que ceux obtenus pour $\beta = \ln(2)$ avec les mêmes paramètres.

En résumé, avec $\beta = 0$ et pour un risque de base constant ($\gamma = 1$) ou croissant ($\gamma = 4/3$), les résultats sont similaires à ceux obtenus avec $\beta = \ln(2)$ avec des écarts-types estimés moyens plus importants et des taux de recouvrement qui s'améliorent souvent pour les approches non paramétriques KM et KMP en τ. Pour un risque de base décroissant ($\gamma = 3/4$), les taux de recouvrement s'améliorent également en τ pour les approches non paramétriques KM et KMP. Avec une probabilité d'exposition plus faible ($p = 0{,}25$) les résultats sont meilleurs pour l'approche paramétrique RCM par rapport à ceux obtenus avec $\beta = \ln(2)$ avec des taux de recouvrement compris entre 94,3 % et 94,5 % pour une taille d'échantillon de $10\,000$ observations.

3.3.5 Risques non proportionnels

Lorsque les données sont générées selon un modèle à risques non proportionnels avec une probabilité d'exposition à 0,50 et $\beta = \ln(2)$ (Tableau 3.23), les approches non paramétriques KM et KMP donnent des résultats similaires au cas des risques proportionnels. En revanche, les approches semi-paramétrique COX et paramétrique RCM reposant sur l'hypothèse des risques proportionnels donnent de mauvais résultats. Avec une taille

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	0,001511	0,082259	0,083244	0,945	0,000093	0,026029	0,026520	0,947
	$\tau/2$	0,001846	0,054592	0,054752	0,959	0,000794	0,017267	0,016809	0,957
	$3\tau/4$	0,001022	0,046325	0,046390	0,947	−0,000120	0,014706	0,014701	0,952
	τ	0,004487	0,055493	0,065578	0,920	0,000098	0,022799	0,025225	0,963
KMP	$\tau/4$	0,001515	0,082325	0,083138	0,945	0,000093	0,026032	0,026529	0,948
	$\tau/2$	0,001883	0,054616	0,054730	0,958	0,000800	0,017272	0,016821	0,957
	$3\tau/4$	0,000884	0,046281	0,046359	0,946	−0,000114	0,014703	0,014700	0,952
	τ	0,003575	0,053048	0,063653	0,897	−0,000083	0,022246	0,025292	0,954
COX	$\tau/4$	0,002155	0,051231	0,050304	0,956	0,000190	0,016195	0,016059	0,950
	$\tau/2$	0,001903	0,045166	0,044351	0,956	0,000167	0,014261	0,014143	0,950
	$3\tau/4$	0,001584	0,038619	0,037885	0,956	0,000142	0,012175	0,012080	0,951
	τ	0,001290	0,032390	0,031695	0,958	0,000119	0,010121	0,010038	0,951
RCM	$\tau/4$	0,002142	0,051417	0,050097	0,959	0,000181	0,016239	0,015982	0,952
	$\tau/2$	0,001890	0,045401	0,044243	0,959	0,000159	0,014323	0,014099	0,952
	$3\tau/4$	0,001588	0,038872	0,037850	0,959	0,000135	0,012251	0,012063	0,952
	τ	0,001362	0,032528	0,031683	0,960	0,000102	0,010217	0,010057	0,952

Tableau 3.20: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base croissant ($\gamma = 4/3$), un paramètre de régression $\beta = 0$ et une probabilité d'exposition égale à 0,5. Les valeurs théoriques du risque attribuable aux temps $\tau/4$, $\tau/2$, $3\tau/4$ et τ sont toutes nulles.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	0,000035	0,047408	0,046678	0,948	−0,000361	0,015028	0,015436	0,948
	$\tau/2$	0,000837	0,031520	0,031422	0,970	0,000203	0,009977	0,009816	0,956
	$3\tau/4$	0,000680	0,026750	0,026867	0,950	−0,000381	0,008495	0,008394	0,947
	τ	0,001780	0,032748	0,038541	0,943	−0,000129	0,013092	0,014941	0,946
KMP	$\tau/4$	0,000037	0,047424	0,046671	0,947	−0,000362	0,015032	0,015446	0,948
	$\tau/2$	0,000814	0,031455	0,031347	0,966	0,000205	0,009979	0,009818	0,958
	$3\tau/4$	0,000662	0,026576	0,026739	0,948	−0,000382	0,008487	0,008385	0,946
	τ	0,000347	0,029102	0,033903	0,896	−0,000362	0,012323	0,014312	0,928
COX	$\tau/4$	0,001675	0,029578	0,028929	0,959	−0,000285	0,009353	0,009366	0,951
	$\tau/2$	0,001277	0,026042	0,025481	0,961	−0,000272	0,008237	0,008248	0,951
	$3\tau/4$	0,000853	0,022233	0,021728	0,962	−0,000254	0,007033	0,007042	0,950
	τ	0,000524	0,018627	0,018227	0,965	−0,000240	0,005848	0,005849	0,950
RCM	$\tau/4$	0,001657	0,029678	0,028815	0,962	−0,000292	0,009377	0,009327	0,950
	$\tau/2$	0,001266	0,026172	0,025420	0,963	−0,000278	0,008271	0,008227	0,951
	$3\tau/4$	0,000853	0,022376	0,021699	0,962	−0,000259	0,007076	0,007037	0,951
	τ	0,000503	0,018703	0,018186	0,961	−0,000242	0,005902	0,005870	0,951

Tableau 3.21: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base croissant ($\gamma = 4/3$), un paramètre de régression $\beta = 0$ et une probabilité d'exposition égale à 0,25. Les valeurs théoriques du risque attribuable aux temps $\tau/4$, $\tau/2$, $3\tau/4$ et τ sont toutes nulles.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	−0,000141	0,142412	0,140261	0,947	0,001059	0,045058	0,046262	0,948
	$\tau/2$	−0,002195	0,094438	0,094099	0,966	−0,000617	0,029915	0,029444	0,958
	$3\tau/4$	−0,002097	0,079879	0,080327	0,947	0,001146	0,025438	0,025158	0,946
	τ	0,000037	0,087767	0,103499	0,904	0,001423	0,037045	0,043831	0,927
KMP	$\tau/4$	−0,000138	0,142545	0,140246	0,947	0,001059	0,045059	0,046250	0,948
	$\tau/2$	−0,002223	0,094532	0,094133	0,966	−0,000615	0,029914	0,029441	0,958
	$3\tau/4$	−0,002118	0,079883	0,080318	0,947	0,001144	0,025441	0,025165	0,946
	τ	−0,001288	0,087398	0,102279	0,892	0,001100	0,036934	0,042927	0,928
COX	$\tau/4$	−0,005090	0,088907	0,086946	0,959	0,000840	0,028038	0,028106	0,950
	$\tau/2$	−0,003880	0,078274	0,076569	0,960	0,000802	0,024691	0,024751	0,951
	$3\tau/4$	−0,002599	0,066829	0,065284	0,961	0,000749	0,021082	0,021132	0,950
	τ	−0,001616	0,055994	0,054757	0,963	0,000710	0,017530	0,017553	0,950
RCM	$\tau/4$	−0,005043	0,089328	0,086615	0,962	0,000859	0,028113	0,027991	0,950
	$\tau/2$	−0,003854	0,078771	0,076399	0,963	0,000819	0,024798	0,024689	0,951
	$3\tau/4$	−0,002605	0,067346	0,065208	0,962	0,000765	0,021214	0,021119	0,951
	τ	−0,001554	0,056298	0,054658	0,961	0,000718	0,017695	0,017615	0,951

Tableau 3.22: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base croissant ($\gamma = 4/3$), un paramètre de régression $\beta = 0$ et une probabilité d'exposition égale à 0,75. Les valeurs théoriques du risque attribuable aux temps $\tau/4$, $\tau/2$, $3\tau/4$ et τ sont toutes nulles.

d'échantillon de 1 000 observations (Tableau 3.23, partie gauche), les estimateurs issus de l'approche semi-paramétrique sont biaisés (biais relatif compris entre 14,6 et 32,6 %) avec de faibles taux de recouvrement (< 93,4 %) sauf en $\tau/2$ où le taux de recouvrement est égal à 94,1 %. Les biais sont plus forts pour l'approche paramétrique RCM (biais relatif compris entre 14,6 et 81,6 %) avec des taux de recouvrement plus faibles encore (< 81 %). L'estimation du paramètre β (Tableau 3.24) est aussi biaisée pour les deux approches (biais relatif égal à 30,3 % et 40,4 % pour les approches paramétrique et semi-paramétrique respectivement). L'estimation du risque de base par la méthode paramétrique n'est pas non plus satisfaisante (Figure 3.7).

Avec $n = 10\,000$ (Tableau 3.23, partie droite), les biais restent élevés et deviennent similaires pour les deux approches semi-paramétrique COX et paramétrique RCM (biais relatifs entre 7,1 et 31,2 % et entre 8,3 et 32,0 % respectivement). Les taux de recouvrement se détériorent suite au resserrement des intervalles de confiance. L'estimation du paramètre β par les deux approches paramétrique et semi-paramétrique reste biaisée (Tableau 3.25). L'estimation du risque de base (Figure 3.8) s'améliore pour l'approche paramétrique par rapport à $n = 1\,000$.

Notons qu'avec une probabilité d'exposition plus faible ou plus importante ($p = 0{,}25$ ou 0,75 respectivement), les taux de recouvrement s'améliorent pour l'approche paramétrique à tous les temps mais restent inférieurs à 93 % en général (Tableaux 3.26 et 3.27).

En somme, lorsque les données sont générées selon un modèle à risques non proportionnels, les approches non paramétriques KM et KMP donnent les mêmes résultats qu'avec les risques proportionnels. Les approches semi-paramétrique COX et paramétrique RCM reposant sur l'hypothèse des risques proportionnels donnent de mauvais résultats. Les estimations du RA pour ces deux approches sont biaisées avec des biais plus importants pour l'approche paramétrique. Les taux de recouvrement sont faibles sauf en $\tau/2$ pour l'approche semi-paramétrique COX et une probabilité d'exposition à 0,50.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	0,001124	0,045053	0,045787	0,954	0,000289	0,014277	0,014126	0,949
	$\tau/2$	0,001330	0,037581	0,037647	0,953	−0,000029	0,011915	0,012154	0,935
	$3\tau/4$	0,001211	0,036543	0,036593	0,953	−0,000301	0,011618	0,011608	0,952
	τ	0,002743	0,043713	0,051764	0,933	−0,000888	0,016362	0,019957	0,950
KMP	$\tau/4$	0,001138	0,045090	0,045739	0,954	0,000291	0,014274	0,014130	0,949
	$\tau/2$	0,001347	0,037587	0,037593	0,956	−0,000024	0,011911	0,012151	0,938
	$3\tau/4$	0,001165	0,036511	0,036518	0,952	−0,000293	0,011612	0,011607	0,956
	τ	0,001685	0,042617	0,049261	0,920	−0,000708	0,016157	0,019107	0,946
COX	$\tau/4$	−0,018761	0,037521	0,037543	0,933	−0,019843	0,011869	0,011939	0,621
	$\tau/2$	0,010548	0,033500	0,033580	0,941	0,009504	0,010588	0,010676	0,847
	$3\tau/4$	0,023376	0,030960	0,031017	0,879	0,022314	0,009775	0,009879	0,368
	τ	0,030360	0,029427	0,029588	0,830	0,029168	0,009323	0,009456	0,127
RCM	$\tau/4$	0,026479	0,048525	0,049191	0,908	−0,017516	0,011688	0,012080	0,672
	$\tau/2$	0,057418	0,044915	0,045594	0,738	0,011082	0,010391	0,010768	0,806
	$3\tau/4$	0,070045	0,042342	0,043042	0,607	0,023478	0,009571	0,009936	0,313
	τ	0,075924	0,040403	0,041050	0,525	0,029848	0,009011	0,009360	0,098

Tableau 3.23: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques non proportionnels avec un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,5. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}181$, $A(\tau/2) = 0{,}133$, $A(3\tau/4) = 0{,}109$ et $A(\tau) = 0{,}093$.

Méthode	Valeur théorique	Minimum	Médiane	Moyenne	Maximum	SEE	SSS	IC empirique
Paramétrique	0,693	0,095	0,488	0,483	0,817	0,115	0,117	0,259 – 0,711
Semi-paramétrique	0,693	0,099	0,414	0,413	0,780	0,096	0,095	0,231 – 0,596

Tableau 3.24: Paramètre β estimé par les méthodes semi-paramétrique COX et paramétrique RCM pour un modèle à risques non proportionnels avec $n = 1\,000$, $\lambda_0 = 0{,}1$ et $p = 0{,}50$

Méthode	Valeur théorique	Minimum	Médiane	Moyenne	Maximum	SEE	SSD	IC empirique
Paramétrique	0,693	0,314	0,422	0,421	0,517	0,030	0,031	0,361 – 0,484
Semi-paramétrique	0,693	0,305	0,409	0,410	0,504	0,030	0,030	0,350 – 0,471

Tableau 3.25: Paramètre β estimé par les méthodes semi-paramétrique COX et paramétrique RCM pour un modèle à risques non proportionnels avec $n = 10\,000$, $\lambda_0 = 0{,}1$ et $p = 0{,}50$

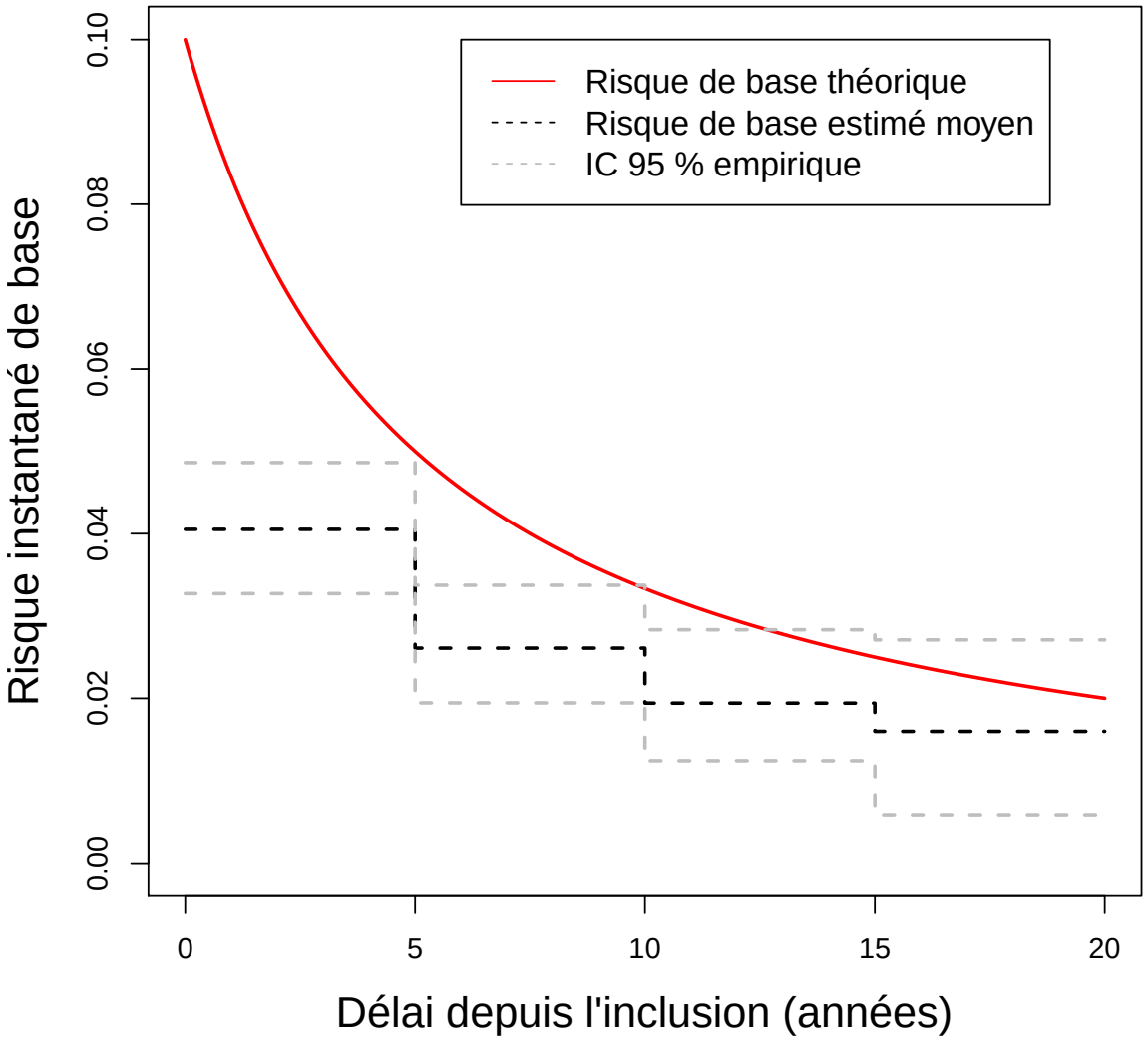


FIGURE 3.7: Estimation moyenne sur les 1 000 simulations du risque instantané de base par la méthode paramétrique RCM pour une durée de suivi divisée en quatre intervalles réguliers pour un modèle à risques non proportionnels avec $n = 1\,000$, $\lambda_0 = 0{,}1$, $p = 0{,}50$

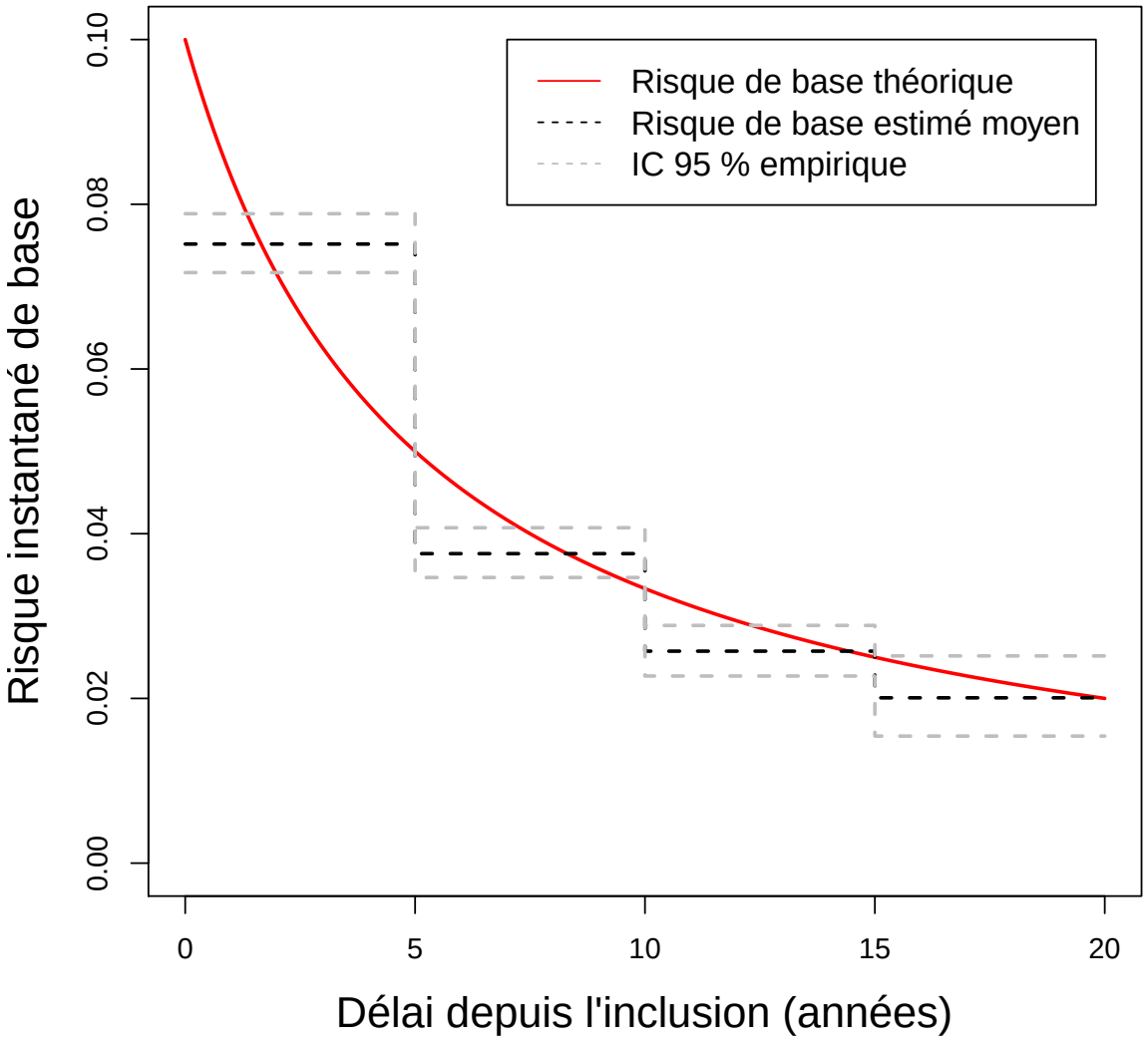


FIGURE 3.8: Estimation moyenne sur les 1 000 simulations du risque instantané de base par la méthode paramétrique RCM pour une durée de suivi divisée en quatre intervalles réguliers pour un modèle à risques non proportionnels avec $n = 10\,000$, $\lambda_0 = 0{,}1$, $p = 0{,}50$

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	0,000462	0,029469	0,029636	0,955	−0,000044	0,009336	0,009236	0,952
	$\tau/2$	0,000140	0,023432	0,023186	0,954	−0,000119	0,007426	0,007401	0,947
	$3\tau/4$	0,000137	0,022198	0,021663	0,959	−0,000187	0,007056	0,007089	0,952
	τ	0,000522	0,026657	0,032997	0,930	−0,000714	0,009557	0,010938	0,948
KMP	$\tau/4$	0,000505	0,029488	0,029571	0,955	−0,000038	0,009330	0,009226	0,951
	$\tau/2$	0,000191	0,023386	0,023131	0,952	−0,000112	0,007417	0,007390	0,946
	$3\tau/4$	0,000198	0,022085	0,021717	0,955	−0,000186	0,007042	0,007083	0,954
	τ	−0,000440	0,024510	0,029188	0,904	−0,000649	0,009327	0,010324	0,944
COX	$\tau/4$	−0,010517	0,024925	0,024526	0,924	−0,010962	0,007890	0,007863	0,695
	$\tau/2$	0,006063	0,021342	0,021016	0,941	0,005771	0,006753	0,006721	0,873
	$3\tau/4$	0,012675	0,019143	0,018770	0,901	0,012435	0,006049	0,006026	0,458
	τ	0,015999	0,017777	0,017416	0,876	0,015753	0,005635	0,005634	0,203
RCM	$\tau/4$	−0,009182	0,023920	0,024920	0,911	−0,009719	0,007569	0,007967	0,722
	$\tau/2$	0,006888	0,020355	0,021244	0,922	0,006542	0,006439	0,006777	0,820
	$3\tau/4$	0,013212	0,018179	0,018919	0,879	0,012945	0,005745	0,006057	0,395
	τ	0,016210	0,016730	0,017320	0,839	0,015968	0,005275	0,005547	0,157

Tableau 3.26: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques non proportionnels avec un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,25. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}100$, $A(\tau/2) = 0{,}071$, $A(3\tau/4) = 0{,}058$ et $A(\tau) = 0{,}049$.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
KM	$\tau/4$	−0,000572	0,069708	0,070140	0,945	−0,000031	0,022081	0,022198	0,951
	$\tau/2$	−0,000369	0,060473	0,061906	0,943	0,000510	0,019198	0,019521	0,946
	$3\tau/4$	−0,000754	0,060071	0,061293	0,942	0,000356	0,019140	0,018818	0,955
	τ	0,001640	0,068678	0,076171	0,923	0,000285	0,026382	0,030524	0,936
KMP	$\tau/4$	−0,000593	0,069753	0,070140	0,946	−0,000038	0,022081	0,022189	0,951
	$\tau/2$	−0,000380	0,060496	0,061948	0,944	0,000504	0,019195	0,019519	0,945
	$3\tau/4$	−0,000734	0,060034	0,061200	0,945	0,000359	0,019140	0,018814	0,951
	τ	0,001477	0,068370	0,075922	0,918	0,000272	0,026393	0,030368	0,939
COX	$\tau/4$	−0,028650	0,057643	0,057265	0,922	−0,027067	0,018204	0,018206	0,683
	$\tau/2$	0,010648	0,053081	0,052772	0,953	0,011831	0,016771	0,016772	0,889
	$3\tau/4$	0,029394	0,050152	0,049930	0,911	0,030160	0,015838	0,015832	0,529
	τ	0,040348	0,048261	0,048121	0,876	0,040477	0,015276	0,015340	0,228
RCM	$\tau/4$	−0,009182	0,023920	0,024920	0,911	−0,023872	0,018227	0,018382	0,746
	$\tau/2$	0,006888	0,020355	0,021244	0,922	0,014128	0,016763	0,016903	0,864
	$3\tau/4$	0,013212	0,018179	0,018919	0,879	0,031973	0,015811	0,015952	0,482
	τ	0,016210	0,016730	0,017320	0,839	0,041728	0,015130	0,015326	0,208

Tableau 3.27: Résultats de simulation du risque attribuable $A(.)$ sous le modèle à risques non proportionnels avec un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,75. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}249$, $A(\tau/2) = 0{,}188$, $A(3\tau/4) = 0{,}155$ et $A(\tau) = 0{,}134$.

3.3.6 Durée de suivi raccourcie

3.3.6.1 Résultats avec les méthodes non paramétriques

Lorsque nous arrêtons le suivi à $\tau/4$ ou $\tau/2$, les estimations par les méthodes non paramétriques KM et KMP à ces temps restent identiques à celles obtenues aux mêmes temps avec un suivi complet quels que soient les paramètres considérés puisque ces méthodes n'utilisent que l'information disponible jusqu'au temps où le RA est estimé.

3.3.6.2 Résultats pour les méthodes semi-paramétrique et paramétrique avec une durée d'étude arrêtée à mi-suivi

Avec un risque de base constant ($\gamma = 1$) et une probabilité d'exposition à 0,50 (Tableau 3.28), lorsque nous censurons à $\tau/2$, les estimations du RA par les méthodes semi-paramétrique COX et paramétrique RCM sont sans biais aux temps $\tau/4$ et $\tau/2$ avec de bons taux de recouvrement pour une taille d'échantillon de 1 000 observations (Tableau 3.28, partie gauche). Les écarts-types estimés moyens par les approches semi-paramétrique et paramétrique augmentent et se rapprochent de ceux estimés par les méthodes non paramétriques. Pour une taille d'échantillon de 10 000 observations (Tableau 3.28, partie droite), les biais deviennent plus faibles avec une augmentation de la précision de l'ordre de $\sqrt{10}$ comme dans le cas d'un suivi complet.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
COX	$\tau/4$	0,001108	0,042166	0,041908	0,956	0,000076	0,013348	0,013252	0,941
	$\tau/2$	0,001261	0,036709	0,036775	0,955	0,000192	0,011607	0,011515	0,943
RCM	$\tau/4$	0,001296	0,041696	0,041947	0,955	0,000108	0,013189	0,013262	0,939
	$\tau/2$	0,001360	0,036272	0,036784	0,951	0,000205	0,011462	0,011516	0,939

Tableau 3.28: Résultats de simulation à mi suivi du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base constant ($\gamma = 1$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,5. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}284$, $A(\tau/2) = 0{,}240$.

Avec une probabilité d'exposition plus faible ($p = 0{,}25$, Tableau 3.29) ou plus forte ($p = 0{,}75$, Tableau 3.30), les résultats obtenus sont satisfaisants avec un rapprochement en moyenne des écarts-types estimés par les méthodes semi-paramétrique et paramétrique de ceux estimés par les approches non paramétriques.

Pour un risque de base décroissant ($\gamma = 3/4$) avec une probabilité d'exposition à 0,50 (Tableau 3.31), les biais sont faibles en $\tau/2$ et $\tau/4$ et les écarts-types estimés moyens se rapprochent de ceux obtenus par les approches non paramétriques avec de bons taux de recouvrement pour $n = 1\,000$ (Tableau 3.31, partie gauche). Lorsque $n = 10\,000$, les biais restent faibles avec des taux de recouvrement qui se détériorent en $\tau/4$ et $\tau/2$ pour les deux approches paramétrique et semi-paramétrique sauf en $\tau/4$ pour l'approche semi-paramétrique.

Avec une plus faible probabilité d'exposition ($p = 0{,}25$, Tableau 3.32), les écarts-types estimés par les méthodes semi-paramétrique et paramétrique se rapprochent en moyenne de ceux estimés par les méthodes non paramétriques. Les taux de recouvrement se détériorent pour l'approche paramétrique quelle que soit la taille de l'échantillon.

Lorsque la prévalence de l'exposition est plus importante ($p = 0{,}75$, Tableau 3.33), les estimations du RA par les méthodes semi-paramétrique et paramétrique sont sans biais. Pour $n = 1\,000$ (Tableau 3.30, partie gauche), les résultats sont satisfaisants avec

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
COX	$\tau/4$	0,000360	0,030540	0,029924	0,956	−0,000230	0,009667	0,009701	0,948
	$\tau/2$	0,000241	0,024475	0,024023	0,956	−0,000128	0,007738	0,007747	0,951
RCM	$\tau/4$	0,000487	0,029607	0,029923	0,950	−0,000214	0,009366	0,009709	0,941
	$\tau/2$	0,000312	0,023619	0,024038	0,949	−0,000121	0,007466	0,007750	0,939

Tableau 3.29: Résultats de simulation à mi-suivi du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base constant ($\gamma = 1$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,25. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}166$, $A(\tau/2) = 0{,}136$.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
COX	$\tau/4$	−0,002156	0,060229	0,059310	0,953	0,000102	0,019039	0,018949	0,957
	$\tau/2$	−0,001315	0,055751	0,054983	0,953	0,000171	0,017637	0,017579	0,954
RCM	$\tau/4$	−0,001933	0,060119	0,059359	0,950	0,000157	0,018989	0,018964	0,955
	$\tau/2$	−0,001195	0,055649	0,054992	0,953	0,000189	0,017592	0,017575	0,952

Tableau 3.30: Résultats de simulation à mi-suivi du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base constant ($\gamma = 1$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,75. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}373$, $A(\tau/2) = 0{,}321$.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
COX	$\tau/4$	0,000802	0,038121	0,038028	0,952	0,000326	0,012065	0,012026	0,938
	$\tau/2$	0,000948	0,033557	0,033580	0,953	0,000322	0,010607	0,010634	0,934
RCM	$\tau/4$	0,004148	0,037106	0,038430	0,938	0,003526	0,011736	0,012133	0,925
	$\tau/2$	0,003320	0,032630	0,033919	0,939	0,002636	0,010311	0,010714	0,925

Tableau 3.31: Résultats de simulation à mi-suivi du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base décroissant ($\gamma = 3/4$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,50. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}269$, $A(\tau/2) = 0{,}231$.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
COX	$\tau/4$	0,000047	0,027013	0,026776	0,950	0,000050	0,008549	0,008403	0,955
	$\tau/2$	−0,000035	0,022105	0,021880	0,954	0,000050	0,006988	0,006867	0,953
RCM	$\tau/4$	0,002422	0,025695	0,027210	0,934	0,002309	0,008128	0,008521	0,922
	$\tau/2$	0,001495	0,020859	0,022170	0,933	0,001535	0,006594	0,006939	0,922

Tableau 3.32: Résultats de simulation à mi-suivi du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base décroissant ($\gamma = 3/4$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,25. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}156$, $A(\tau/2) = 0{,}130$.

de bons taux de recouvrement en $\tau/4$ et $\tau/2$ et les écarts-types estimés se rapprochent en moyenne de ceux estimés par les méthodes non paramétriques. Pour $n = 10\,000$ (Tableau 3.30, partie droite), les estimations du RA restent sans biais et les écarts-types estimés moyens se rapprochent également de ceux estimés par les méthodes non paramétriques. Pour l'approche paramétrique, le taux de recouvrement en $\tau/4$ est inférieur à la valeur nominale.

Enfin, pour un risque de base croissant ($\gamma = 4/3$) avec une probabilité d'exposition à 0,50, 0,25 ou 0,75 (Tableaux 3.34, 3.35 et 3.36 respectivement), les résultats obtenus sont satisfaisants avec des écarts-types estimés moyens assez proches de ceux obtenus par les méthodes non paramétriques.

3.3.6.3 Résultats pour les méthodes semi-paramétrique et paramétrique avec une durée d'étude arrêtée au quart de suivi

Lorsque nous censurons à $\tau/4$, les estimations du RA obtenues par les méthodes semi-paramétrique COX et paramétrique RCM sont sans biais avec de bons taux de recouvrement en $\tau/4$, pour un risque de base constant ($\gamma = 1$) et une probabilité d'exposition à 0,50 (Tableau 3.37). Les écarts-types estimés par les méthodes semi-paramétrique et paramétrique se rapprochent en moyenne de ceux estimés par les méthodes non pa-

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
COX	$\tau/4$	−0,001232	0,055381	0,055890	0,945	0,000021	0,017507	0,017663	0,948
	$\tau/2$	−0,000480	0,051499	0,052277	0,943	0,000081	0,016283	0,016430	0,948
RCM	$\tau/4$	0,002377	0,054482	0,056223	0,937	0,003563	0,017211	0,017765	0,932
	$\tau/2$	0,002304	0,050723	0,052627	0,937	0,002841	0,016029	0,016534	0,937

Tableau 3.33: Résultats de simulation à mi-suivi du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base décroissant ($\gamma = 3/4$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,75. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}356$, $A(\tau/2) = 0{,}310$.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
COX	$\tau/4$	0,001458	0,047231	0,046736	0,953	0,000082	0,014957	0,014803	0,948
	$\tau/2$	0,001599	0,041001	0,040631	0,950	0,000194	0,012964	0,012811	0,946
RCM	$\tau/4$	−0,000500	0,047055	0,046511	0,949	−0,002059	0,014890	0,014718	0,948
	$\tau/2$	0,000574	0,040885	0,040498	0,950	−0,000950	0,012919	0,012766	0,944

Tableau 3.34: Résultats de simulation à mi-suivi du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base croissant ($\gamma = 4/3$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,50. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}299$, $A(\tau/2) = 0{,}250$.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
COX	$\tau/4$	0,000801	0,034934	0,034434	0,958	−0,000242	0,011067	0,011030	0,952
	$\tau/2$	0,000599	0,027744	0,027380	0,961	−0,000143	0,008775	0,008724	0,957
RCM	$\tau/4$	−0,000645	0,034191	0,034100	0,954	−0,001755	0,010824	0,010933	0,942
	$\tau/2$	−0,000095	0,027155	0,027225	0,957	−0,000894	0,008584	0,008683	0,953

Tableau 3.35: Résultats de simulation à mi-suivi du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base croissant ($\gamma = 4/3$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,25. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}176$, $A(\tau/2) = 0{,}143$.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
COX	$\tau/4$	−0,001487	0,066424	0,065350	0,944	−0,000128	0,020988	0,020875	0,957
	$\tau/2$	−0,000551	0,061552	0,060635	0,951	−0,000021	0,019459	0,019313	0,954
RCM	$\tau/4$	−0,003645	0,066758	0,065197	0,948	−0,002507	0,021083	0,020806	0,951
	$\tau/2$	−0,001755	0,061868	0,060512	0,954	−0,001348	0,019550	0,019258	0,958

Tableau 3.36: Résultats de simulation à mi-suivi du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base croissant ($\gamma = 4/3$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,75. Les valeurs théoriques du risque attribuable sont : $A(\tau/4) = 0{,}390$, $A(\tau/2) = 0{,}334$.

ramétriques. Nous obtenons les mêmes conclusions pour une probabilité d'exposition plus faible ($p = 0{,}25$, Tableau 3.38) ou plus élevée ($p = 0{,}75$, Tableau 3.39).

Pour un risque de base décroissant ($\gamma = 3/4$) avec une probabilité d'exposition à 0,50 (Tableau 3.40), les estimations du RA restent sans biais. Les résultats sont satisfaisants pour l'approche semi-paramétrique mais le sont moins pour l'approche paramétrique où le taux de recouvrement en $\tau/4$ est égal à 93,5 % et 91,8 % pour $n = 1\,000$ et $10\,000$ respectivement. Les écarts-types estimés moyens se rapprochent également de ceux des méthodes non paramétriques en $\tau/4$.

Avec une probabilité d'exposition plus faible ($p = 0{,}25$, Tableau 3.41) comme avec une probabilité d'exposition importante ($p = 0{,}75$, Tableau 3.42), les estimations du RA restent sans biais. Les résultats sont satisfaisants pour l'approche semi-paramétrique mais le sont moins pour l'approche paramétrique et une taille d'échantillon plus importante où le taux de recouvrement est égal à 92,0 %. Les écarts-types estimés moyens se rapprochent également de ceux des méthodes non paramétriques en $\tau/4$.

Lorsque le risque de base est croissant ($\gamma = 4/3$, Tableaux 3.43, 3.44 et 3.45), les résultats sont similaires à ceux obtenus pour un risque de base constant ($\gamma = 1$) quelle que soit la probabilité d'exposition.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
COX	$\tau/4$	0,001043	0,051759	0,051582	0,943	0,000010	0,016374	0,016031	0,946
RCM	$\tau/4$	0,001185	0,051380	0,051600	0,942	0,000040	0,016245	0,016039	0,944

Tableau 3.37: Résultats de simulation au quart de suivi du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base constant ($\gamma = 1$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,50. La valeur théorique du risque attribuable en $\tau/4$ est 0,284.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
COX	$\tau/4$	0,000408	0,036873	0,036106	0,956	−0,000350	0,011665	0,011632	0,950
RCM	$\tau/4$	0,000535	0,036113	0,036126	0,952	−0,000334	0,011417	0,011637	0,949

Tableau 3.38: Résultats de simulation au quart de suivi du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base constant ($\gamma = 1$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,25. La valeur théorique du risque attribuable en $\tau/4$ est 0,166.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
COX	$\tau/4$	−0,001912	0,074752	0,073415	0,960	0,000494	0,023643	0,023764	0,955
RCM	$\tau/4$	−0,001770	0,074674	0,073398	0,960	0,000539	0,023601	0,023749	0,953

Tableau 3.39: Résultats de simulation au quart de suivi du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base constant ($\gamma = 1$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,75. La valeur théorique du risque attribuable en $\tau/4$ est 0,373.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
COX	$\tau/4$	0,001137	0,043862	0,044425	0,947	0,000163	0,013877	0,013782	0,948
RCM	$\tau/4$	0,006320	0,042795	0,045111	0,935	0,005245	0,013535	0,013987	0,918

Tableau 3.40: Résultats de simulation au quart de suivi du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base décroissant ($\gamma = 3/4$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,5. La valeur théorique du risque attribuable en $\tau/4$ est 0,269.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
COX	$\tau/4$	0,000285	0,030666	0,030374	0,952	−0,000160	0,009700	0,009597	0,950
RCM	$\tau/4$	0,003793	0,029436	0,031110	0,936	0,003254	0,009307	0,009808	0,920

Tableau 3.41: Résultats de simulation au quart de suivi du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base décroissant ($\gamma = 3/4$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,25. La valeur théorique du risque attribuable en $\tau/4$ est 0,156.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
COX	$\tau/4$	−0,000921	0,064327	0,063936	0,954	0,000212	0,020344	0,020677	0,945
RCM	$\tau/4$	0,004910	0,063174	0,064545	0,948	0,006077	0,019970	0,020852	0,923

Tableau 3.42: Résultats de simulation au quart de suivi du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base décroissant ($\gamma = 3/4$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,75. La valeur théorique du risque attribuable en $\tau/4$ est 0,356.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
COX	$\tau/4$	0,000521	0,063756	0,063799	0,949	0,000028	0,020180	0,019910	0,956
RCM	$\tau/4$	−0,002812	0,063869	0,063275	0,947	−0,003489	0,020199	0,019734	0,955

Tableau 3.43: Résultats de simulation au quart de suivi du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base croissant ($\gamma = 4/3$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,5. La valeur théorique du risque attribuable en $\tau/4$ est 0,299.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
COX	$\tau/4$	0,000534	0,046249	0,044985	0,959	−0,000364	0,014658	0,014565	0,958
RCM	$\tau/4$	−0,001807	0,045772	0,044354	0,958	−0,002792	0,014493	0,014348	0,954

Tableau 3.44: Résultats de simulation au quart de suivi du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base croissant ($\gamma = 4/3$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,25. La valeur théorique du risque attribuable en $\tau/4$ est 0,176.

Méthode		$n = 1\,000$				$n = 10\,000$			
d'estimation	Temps	Biais	SEE	SSD	PC	Biais	SEE	SSD	PC
COX	$\tau/4$	−0,001070	0,090759	0,088571	0,948	0,000438	0,028689	0,029375	0,953
RCM	$\tau/4$	−0,004819	0,091495	0,088178	0,953	−0,003519	0,028903	0,029240	0,948

Tableau 3.45: Résultats de simulation au quart de suivi du risque attribuable $A(.)$ sous le modèle à risques proportionnels avec un risque de base croissant ($\gamma = 4/3$), un paramètre de régression $\beta = \ln(2)$ et une probabilité d'exposition égale à 0,75. La valeur théorique du risque attribuable en $\tau/4$ est 0,390.

En résumé, lorsque nous censurons à $\tau/2$ ou $\tau/4$ les résultats obtenus pour les approches non paramétriques KM et KMP sont identiques à ceux obtenus avec un suivi complet. Les estimations du RA pour les approches semi-paramétrique et paramétrique sont sans biais avec des écarts-types estimés moyens qui se rapprochent de ceux obtenus par les approches non paramétriques lorsque nous censurons à $\tau/2$ et deviennent identiques lorsque nous censurons à $\tau/4$. Les taux de recouvrement sont satisfaisants pour l'approche semi-paramétrique sauf en $\tau/2$ avec une probabilité d'exposition à 0,50 et une taille d'échantillon de 10 000 observations lorsque nous censurons à $\tau/2$. Les taux de recouvrement obtenus pour l'approche paramétrique sont également satisfaisants sauf dans le cas d'un risque de base décroissant ($\gamma = 3/4$) où ils sont souvent inférieurs à la valeur attendue.

3.4 Application à des données de cohorte

3.4.1 Présentation de la cohorte E3N

L'étude E3N, Étude Épidémiologique auprès de femmes de la Mutuelle Générale de l'Éducation Nationale (MGEN), est une enquête de cohorte prospective auprès de 98 995 femmes nées entre 1925 et 1950 et suivies depuis 1990 [87,88]. Elle a pour objectif principal d'identifier et d'analyser le rôle de certains facteurs de risque dans la survenue des cancers de la femme. Le suivi est effectué par auto-questionnaire envoyé par courrier postal sous format papier, tous les deux à trois ans environ. Au total, 11 questionnaires ont été envoyés à ce jour.

L'identification des cas de cancer du sein est principalement réalisée aux questionnaires de suivi. Quelques cas de cancers ont été identifiés grâce aux informations transmises par les proches, grâce aux investigations auprès des médecins des participantes mais également grâce au registre des causes de décès.

Le recueil des informations sur les traitements hormonaux de la ménopause (THM) a débuté au deuxième questionnaire (Q2, envoyé en janvier 1992). Il est mis à jour

dans chacun des questionnaires suivants accompagnés d'un livret de photos aidant à la mémorisation des participants [89].

Les différents traitements ont été regroupés en quatre classes :

— Estrogène seul

— Estrogène + Progestérone micronisée ou dydrogestérone

— Estrogène + autre progestatif : chlormadinone acétate, cyproterone acétate, démégestone, dienogest, drospirénone, ethynodiol acétate, gestodène, lévonorgestrel, lynestrénol, médrogestone, médroxyprogestérone acétate, megestrol acétate, noménegestrol acétate, noréthistérone acétate et promégestone

— Autre THM : cette catégorie comprend la tibolone, les THM de type non spécifié, les traitements injectables et les THM associant un estrogène et un androgène.

Un travail séparé effectué sur les données de la cohorte E3N a estimé que 14,5 % des cas de cancer du sein chez les femmes ménopausées étaient attribuables à une utilisation récente de THM après un suivi de 15 ans [86].

3.4.2 Population et méthodes

Sur les 98 995 femmes de la cohorte recrutées en 1990, nous avons exclu celles non ménopausées à Q2, celles sans suivi au-delà de Q2, celles atteintes d'un cancer du sein avant Q2 et celles qui n'ont pas répondu au questionnaire Q2. Notre étude inclut ainsi 38 359 femmes ménopausées et sans antécédent de cancer à Q2.

Nous avons utilisé le temps de suivi, calculé comme la différence entre la date d'entrée dans l'étude et la date de la dernière réponse avant cancer ou décès, comme échelle de temps dans nos analyses pour le calcul du RA mais aussi l'âge comme échelle de temps pour vérifier la compatibilité des estimations du HR (*hazard ratio*) avec celles obtenues dans l'étude originale [85]. Nous avons estimé les fonctions de survie en utilisant un modèle paramétrique de Weibull et la méthode non paramétrique de Kaplan-Meier.

Après vérification de l'hypothèse des risques proportionnels avec les résidus de Schoenfeld, nous avons estimé les risques relatifs de cancer du sein associés à l'utilisation de THM

à l'inclusion en utilisant un modèle de Cox à risques proportionnels. Enfin, nous avons estimé le RA à l'exposition aux THM (avoir utilisé au moins une fois des THM avant Q2) et son intervalle de confiance en utilisant les quatre méthodes choisies pour l'étude de simulation, à savoir les deux approches non paramétriques KM et KMP [6], l'approche semi-paramétrique COX [6] et l'approche paramétrique RCM [3]. Nous avons aussi utilisé l'approche de Spiegelman *et al.* [31] pour estimer un RA unique à l'échelle de la cohorte.

3.4.3 Résultats

Les 38 359 participantes sélectionnées ont été suivies en moyenne pendant 14,0 années (étendue : 0,0 à 16,4). Parmi ces femmes, 17 185 (44,8 %) étaient considérées comme exposées à l'inclusion. L'âge moyen (écart-type) à l'inclusion était de 56,4 (5,3) années. Au cours du suivi, un cancer du sein invasif a été diagnostiqué chez 2 228 femmes parmi lesquelles 1 106 n'avaient jamais utilisé de THM avant Q2 (non exposées). Le nombre de cas incidents de cancer du sein invasif était de 471 à 4 ans de suivi, 580 à 8 ans, 655 à 12 ans et 522 à 16 ans.

En utilisant un modèle paramétrique de Weibull, les paramètres de position et d'échelle sont estimés à 1,2 et 178,2 respectivement. Les estimations de la fonction de survie chez les exposées et chez les non exposées selon un modèle paramétrique de Weibull et selon l'approche non paramétrique de Kaplan-Meier sont très proches (Figure 3.9).

Le test des résidus de Schoenfeld est non significatif ($p = 0{,}7$), ce qui ne contredit pas l'hypothèse des risques proportionnels. En utilisant un modèle de Cox à risques proportionnels, le risque estimé de développer un cancer du sein invasif est 1,22 fois plus important (IC 95 %, 1,13 à 1,33) chez les exposées aux THM que chez les non exposées à l'inclusion. Avec l'âge comme échelle de temps, nous estimons un HR à 1,24 (IC 95 % 1,14 à 1,34) proche de celui obtenu avec le temps de suivi comme échelle de temps (Tableau 3.46).

Les estimations du RA par les approches non paramétriques KM et KMP sont pratiquement identiques (Figure 3.10). Elles sont croissantes jusqu'à environ 12 ans de suivi.

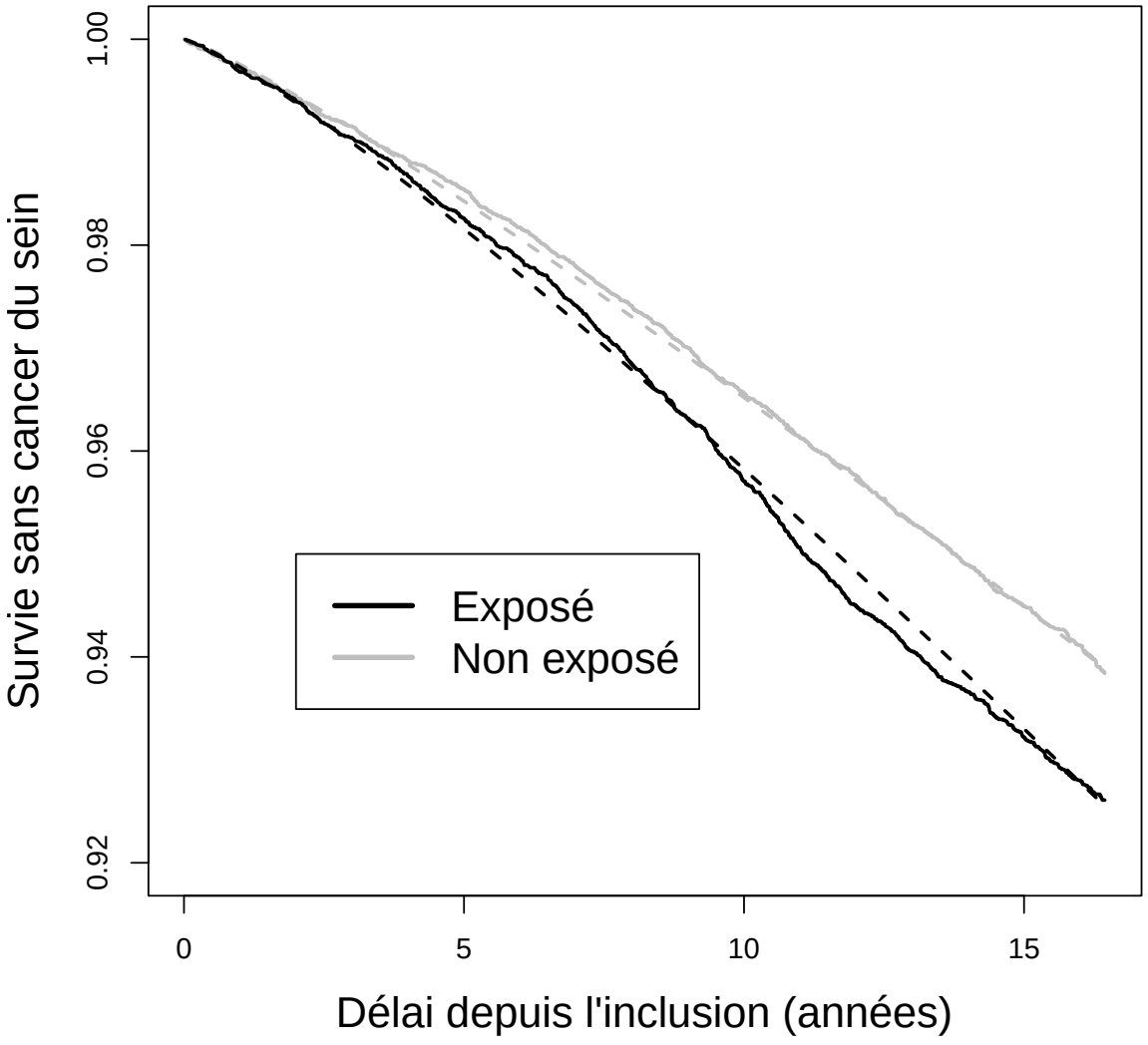


FIGURE 3.9: Estimation de la fonction de survie chez les exposées et chez les non exposées selon l'approche non paramétrique de Kaplan-Meier (trait plein) et une approche paramétrique de Weibull (trait pointillé), cohorte E3N, 1992-2008

Échelle de temps	β	Écart-type	HR	IC à 95 %
Temps de suivi	0,04	0,20	1,22	1,13 – 1,33
Âge	0,04	0,21	1,24	1,14 – 1,34

Tableau 3.46: Estimation du risque relatif associé à l'utilisation de THM par le modèle de Cox avec l'âge et le temps de suivi comme échelle de temps, cohorte E3N, 1992-2008

Par exemple, pour l'approche KM, elles varient de 5,5 % (IC 95 %, -2,7 % à 13,6 %) après 4 ans de suivi à 12,0 % (IC 95 %, 7,8 % à 16,2 %) après 12 ans de suivi. Ensuite, ces estimations décroissent et tendent vers celles des méthodes semi-paramétrique COX et paramétrique RCM avec une valeur estimée à 9,2 % (IC 95 %, 5,4 % à 13,0 %) après 16 ans de suivi. En comparaison, lorsque nous utilisons les méthodes semi-paramétrique et paramétrique, les estimations sont légèrement décroissantes au cours du suivi, allant de 9,0 % (IC 95 %, 5,3 % à 12,8 %) à 8,8 % (IC 95 %, 5,1 % à 12,4 %) et de 8,9 % (IC 95 %, 5,2 % à 12,6 %) à 8,7 % (IC 95 %, 5,0 % à 12,3 %) respectivement. Ainsi, après 16 ans de suivi, la proportion de cas de cancer du sein invasif attribuable à l'utilisation de THM à l'inclusion est proche de 9 % quelle que soit la méthode considérée. Les estimations du RA par les approches non paramétriques ont une précision moindre avec des intervalles de confiance plus larges que celles des méthodes semi-paramétrique et paramétrique dans la première moitié de suivi. Par exemple, après 8 ans de suivi, le RA est estimé à 8,9 % (IC 95 %, 3,5 % à 14,4 %) et 9,0 % (IC 95 %, 5,2 % à 12,7 %) par les approches KM et COX respectivement.

Enfin, avec la méthode paramétrique de Spiegelman *et al.* [31], la proportion de cas de cancer du sein invasif attribuables à l'utilisation de THM à l'inclusion est estimée à 9,2 % (IC 95 %, 5,4 % à 12,7 %) pour l'ensemble du suivi.

3.5 Discussion

3.5.1 Synthèse des résultats de simulation

Ce premier travail compare différentes méthodes d'estimation du RA proposées lorsque la probabilité de la maladie est interprétée comme une fonction de répartition [3, 6, 7]. Notre étude de simulation montre que les estimateurs du RA sont essentiellement sans biais pour toutes les approches que nous considérons lorsque les temps d'événement sont générés selon un modèle à risques proportionnels. La variance estimée est proche de la variance empirique avec des taux de recouvrement satisfaisants sauf en fin d'étude pour

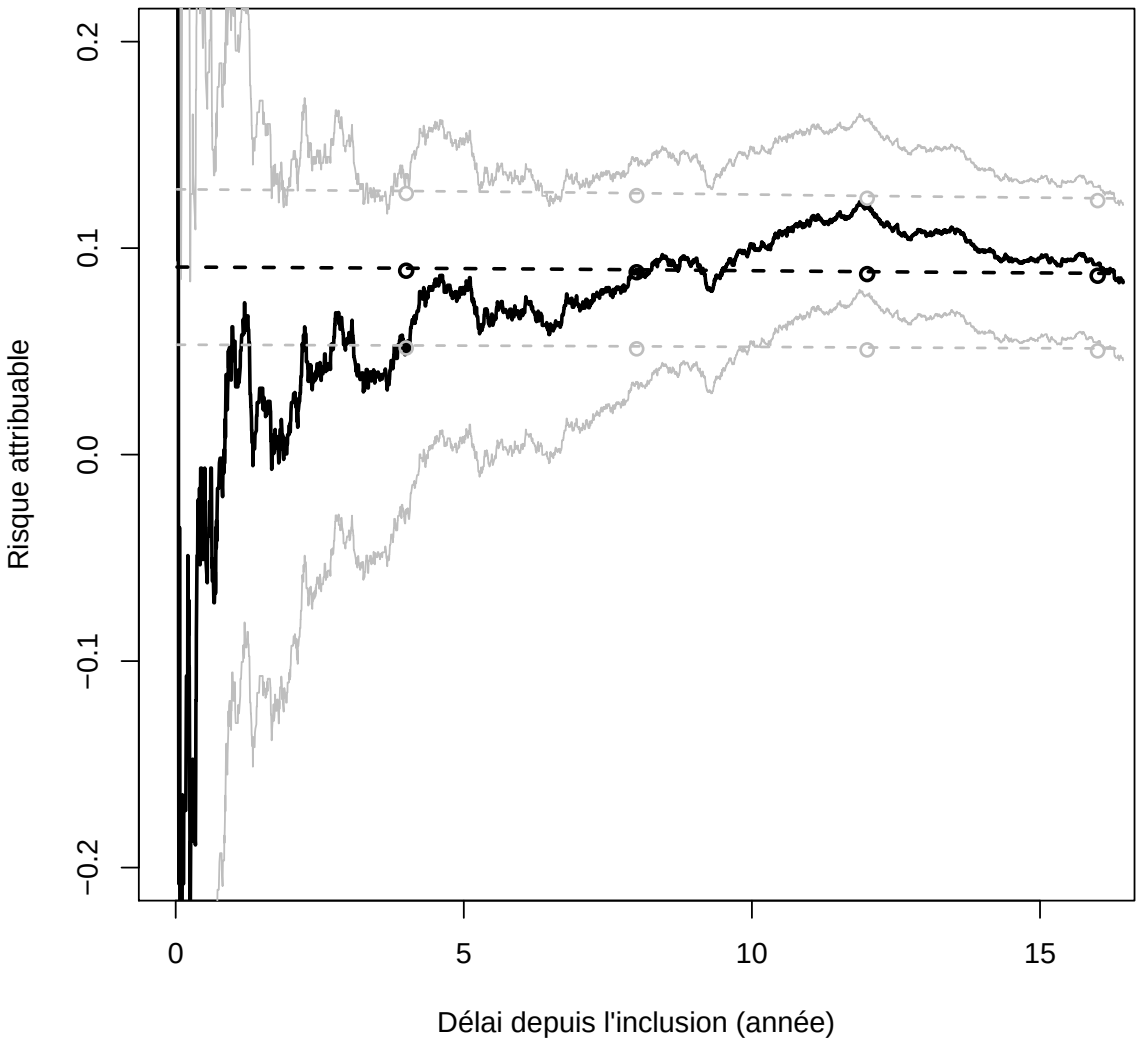


FIGURE 3.10: Estimation du risque de cancer du sein invasif attribuable à l'utilisation de traitements hormonaux de la ménopause à l'inclusion, cohorte E3N, 1992-2008.
Les courbes noires en trait plein et pointillé représentent les estimations selon les approches KM et COX respectivement, les cercles noirs représentent les estimations pour la méthode paramétrique RCM; les courbes en gris (trait plein et pointillé) et les cercles en gris représentent les intervalles de confiance à 95 % correspondants. L'approche KMP n'est pas représentée car elle coïncide avec l'approche KM à l'échelle de la figure.

les méthodes non paramétriques et des échantillons de taille plus petite. Pour un risque de base dépendant du temps (décroissant ou croissant), les estimations par l'approche paramétrique s'avèrent robustes malgré la mauvaise spécification du risque de base. Cependant, quand nous générons les données selon un modèle à risques non proportionnels, les estimations obtenues par les approches semi-paramétrique et paramétrique sont biaisées avec de faibles taux de recouvrement.

3.5.2 Comparaison aux travaux antérieurs

L'étude de simulation que nous avons réalisée constitue la première comparaison systématique de méthodes non paramétriques, semi-paramétrique et paramétrique pour l'estimation du RA dans divers scénarios : risques proportionnels ou non, risque de base constant ou non, probabilités d'exposition variables, forces d'association et tailles d'échantillon variées. Chen *et al.* [6] ont réalisé des études de simulation pour les méthodes KM, KMP et les modèles de transformation en générant les temps d'événement à partir de modèles à risques proportionnels et non proportionnels, avec un paramètre de régression $\beta = 1$ et 40 % d'exposés ($p = 0{,}40$) pour 1 000 individus. Comme ces auteurs, nous avons trouvé que, sous l'hypothèse d'une censure indépendante, les résultats des approches non paramétriques KM et KMP sont très proches. Des différences apparaissaient entre les deux méthodes non paramétriques lorsque la censure est dépendante des covariables [6], scénario que nous n'avons pas évalué dans notre étude de simulation.

Aussi comme Chen *et al.* [6], lorsque nous générons les temps d'événement sous l'hypothèse de risques proportionnels, nous trouvons que les estimations obtenues par les modèles non paramétriques et semi-paramétrique sont sans biais, que les estimateurs non paramétriques ont une variance plus grande que les estimateurs semi-paramétriques et que les variances estimées reflètent correctement la vraie variance sauf en τ pour les approches non paramétriques et une taille d'échantillon de 1 000 observations. Les approches non paramétriques donnent de meilleurs résultats (respectivement moins bons) lorsque la prévalence de l'exposition est plus faible (respectivement plus importante). En

effet, lorsque la proportion de sujets non exposés est faible, l'estimateur de Kaplan-Meier pour la fonction de survie chez les non exposés est instable et inéfficace [6]. Ce comportement général est observé dans nos simulations lorsque les temps d'événement sont générés selon un risque instantané de base constant, croissant ou décroissant. Cependant, notons que, pour une taille d'échantillon plus grande, les écarts entre les écarts-types estimés et empirique tendent à diminuer, avec des probabilités de couverture en τ satisfaisantes pour les approches non paramétriques.

Pour les risques non proportionnels, nous avons généré les temps d'événement en utilisant un modèle de transformation considéré par Chen *et al.* [6] et trouvons des résultats cohérents pour les approches non paramétriques et similaires à ceux obtenus pour le modèle à risques proportionnels. Cependant, alors que Chen *et al.* [6] ont appliqué le même modèle à risques non proportionnels pour la génération et l'analyse (estimation du RA) des données, nous avons généré nos données sous l'hypothèse d'un modèle à risques non proportionnels et estimé le RA en utilisant un modèle de Cox à risques proportionnels (mal spécifié donc). Ceci explique la moindre performance que nous observons lorsque l'hypothèse des risques proportionnels n'est pas vérifiée, contrairement aux résultats satisfaisants obtenus par Chen *et al.* [6].

Une autre originalité de notre travail est l'évaluation à l'aide de simulations de l'approche paramétrique proposée par Laaksonen *et al.* [3] pour estimer le RA et la comparaison de cette dernière aux approches non paramétriques et semi-paramétrique. Quand les données générées vérifient l'hypothèse des risques proportionnels, les résultats des approches paramétrique et semi-paramétrique sont généralement proches dans nos simulations comme dans l'exemple d'application. Remarquons que l'approche paramétrique apparaît robuste pour prendre en compte un risque instantané de base décroissant ou croissant (au lieu de constant) et des risques proportionnels. Cependant, de même que l'approche semi-paramétrique basée sur le modèle de Cox, l'approche paramétrique s'avère sensible à l'hypothèse des risques proportionnels, donnant des résultats médiocres dans nos simulations lorsque cette hypothèse n'est pas vérifiée.

Enfin, comme Chen *et al.* [6] dans leurs simulations et leur exemple, nous observons une plus grande imprécision des estimateurs non paramétriques au début du suivi, ce qui peut expliquer les valeurs négatives du RA estimé en début de suivi. Ce résultat pouvait être attendu car l'estimation des fonctions de survie dépend des informations disponibles jusqu'au temps considéré et peu d'événements sont survenus jusque là. Ceci diffère des méthodes semi-paramétrique et paramétrique qui ont l'avantage d'estimer le paramètre β en utilisant les informations disponibles sur la totalité de la durée de suivi. De façon cohérente, avec un suivi plus court, les variances estimées par les approches semi-paramétrique et paramétrique sont plus grandes et plus proches de celles estimées par les méthodes non paramétriques.

3.5.3 Comparaison entre l'étude de simulation et l'application aux données réelles

Nous avons choisi nos paramètres de simulation pour nous rapprocher de cohortes épidémiologiques réelles. Celles-ci incluent souvent plusieurs milliers de participants suivis pendant plusieurs années. Dans notre application, l'hypothèse des risques proportionnels semble vérifiée. L'estimateur de Kaplan-Meier des fonctions de survie donne des résultats proches de l'approche paramétrique de Weibull. Si l'on considére une distribution de Weibull, le risque de base estimé est croissant et le paramètre de forme est compris entre les valeurs $\gamma = 1$ et $4/3$ que nous avons considérées dans notre étude de simulation. En revanche, comme dans beaucoup de cohortes épidémiologiques, le pourcentage de censure était très important (94,2 %) et beaucoup plus élevé dans notre exemple que dans les données simulées. Le faible nombre d'événements observé dans notre application peut expliquer la croissance apparente des estimations du RA par les méthodes non paramétriques jusqu'aux trois quarts de suivi, ce qui reste compatible néanmoins avec la tendance à la décroissance qui serait davantage attendue avec un RR constant. Chen *et al.* [6] observaient le même phénomène jusqu'au tiers du suivi dans leur application.

En utilisant l'approche proposée par Spiegelman *et al.* [31], l'estimation du RA global pour une exposition aux THM à l'inclusion était de 9,2 % dans la cohorte E3N. Ce résultat est proche de ceux obtenus en fin de suivi pour les méthodes non paramétriques KM et KMP et en début de suivi pour les approches paramétrique RCM et semi-paramétrique COX. Dans une publication récente, Dartois *et al.* [86] ont obtenu une estimation du RA de cancer du sein invasif plus élevée à 14,5 % (IC 95 %, 9,2 à 19,6 %) pour une exposition récente (dans les douze derniers mois) aux THM et chez les femmes ménopausées à partir des données de la cohorte E3N, en utilisant la même approche proposée par Spiegelman *et al.* [31] mais une analyse plus fine, ajustée sur d'autres facteurs de risque de cancer du sein et l'exposition aux THM comme variable dépendante du temps. Lorsque nous avons considéré l'utilisation des THM comme variable dépendante du temps dans notre modèle, nous avons trouvé des estimations de RR et RA plus proches de celles de Dartois *et al.* [86].

3.5.4 Limites de l'étude

Tout d'abord, nous n'avons pas évalué le RA ajusté sur des covariables. L'ajustement sur plusieurs variables est une pratique courante en épidémiologie pour réduire le biais de confusion, en particulier l'ajustement sur l'âge qui peut aussi être utilisé comme échelle de temps [79, 90]. Dans notre exemple, en utilisant des analyses non ajustées, nous avons trouvé une association statistiquement significative entre l'utilisation de THM avant l'inclusion et le risque de cancer du sein, en accord avec les résultats de l'étude originale où les auteurs ont utilisé des modèles de Cox plus complexes avec l'âge comme échelle de temps et l'ajustement sur d'autres covariables [85]. Les logiciels disponibles permettent l'ajustement sur des covariables pour les approches semi-paramétrique et paramétrique [3, 6] de même que l'approche de Spiegelman *et al.* [31]. Les approches non paramétriques en revanche requièrent un nombre de covariables limité et ne permettent pas l'ajustement sur des variables continues. De plus, il faudrait adapter les programmes disponibles pour prendre en compte la troncature à gauche résultant de l'utilisation de l'âge comme échelle

de temps. Idéalement, il faudrait aussi adapter le programme de Laaksonen *et al.* [80] afin de s'affranchir de la contrainte des intervalles de temps réguliers; il serait préférable de pouvoir définir les intervalles de temps en fonction de la distribution temporelle des événements de façon à éviter une mauvaise estimation du risque de base due à des intervalles trop larges ou à un nombre d'événements trop faible.

Dans notre exemple, nous avons considéré comme exposées les femmes qui avaient reçu des THM avant l'inclusion seulement, alors que l'exposition aux THM peut varier durant le suivi. D'autres travaux méthodologiques sont nécessaires pour prendre en compte les expositions dépendantes du temps. Seule l'approche globale de Spiegelman *et al.* [31] permet actuellement d'estimer le RA à une exposition dépendante du temps.

Enfin, nous avons ignoré les risques concurrents dus aux décès et aux autres types de cancer (6,46 % pour les autres cancers, 0,12 % pour les cancers sans date connue, 0,77 % pour les cancers du sein *in situ* et 3,85 % pour les décès, ce qui représente au total 11,2 % pour nos 94,2 % de censure observés), ce qui pourrait biaiser notre estimation du risque de cancer du sein attribuable à l'utilisation des THM [73]. Une extension permettant de prendre en compte les risques concurrents a été proposée pour la méthode paramétrique [73] et discutée pour la méthode semi-paramétrique [91] mais pas implémentée à notre connaissance (communication personnelle de Cynthia Crowson, 15 janvier 2016).

En conclusion, les estimateurs du RA ont des performances satisfaisantes pour les quatre méthodes considérées quand l'hypothèse des risques proportionnels est vérifiée. Les estimateurs du RA issus des approches semi-paramétrique et paramétrique ne sont pas robustes en revanche en cas de non respect de cette hypothèse. Le manque de précision constitue un problème pour les méthodes non paramétriques dans les cohortes où le nombre d'événements est faible. Dans la pratique, si l'hypothèse des risques proportionnels est vérifiée, il paraît préférable d'utiliser les approches semi-paramétrique ou paramétrique.

Chapitre 4

Estimation de la fraction préventive dans une cohorte

4.1 Problématique et objectifs

En pharmaco-épidémiologie, un des objectifs importants est d'estimer l'impact réel de l'exposition médicamenteuse sur les maladies. Lorsque cette exposition est associée à une réduction du risque d'avoir la maladie, la fraction préventive (FP) peut être utilisée pour estimer la proportion de cas de maladie qui pourraient être évités. Aucune définition ni méthode d'estimation de la FP n'a été proposée dans le contexte de l'analyse de survie.

L'objectif de ce travail est de proposer une définition de la FP dans le contexte de l'analyse de survie avec une application aux données de la cohorte Trois Cités (3C) sur l'association entre les traitements hypolipémiants et le risque d'accident vasculaire cérébral (AVC).

Pour illustrer notre propos, nous avons utilisé les données récemment publiées [92] de la cohorte 3C pour estimer la proportion de cas d'AVC qui pourraient être évités en utilisant les traitements hypolipémiants en considérant deux méthodes : une méthode semi-paramétrique basée sur le modèle de Cox à risques proportionnels COX [6, 8] et une méthode paramétrique basée sur le modèle à risque instantané de base constant par morceaux RCM [3].

4.2 Proposition d'estimateurs de la fraction préventive pour des données de survie

4.2.1 Définition de la fraction préventive en survie

Dans le contexte de l'analyse de survie, comme pour le RA, plusieurs interprétations des probabilités $\mathbf{P}(D)$ et $\mathbf{P}(D|\bar{E})$ peuvent être envisagées [3–8]. Ainsi, lorsque nous interprétons les probabilités de maladie comme des fonctions de répartition [3, 6–8], la FP peut être définie en utilisant la fonction de survie marginale $S(t)$ et la fonction de survie obtenue sous l'hypothèse que tous les sujets sont non exposés $S_0(t)$:

$$FP(t) = \frac{\mathbf{P}(T \leq t|Z=0) - \mathbf{P}(T \leq t)}{\mathbf{P}(T \leq t|Z=0)} = \frac{F(t|Z=0) - F(t)}{F(t|Z=0)} = \frac{S(t) - S_0(t)}{1 - S_0(t)}.$$

Alternativement, lorsque nous interprétons les probabilités de maladie comme des fonctions de risque instantané [4, 5], la FP instantanée peut être définie comme :

$$\psi(t) = \frac{\lambda(t|Z=0) - \lambda(t)}{\lambda(t|Z=0)} = 1 - \frac{\lambda(t)}{\lambda_0(t)}$$

où $\lambda(t)$ est le risque instantané au temps t et $\lambda_0(t) = \lambda(t|Z=0)$.

Dans cette partie, nous nous sommes basés sur la première définition de la FP dans le contexte de la survie. En effet, l'utilisation des risques instantanés pour définir le RA apparaît moins fréquente dans la littérature. Ainsi, comme dans le cas général, la FP est également fonction du RA :

$$FP(t) = 1 - \frac{1}{1 - A(t)} = -\frac{A(t)}{1 - A(t)} \tag{4.1}$$

où $A(t) = \dfrac{S_0(t) - S(t)}{1 - S(t)}$.

4.2.2 Méthodes d'estimation de la fraction préventive en survie

Plusieurs programmes sont proposés pour calculer le RA dans le contexte de l'analyse de survie [75, 80, 84]. Par conséquent, nous pouvons estimer la FP en utilisant la relation

avec le RA (équation 4.1) et les programmes existants :

$$\widehat{FP}(t) = -\frac{\hat{A}(t)}{1-\hat{A}(t)}.$$

Sa variance peut également être estimée par delta méthode [11] :

$$\widehat{\mathrm{Var}}\{\widehat{FP}(t)\} = \frac{\widehat{\mathrm{Var}}\{\hat{A}(t)\}}{\{1-\hat{A}(t)\}^4}.$$

Alternativement, l'utilisation d'une approche directe pour estimer la FP et sa variance peut être considérée. Comme pour le RA, un estimateur naturel consiste à remplacer $S(t)$ et $S_0(t)$ par leurs estimateurs respectifs $\hat{S}(t)$ et $\hat{S}_0(t)$ obtenus selon différentes approches :

$$\widehat{FP}(t) = \frac{\hat{S}(t)-\hat{S}_0(t)}{1-\hat{S}_0(t)}.$$

En nous basant sur les travaux de Chen *et al.*, nous pouvons calculer directement l'expression de la variance de $\widehat{FP}(t)$ pour des données de survie. En effet,

$$\begin{aligned}\sqrt{n}\{\widehat{FP}(t) - FP(t)\} &= \sqrt{n}\left\{\frac{\hat{S}(t)-\hat{S}_0(t)}{1-\hat{S}_0(t)} - \frac{S(t)-S_0(t)}{1-S_0(t)}\right\} \\ &= \frac{\sqrt{n}\{\hat{S}(t)-S(t)\}}{\{1-S_0(t)\}} - \sqrt{n}\{\hat{S}_0(t)-S_0(t)\}\frac{1-\hat{S}(t)}{\{1-S_0(t)\}\{1-\hat{S}_0(t)\}}.\end{aligned}$$

Notons P_n la mesure empirique sur les données observées et P la distribution théorique. Chen *et al.* ont montré que $\sqrt{n}\{\hat{S}(t)-S(t)\}$ et $\sqrt{n}\{\hat{S}_0(t)-S_0(t)\}$ sont asymptotiquement équivalents à $\sqrt{n}\,(P_n-P)\,\eta_2(t)$ et $\sqrt{n}\,(P_n-P)\,\eta_1(t)$ respectivement où les fonctions $\eta_1(t)$ et $\eta_2(t)$ dépendent des méthodes d'estimation de $S_0(t)$ et $S(t)$ respectivement. Les expressions de $\eta_1(t)$ et $\eta_2(t)$ correspondent à celles citées dans le chapitre 2 dans le cas où les fonctions de survie sont estimées en utilisant une approche semi-paramétrique ou non paramétrique. Ainsi, sous certaines conditions de régularité, $\sqrt{n}\{\widehat{FP}(t) - FP(t)\}$ est asymptotiquement équivalent à

$$\sqrt{n}\,(P_n-P)\,\frac{1}{1-S_0(t)}\left\{\eta_2(t) - \frac{1-S(t)}{1-S_0(t)}\eta_1(t)\right\}.$$

Par conséquent, $\sqrt{n}\{\widehat{FP}(t) - FP(t)\}$ converge faiblement vers un processus Gaussien de moyenne nulle et de matrice de variance-covariance $\mathbf{E}\{\Omega(t)\Omega^T(s)\}$ entre les temps t et s avec :

$$\Omega(t) = \frac{1}{1 - S_0(t)} \left\{ \eta_2(t) - \frac{1 - S(t)}{1 - S_0(t)} \eta_1(t) \right\}.$$

Un estimateur consistant de la fonction de covariance est :

$$\hat{\sigma}_{FP}(s,t) = \frac{1}{n} \sum_{i=1}^{n} \hat{\Omega}_i(t) \hat{\Omega}_i^T(s)$$

où

$$\hat{\Omega}_i(t) = \frac{1}{1 - \hat{S}_0(t)} \left\{ \hat{\eta}_{2i}(t) - \frac{1 - \hat{S}(t)}{1 - \hat{S}_0(t)} \hat{\eta}_{1i}(t) \right\}$$

et $\hat{\eta}_{1i}(t)$ et $\hat{\eta}_{2i}(t)$ sont les estimations de $\eta_1(t)$ et $\eta_2(t)$ respectivement pour le i-ème sujet au temps t.

Pour construire les IC de la FP, nous pouvons utiliser la transformation logarithme, $\ln\{1 - FP(t)\}$, comme cela a été proposé pour le RA [3, 6].

Les mêmes approches que celles considérées pour l'estimation du RA peuvent être envisagées pour la FP dans le contexte de la survie. Pour notre application, nous avons considéré deux approches pour estimer la FP définie à l'aide des fonctions de survie : une approche semi-paramétrique basée sur le modèle de Cox à risques proportionnels [6, 8] et une approche paramétrique basée sur un modèle à risque de base constant par morceaux [3].

4.3 Application aux données de la cohorte 3C

4.3.1 Présentation de la cohorte 3C

L'étude des Trois Cités est une étude de cohorte prospective dont l'objectif est d'étudier les facteurs de risque qui favorisent la survenue de certaines maladies comme les AVC ou la maladie d'Alzheimer mais aussi d'étudier le rôle précis des facteurs de risque vasculaires (hypertension artérielle) dans la perte des capacités intellectuelles [93, 94].

La cohorte a inclus 9 294 personnes âgées de 65 ans ou plus, tirées au sort sur les listes électorales de trois villes françaises : Bordeaux, Dijon et Montpellier entre mars 1999 et mars 2001. A leur inclusion comme à chacun des examens de suivi (environ tous les deux

ans), les participants de l'étude ont subi un entretien en face à face au cours duquel les données (caractéristiques démographiques, niveaux d'études et socioéconomique, consommation de tabac et d'alcool, habitudes alimentaires, antécédents médicaux notamment vasculaires et consommation de médicaments) ont été recueillies.

À chaque examen de suivi, les participants ou les représentants des participants décédés ont été systématiquement interrogés sur la survenue éventuelle d'événements médicaux graves et d'hospitalisations depuis le dernier contact. Pour les participants ayant rapporté une possible maladie cardiaque (coronarienne) ou un AVC, toutes les informations cliniques disponibles ont été rassemblées et un comité d'experts a passé en revue ces dernières afin de classer chaque événement en utilisant le code CIM-10 (Classification internationale des maladies, 10-ème révision).

Au domicile des participants, les enquêteurs ont rassemblé des informations sur tous les médicaments utilisés pendant le mois précédent. Il a été demandé aux participants de montrer leurs prescriptions et les boîtes de médicament. Les noms des médicaments ont été codés selon la classification anatomique, thérapeutique et chimique (ATC) de l'Organisation Mondiale de la Santé (OMS). Les hypolipémiants (ATC C10A) incluent les statines (ATC C10AA), les fibrates (ATC C10AB) et d'autres médicaments non considérés dans cette étude. L'utilisation des médicaments antithrombotiques (ATC B01A) a aussi été enregistrée.

En pharmaco-épidémiologie, la détermination précise de l'exposition médicamenteuse est fondamentale, en particulier dans les études dont l'objectif final vise à évaluer l'association entre une exposition médicamenteuse et un événement d'intérêt [95]. La validité des données sur l'exposition aux hypolipémiants a été évaluée dans l'étude 3C. Ainsi, Noize *et al.* [96, 97] ont montré un très bon accord entre l'utilisation des médicaments hypolipémiants rapportée par les participants et les remboursements enregistrés dans la base de données de l'assurance maladie.

4.3.2 Population et méthodes

Nous avons reproduit les critères de sélection retenus pour définir la population d'une étude récemment publiée et portant sur l'association entre l'utilisation des hypolipémiants et les maladies cardiovasculaires chez les participants de la cohorte 3C [92]. Cette étude rapportait une association statistiquement significative entre l'utilisation de statines et le risque d'AVC [92].

Sur les 9 294 participants, 1 439 n'étaient pas éligibles à cette étude car ils avaient déclaré des antécédents de maladie cardiovasculaire à l'inclusion : 1 017 avaient déclaré une maladie coronarienne, 330 un AVC et 92 les deux. Les 113 participants utilisant des médicaments hypolipémiants autres que statines et fibrates ont également été exclus de l'étude. Sur les 7 742 participants restants, 258 étaient perdus de vue. La population d'étude comprenait au final 7 484 participants.

Comme dans l'étude originale [92], nous avons considéré comme exposés les participants qui ont déclaré avoir utilisé des statines ou des fibrates à l'inclusion et comme non exposés ceux qui n'ont utilisé aucun des deux produits à l'inclusion. Parmi les participants inclus, 2 048 (27,4 %) ont utilisé des hypolipémiants (statines ou fibrates) à l'inclusion dans la cohorte.

Dans ce travail, nous nous sommes intéressés à l'incidence de l'AVC comme premier événement cardiovasculaire. Ainsi, pour un suivi maximal de 12,9 années, 292 (3,9 %) cas d'AVC ont été diagnostiqués. Nous avons également considéré 1 791 événements concurrents (maladie coronarienne ou décès par autres causes).

Nous avons considéré deux approches pour estimer la FP : une approche semi-paramétrique basée sur le modèle de Cox à risques proportionnels [6, 8] et une approche paramétrique basée sur le modèle à risque de base constant par morceaux [3]. En outre, parce que l'utilisation des hypolipémiants et les autres facteurs de risque de l'AVC pourraient aussi être liés à la maladie coronarienne et aux décès par autres causes, nous avons considéré une extension de l'approche paramétrique basée sur les risques spécifiques par cause pour prendre en compte la maladie coronarienne comme premier événement vas-

culaire et les décès par autres causes comme des événements concurrents (cf. figure 4.1) dans le modèle paramétrique à risque constant par morceaux [73].

Pour rester cohérent avec l'analyse publiée sur les données de la cohorte 3C [92], nous considérons dans un premier temps un modèle simple (modèle 1) ajusté sur le sexe, le centre d'étude (Bordeaux, Dijon, Montpellier) et l'âge (en 10 classes, déciles) avec le temps de suivi comme échelle de temps. Ensuite, nous considérons un modèle plus complexe (modèle 2) ajusté, en plus des covariables du modèle 1, sur des facteurs de confusion potentiels : diabète (oui, non), indice de masse corporelle (IMC < 25, 25-29, ≥ 30 kg/m^2), statut tabagique (jamais fumé, ex-fumeur, fumeur actuel), consommation d'alcool (non buveur, ex-buveur, buveur actuel), hypertension artérielle (oui, non), troubles du rythme cardiaque (oui, non), utilisation de médicaments antithrombotiques (oui, non), concentration des triglycérides (tertiles) et rapport entre cholestérol LDL (pour *low density lipoprotein*) et HDL (pour *high density lipoprotein*) (tertiles).

Nous avons évalué en premier lieu l'association entre l'utilisation des hypolipémiants et le risque d'AVC. Pour cela, nous avons utilisé la méthode de Kaplan-Meier [47] pour estimer les fonctions de survie sans AVC chez les exposés et chez les non exposés aux hypolipémiants à l'inclusion et le test du log rank pour comparer les deux courbes de survie. Nous avons estimé le HR de développer un AVC chez les exposés aux hypolipémiants par rapport aux non exposés et son IC à 95 % à l'aide du modèle de Cox à risques proportionnels et du modèle paramétrique exponentiel par morceaux avec le temps de suivi comme échelle de temps. Pour le modèle de Cox, nous avons également utilisé l'âge comme échelle de temps pour vérifier la compatibilité des estimations du HR avec celles obtenues avec le temps de suivi comme échelle de temps après ajustement sur l'âge en déciles. Pour les deux modèles paramétrique et semi-paramétrique, une fonction log-linéaire a été supposée pour l'association entre les facteurs de risque et la fonction de risque et nous avons testé l'hypothèse des risques proportionnels en examinant l'interaction entre l'exposition et le temps de suivi divisé en quatre intervalles de trois ans.

Dans un deuxième temps, nous avons estimé la FP d'AVC associée à l'utilisation des hypolipémiants en fonction du temps de suivi. Nous avons utilisé l'estimation du RA et de

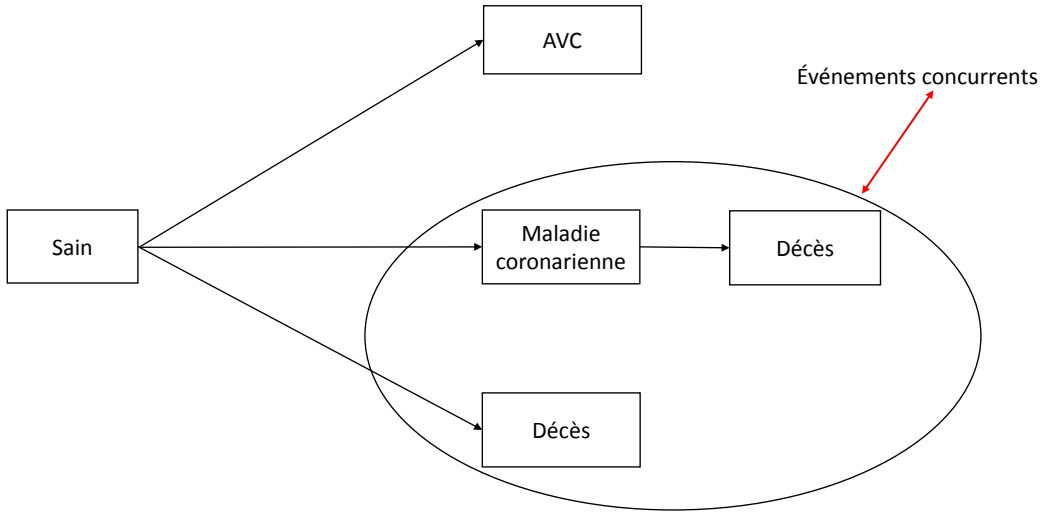


FIGURE 4.1: Événements concurrents pour l'analyse de l'association entre utilisation des statines et risque d'accident vasculaire cérébral (AVC) comme premier événement cardio-vasculaire

sa variance pour en déduire la FP selon l'équation 4.1 et avons appliqué la delta méthode pour estimer la variance de la FP [11]. Les IC de la FP ont été construits en utilisant la transformation logarithmique.

Les résultats sont présentés en quatre temps également répartis (après 3, 6, 9 et 12 ans de suivi). Pour l'approche semi-paramétrique, les estimateurs sont obtenus en des temps réellement observés dans la cohorte. Nous avons donc retenu les valeurs inférieures les plus proches des temps considérés.

Tous les calculs ont été effectués avec les logiciels R (version 3.2.2, R Core Team, Vienna, Austria) et SAS (version 9.3, SAS Institute Inc., Cary, NC, USA). Pour l'approche semi-paramétrique, nous avons utilisé le package `paf` développé par Chen [75]. Les résultats ne sont pas présentés pour le package `AF` de Dahlqwist *et al.* [84] parce que les estimations ponctuelles étaient pratiquement identiques et les variances proches de celles obtenues avec le package `paf` de Chen. Pour l'approche paramétrique, nous avons utilisé un ensemble de macros développées par Laaksonen *et al.* [80].

4.3.3 Résultats

Parmi les 7 484 participants, l'âge moyen était de 73,9 ans, 63,0 % étaient des femmes et 2 048 (27,4 %) avaient déclaré avoir utilisé des hypolipémiants (13,5 % des statines et 13,8 % des fibrates) à l'inclusion. Cinq participants ont utilisé les deux. En comparaison avec les non exposés, les utilisateurs de statines ou fibrates étaient plus jeunes, davantage susceptibles d'être des femmes et avaient un niveau d'éducation plus bas [92]. Les utilisateurs de médicaments hypolipémiants avaient aussi une pression artérielle et un IMC plus élevés. Ils avaient plus souvent de l'hypertension artérielle, du diabète et des troubles du rythme cardiaque. Ils utilisaient plus souvent des antihypertenseurs et des antithrombotiques [92]. En comparaison avec les utilisateurs de statines, les utilisateurs de fibrates étaient plus âgés, avaient une consommation d'alcool et une pression artérielle diastolique plus basses et prenaient moins de traitements antithrombotiques [92]. Les taux de cholestérol et de triglycérides étaient plus bas chez les utilisateurs de fibrates que chez les utilisateurs de statines [92].

La figure 4.2 montre que la survie à l'AVC est statistiquement significativement meilleure chez les exposés aux médicaments hypolipémiants que chez les personnes non exposées (test du log rank $p = 0{,}002$).

En utilisant le temps de suivi comme échelle de temps, avec le modèle 1, les HR associés à l'exposition aux hypolipémiants estimés selon un modèle de Cox à risques proportionnels (HR, 0,692; IC 95 %, 0,518 à 0,924) et un modèle paramétrique exponentiel par morceaux (HR, 0,698; IC 95 %, 0,522 à 0,932) sont significativement inférieurs à 1 et proches l'un de l'autre (Tableau 4.1). L'ajustement sur d'autres facteurs de confusion potentiels (modèle 2) tend à renforcer les associations entre hypolipémiants et risque d'AVC, les HR s'éloignant de 1 (HR, 0,646; IC 95 %, 0,478 à 0,873 et HR, 0,653; IC 95 %, 0,483 à 0,882, respectivement (Tableau 4.1)). L'hypothèse des risques proportionnels n'est pas rejetée pour les deux approches Cox et paramétrique et les deux modèles 1 et 2 (test de rapport de vraisemblance, $p > 0{,}5$ pour tous les modèles). Nous obtenons des estimations proches en utilisant l'âge comme échelle de temps pour le modèle de Cox à

risques proportionnels (HR, 0,710; IC 95 %, 0,531 à 0,948 et HR 0,659; IC 95 %, 0,488 à 0,891 pour les modèles 1 et 2 respectivement).

Lorsque nous considérons un modèle semi-paramétrique simple (modèle 1), l'estimation de la proportion de cas d'AVC évités par l'utilisation de statines ou fibrates est décroissante en fonction du temps, passant de 7,86 % (IC 95 %, 2,57 % à 12,87 %) après 3 ans de suivi à 7,61 % (IC 95 %, 2,41 % à 12,53 %) après 12 ans de suivi (Tableau 4.1). Avec le modèle 2, l'estimation de la FP augmente, en cohérence avec les estimations du HR plus éloignées de 1. Elle reste décroissante en fonction du temps de suivi passant de 9,41 % (IC 95 %, 4,03 % à 14,50 %) après 3 ans de suivi à 9,05 % (IC 95 %, 3,78 % à 14,03 %) après 12 ans de suivi (Tableau 4.1). Les estimations obtenues par la méthode paramétrique ne diffèrent pas notablement de celles obtenues par la méthode semi-paramétrique (Tableau 4.1).

Enfin, lorsque nous considérons la maladie cardiaque et le décès par autres causes comme événements concurrents avec l'approche paramétrique, les estimations de la FP sont très proches de celles obtenues en ignorant les événements concurrents pour les deux modèles d'ajustement (Tableau 4.1).

4.4 Discussion

4.4.1 Synthèse des résultats

Suite à une récente publication d'une réduction d'un tiers du risque d'AVC statistiquement significative chez les exposés aux hypolipémiants comparés aux non utilisateurs d'hypolipémiants à l'inclusion dans la cohorte 3C composée de personnes âgées [92], nous avons estimé que jusqu'à 9 % des cas d'AVC pourraient être évités en utilisant les statines ou les fibrates. Cette proportion est loin d'être négligeable en raison de l'incidence et de la gravité des AVC chez les personnes âgées.

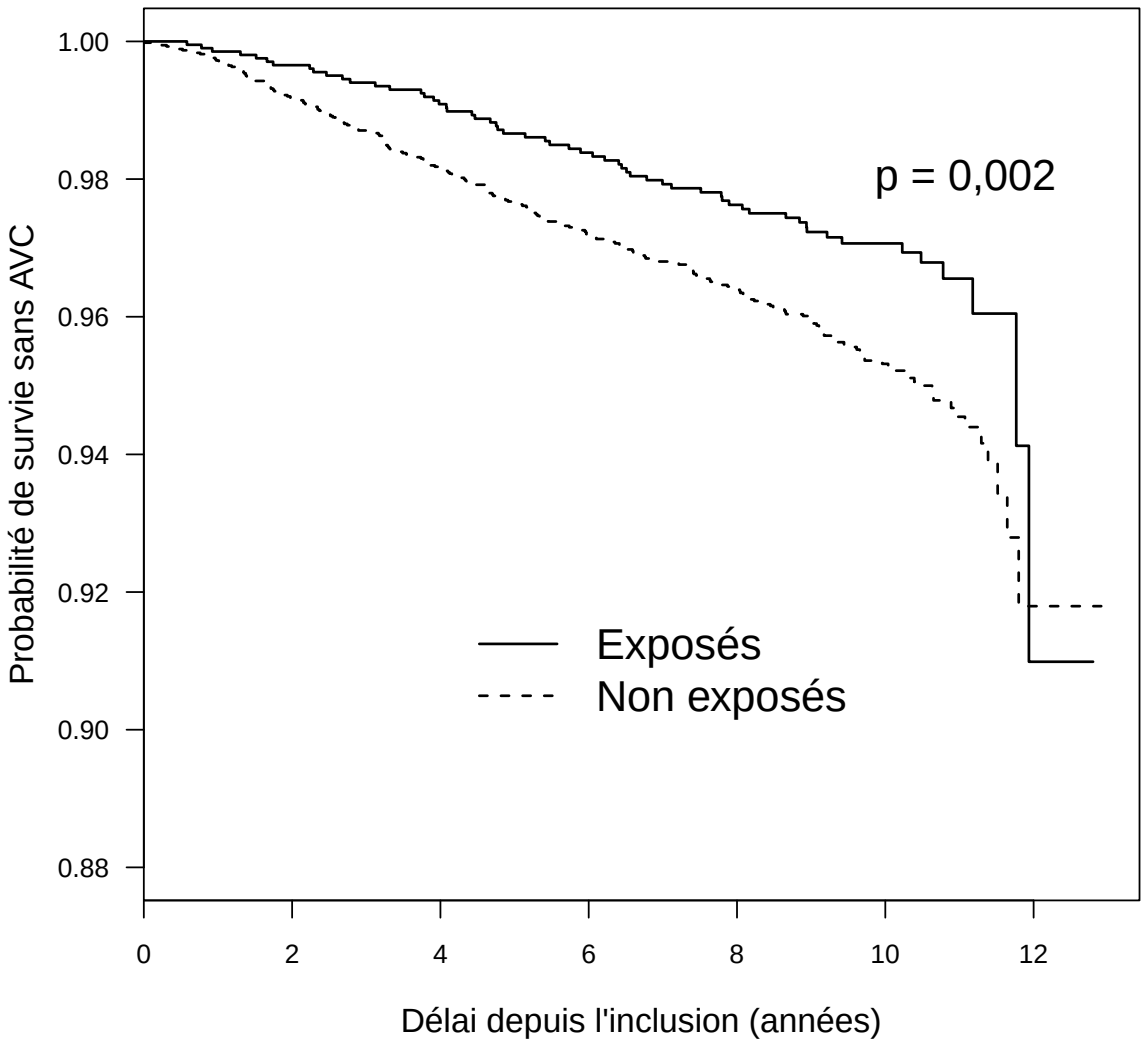


FIGURE 4.2: Estimations de Kaplan-Meier des fonctions de survie chez les exposés et les non exposés aux hypolipémiants à l'inclusion, Cohorte 3C, 1999-2011

Estimation	Modèle 1			Modèle 2		
méthode	HR (IC 95 %)	FP	IC 95 %	HR (IC 95 %)	FP	IC 95 %
COX	0.692 (0.518 – 0.924)	0.0786	0.0257 – 0.1287	0.646 (0.478 – 0.873)	0.0941	0.0403 – 0.1450
		0.0782	0.0254 – 0.1281		0.0935	0.0398 – 0.1441
		0.0777	0.0251 – 0.1275		0.0928	0.0394 – 0.1433
		0.0761	0.0241 – 0.1253		0.0905	0.0378 – 0.1403
RCM	0.698 (0.522 – 0.932)	0.0772	0.0238 – 0.1276	0.653 (0.483 – 0.882)	0.0922	0.0372 – 0.1441
		0.0767	0.0235 – 0.1270		0.0916	0.0368 – 0.1432
		0.0763	0.0233 – 0.1264		0.0910	0.0364 – 0.1424
		0.0755	0.0229 – 0.1253		0.0899	0.0358 – 0.1410
RCM*	0.698 (0.522 – 0.932)	0.0773	0.0237 – 0.1279	0.653 (0.483 – 0.882)	0.0922	0.0370 – 0.1443
		0.0769	0.0232 – 0.1277		0.0916	0.0362 – 0.1438
		0.0767	0.0227 – 0.1277		0.0910	0.0353 – 0.1434
		0.0764	0.0219 – 0.1278		0.0900	0.0338 – 0.1429

Tableau 4.1: Estimation de la fraction préventive d'AVC associée à l'utilisation d'hypolipémiants à l'inclusion, après 3, 6, 9 et 12 ans de suivi, Cohorte 3C, 1999-2011
RCM* correspond à l'approche paramétrique en considérant la maladie cardiaque coronarienne et le décès pour autres causes comme événements concurrents

4.4.2 Comparaison aux travaux antérieurs

À notre connaissance, notre étude est la première à proposer une estimation de la FP dans le cadre des études de cohorte et dans le contexte de l'analyse de survie. Notre étude complète l'étude originale dans laquelle les HR d'AVC chez les utilisateurs d'hypolipémiants ont été rapportés [92], en apportant des estimations de la FP qui tiennent compte des HR estimés, ainsi que de la prévalence à 27,4 % de l'exposition aux hypolipémiants à l'inclusion. Jusqu'ici, les mesures de la FP considérées pour les études de cohorte ont supposé des temps de suivi fixes pour tous les sujets, ignorant ainsi la censure due aux perdus de vue et les risques concurrents [45]. La prise en compte du temps auquel est évalué le risque, dans le contexte de l'analyse de survie avec des risques concurrents, est importante dans un cadre longitudinal parce que la prévalence des facteurs de risques change au cours du suivi.

Comme pour les fonctions du RA dans d'autres exemples avec des expositions délétères (RR > 1), la FP est également une fonction décroissante avec le temps de suivi en présence d'une exposition protectrice (RR constant < 1) en cas de risques proportionnels. Dans notre exemple d'application, nous avons observé de faibles variations des estimations de la FP dans le temps que ce soit avec l'approche semi-paramétrique ou paramétrique. Davantage de variation pourrait être attendue dans le cas d'événements plus fréquents comme dans les cohortes cliniques où des patients atteints d'une maladie sont suivis [3, 5, 6, 73].

Dans ce travail, nous avons utilisé deux méthodes pour estimer la FP en fonction du temps lorsque les probabilités de maladie sont interprétées comme des fonctions de répartition, ce qui nous a permis de définir la FP à l'aide des fonctions de survie. La première approche est une méthode semi-paramétrique basée sur le modèle de Cox à risques proportionnels, la seconde est une méthode paramétrique basée sur un modèle à risque de base constant par morceaux (modèle exponentiel par morceaux). Les méthodes non paramétriques basées sur l'estimateur de Kaplan-Meier de la fonction de survie qui ont été proposées pour estimer le RA [6] pourraient être utilisées pour estimer la FP. Cependant, de telles approches seraient limitées par la nature (catégorielle seulement) et le nombre de covariables qui pourraient être prises en considération. Par conséquent, ces méthodes ne conviennent pas pour une application à l'étude 3C ou des études similaires avec de nombreux facteurs de confusion potentiels. Une définition alternative de la FP en fonction du temps pourrait être obtenue lorsque les probabilités de la maladie sont interprétées comme des fonctions de risque instantané comme pour le RA [4, 5] mais cette définition diffère sur le plan conceptuel de la définition générale et a reçu moins d'attention dans la littérature.

Dans cette étude, nous avons estimé la variance de la FP en appliquant la delta méthode à l'estimateur du RA et sa variance associée obtenus à partir de programmes SAS et R disponibles. Alternativement, nous pourrions également adapter ces programmes pour calculer la FP directement en utilisant les fonctions de survie $S_0(.)$ et $S(.)$, ainsi que la variance associée en adaptant les approches proposées pour le RA. Pour l'approche

semi-paramétrique, Chen *et al.* [6] se sont basés sur des développements asymptotiques alors que Sjölander *et al.* [8] ont combiné une approche sandwich et la delta méthode. Pour l'approche paramétrique, Laaksonen *et al.* [3] ont obtenu une estimation de la variance en utilisant la delta méthode. L'utilisation des approches directes a donné des résultats pour l'estimation de la FP et son IC très proches de ceux obtenus en utilisant l'équation 4.1 et la delta méthode.

Plusieurs auteurs [3, 6] ont noté que l'approximation normale de la distribution du RA estimé peut ne pas être exacte en cas d'échantillon de petite taille et recommandent l'utilisation d'une transformation logarithmique, $\ln\{1-A(t)\}$. Il pourrait en être de même pour les estimations de la FP. Pour cette raison, nous avons également considéré une transformation logarithmique, $\ln\{1-A(t)\} = -\ln\{1-FP(t)\}$, qui est conforme aux recommandations données pour le RA [3, 6].

Comme observé dans nos résultats, lorsque nous avons pris en compte la maladie coronarienne et le décès par autres causes comme des risques concurrents pour l'approche paramétrique, les estimations de la FP étaient très proches de celles obtenues en ignorant les risques concurrents. Nous n'avons observé aucune association statistiquement significative entre l'utilisation des hypolipémiants et les événements concurrents dans notre exemple (HR, 0,960; IC 95 %, de 0,864 à 1,073 et HR, 0,932 IC 95 %, de 0,833 à 1,042 pour les modèles 1 et 2 respectivement avec l'approche paramétrique). L'événement d'intérêt dans cette étude était l'AVC comme premier événement cardiovasculaire. Ignorer les risques concurrents dus à la maladie coronarienne et au décès par autres causes implique que les estimations de la FP obtenues ne sont valables que sous l'hypothèse que personne ne meure ou ne développe une maladie coronarienne comme premier événement cardiovasculaire au cours du suivi [73]. Le biais causé en censurant la durée d'observation à la survenue d'une maladie coronarienne et d'un décès par autres causes était très faible dans notre exemple du fait de l'absence d'association entre les hypolipémiants et ces événements concurrents. Il pourrait devenir plus important si l'association était plus forte et le suivi plus long [73].

4.4.3 Limites de l'étude

Tout d'abord, dans leur analyse, Alpérovitch *et al.* [92] ont utilisé l'âge comme échelle de temps dans les modèles de régression de Cox pour estimer les HR à partir des données de la cohorte 3C. Ils ont trouvé des HR de 0,71 (IC 95 %, 0,53 à 0,95) et 0,66 (IC 95 %, 0,49 à 0,90) pour les modèles simple et complet, respectivement. Les logiciels disponibles actuellement pour l'estimation du RA comme fonction du temps ne permettent pas la troncature à gauche résultant de l'utilisation de l'âge comme échelle de temps. Pour cette raison, nous avons utilisé dans notre étude le temps de suivi comme échelle de temps et avons ajusté nos modèles sur l'âge d'inclusion. Nous avons trouvé des estimations de HR proches de celles de l'étude originale basée sur les mêmes données (0,69 et 0,65 avec les modèles de Cox 1 et 2 respectivement).

Deuxièmement, nous n'avons pas considéré les risques concurrents pour l'approche semi-paramétrique. En effet, les auteurs qui ont proposé cette approche n'ont pas abordé cette question [6,8]. Il serait néanmoins intéressant d'étendre l'approche semi-paramétrique afin de prendre en compte les risques concurrents dans l'estimation de la FP comme du RA [91].

Enfin, nous ne pouvons pas exclure que notre estimation de la FP puise être entachée des biais inhérents aux études pharmaco-épidémiologiques, en particulier le biais d'indication. Celui-ci se produit lorsque le médicament considéré est prescrit plus souvent chez les sujets à risque élevé de l'événement d'intérêt que chez les sujets à faible risque. De plus, un biais de sélection est possible, les participants à l'étude 3C différant quelque peu de la population générale française. Ils avaient en effet plus souvent suivi un enseignement supérieur, une situation économique favorable et un meilleur fonctionnement cognitif. Globalement, ils avaient un mode de vie sain, en particulier de bonnes habitudes alimentaires, ce qui pourrait contribuer à une réduction du risque vasculaire [98]. Cependant, il est peu probable que ces spécificités des participants de la cohorte 3C aient un effet majeur sur l'estimation du risque d'AVC associé à l'utilisation d'hypolipémiants. L'ensemble de ces biais a été écarté comme explication première du risque d'AVC diminué chez les utilisa-

teurs d'hypolipémiants comparés aux non utilisateurs dans l'étude originale [92] et l'étude détaillée de ces biais dépasse le cadre de notre étude.

En conclusion, notre étude a montré que la FP pouvait être estimée pour évaluer l'impact des médicaments bénéfiques dans les études de cohorte tout en tenant compte des facteurs de confusion potentiels et des risques concurrents. L'utilisation des statines ou fibrates était associée à environ 9 % de cas d'AVC évités chez les personnes âgées.

Chapitre 5

Conclusion et perspectives

5.1 Synthèse générale des travaux réalisés

L'objectif de cette thèse était de proposer une définition et des méthodes d'estimation du RA et de la FP dans les études de cohorte.

Dans notre premier travail, nous avons réalisé une étude de simulation afin de comparer différentes méthodes d'estimation du RA dans le contexte de l'analyse de survie. Dans le cas où l'hypothèse des risques proportionnels est vérifiée, cette étude a montré que les estimateurs du RA sont sans biais pour les quatre méthodes considérées, à savoir les deux approches non paramétriques KM et KMP basées sur l'estimateur de Kaplan-Meier, l'approche semi-paramétrique COX basée sur le modèle de Cox à risques proportionnels et l'approche paramétrique RCM basée sur le modèle à risque de base constant par morceaux. Les approches semi-paramétrique et paramétrique sont plus précises en début de suivi contrairement aux approches non paramétriques. Dans le cas où l'hypothèse des risques proportionnels n'est pas vérifiée, les estimateurs issus des approches semi-paramétrique et paramétrique sont biaisés. L'application aux données de la cohorte E3N a montré qu'environ 9 % des cas de cancer du sein invasif seraient attribuables à une utilisation des traitements hormonaux de la ménopause à l'inclusion.

Dans notre second travail, nous avons proposé une définition de la FP dans le cadre de l'analyse de survie. Nous avons aussi adapté les programmes existants pour le RA à

la FP afin d'estimer directement la variance et les IC. Nous avons aussi montré qu'il est possible d'utiliser les programmes existants pour estimer la FP en utilisant la relation avec le RA et sa variance par delta méthode. Nous avons ainsi montré que la FP peut être estimée pour évaluer l'impact des médicaments bénéfiques dans les études de cohorte tout en tenant compte des facteurs de confusion potentiels et des risques concurrents. L'application aux données de la cohorte 3C a montré que l'utilisation des hypolipémiants pourrait permettre d'éviter environ 9 % de cas d'AVC chez les personnes âgées.

5.2 Logiciels

Dans la dernière décennie, plusieurs programmes permettant de calculer le RA ont été proposés pour les études de cohorte et dans le contexte de l'analyse de survie. Le calcul de la FP peut être déduit du RA mais les programmes existants pourraient être développés pour être plus largement applicables.

L'utilisation de l'âge comme échelle de temps est une pratique courante en épidémiologie pour l'analyse des facteurs de risque de maladies chroniques [79, 90]. Par conséquent, une extension des programmes existants s'avère nécessaire pour permettre une plus large utilisation du RA et de la FP.

Laaksonen *et al.* [80] ont proposé un ensemble de macros permettant d'estimer le RA avec un modèle paramétrique par morceaux avec des pas de temps réguliers. Ces macros ne sont pas adaptées à des applications où le risque de base décroît très vite. En effet, dans ce cas, un découpage trop large en un nombre limité d'intervalles de temps égaux rend insuffisamment compte des variations au début du suivi tandis qu'un découpage trop fin en un nombre plus important d'intervalles de temps égaux est susceptible d'engendrer des problèmes d'estimation à cause de l'absence d'événements observés dans les intervalles en fin de suivi. Une extension de ces macros permettant de définir des pas de temps irréguliers en fonction du temps de suivi s'avère donc nécessaire.

Enfin, Laaksonen *et al.* [73] ont également proposé un programme permettant de prendre en compte les événements concurrents dans l'estimation du RA pour l'approche

paramétrique. Des développements analogues pour l'approche semi-paramétrique basée sur le modèle de Cox à risques proportionnels seraient nécessaires pour une plus large utilisation de ces mesures d'impact en épidémiologie quand on s'intéresse à l'incidence d'une maladie et non au décès.

5.3 Extensions en vue d'une utilisation en pharmaco-épidémiologie

Dans les deux applications considérées pour cette thèse, l'exposition d'intérêt était la prise d'un médicament mesurée à l'inclusion et représentée par une variable binaire. Cette exposition, supposée fixe à l'inclusion dans notre travail, peut changer au cours du suivi. En effet, l'exposition en pharmaco-épidémiologie est souvent de nature dynamique : la prise d'un médicament n'est pas nécessairement constante ni ponctuelle dans le temps, tandis que la relation entre cette exposition et la réponse peut être gouvernée par une fonction de risque complexe avec effet variable dans le temps. L'estimation du RA et de la FP en pharmaco-épidémiologie pose ainsi des problèmes spécifiques pour lesquels des travaux complémentaires à ce travail de thèse seront nécessaires [99].

L'approche de Spiegelman *et al.* [31] pourrait être envisagée dans ce cas, car elle permet de prendre en compte dans le modèle une exposition dépendante du temps. Cependant, cette approche donne une estimation du RA global à l'échelle de la cohorte. Des méthodes d'estimation prenant en compte la nature dynamique de l'exposition en pharmaco-épidémiologie s'avèrent donc nécessaires afin de mieux estimer l'impact réel de l'exposition médicamenteuse au cours du suivi. Le besoin d'étendre les approches proposées actuellement à une exposition dépendante du temps a également été souligné par d'autres auteurs [5–8, 91].

Par ailleurs, le contexte spécifique des maladies infectieuses soulève des difficultés supplémentaires liées au défaut d'indépendance entre sujets. En effet, un sujet exposé à un médicament anti-infectieux peut devenir un cas (c'est-à-dire porteur du pathogène

transmissible) du fait de son exposition et, par transmission du pathogène à un autre sujet, faire que celui-ci devienne à son tour un cas. La survenue de l'événement chez un sujet constitue donc une exposition supplémentaire pour les autres sujets encore indemnes. Plus le nombre de sujets déjà porteurs est important (on parle de forte pression de colonisation), plus le risque de transmission aux sujets encore non porteurs devient grand. Cette non indépendance entre individus pose des difficultés pour l'estimation du risque d'acquisition de pathogène attribuable à l'exposition aux anti-infectieux. En effet, les modèles de régression classiques permettant la prise en compte d'expositions multiples et d'éventuels effets modificateurs dans le cas de maladies non transmissibles [2] ne sont pas adaptés aux maladies transmissibles. Il est primordial néanmoins de prendre en compte les cas secondaires (acquisition de résistance par transmission du pathogène) pour éviter une estimation biaisée du risque d'acquisition de résistance attribuable aux antibiotiques [100]. Des développements méthodologiques s'avèrent nécessaires pour prendre en compte les cas secondaires dans l'estimation du risque attribuable infectieux.

Enfin, alors que des méthodes d'estimation du risque attribuable existent pour les principaux schémas d'étude épidémiologique (transversal, cas-témoins, cohorte et cas-cohorte) [2], il est important de noter que d'autres schémas sont fréquemment utilisés en pharmaco-épidémiologie en général comme en épidémiologie de la résistance bactérienne en particulier. Ces schémas reposent sur des séries de cas sans témoins [101] ou bien comparent deux types de cas (porteurs de bactéries sensibles ou résistantes) avec des témoins [102]. Le schéma par série de cas suscite un intérêt grandissant en pharmaco-épidémiologie car il a l'avantage de permettre l'estimation d'une incidence relative proche du RR [103] et de permettre un contrôle naturel du biais de confusion lié aux caractéristiques individuelles fixes dans le temps [104–107]. Cette approche consiste à ne considérer que les cas pour la mesure de l'association entre une exposition intermittente, telle que rencontrée en vaccinologie, et des événements à survenue aiguë, rares, potentiellement récurrents [107]. Cette méthodologie s'applique quand l'événement ne peut plausiblement apparaître que dans un court délai après l'exposition. Ainsi, on peut comparer le risque pour chaque individu de développer l'événement pendant la période à risque, pendant laquelle l'exposition

est supposée pouvoir entraîner la survenue de l'événement, à son risque de développer l'événement pendant la période témoin, pendant laquelle l'exposition ne peut pas entraîner la survenue de cet événement. Un travail a été réalisé par Adrien Allorant pour son mémoire de Master 2 Recherche en 2016 proposant une définition du RA pour les séries de cas et sa validation [108].

5.4 Conclusion générale

Ce travail de thèse s'inscrit dans un programme de recherche plus vaste sur l'estimation du RA en pharmaco-épidémiologie. Cette thèse vise à encourager l'utilisation du RA et de la FP pour estimer l'impact réel de l'exposition médicamenteuse sur la survenue de maladie. Cette thèse a permis de comparer différentes méthodes d'estimation du RA dans le contexte de l'analyse de survie. Elle a aussi permis de proposer une définition et des méthodes d'estimation de la FP dans le contexte de l'analyse de survie.

Bibliographie

[1] ML Levin. The occurrence of lung cancer in man. *Acta Unio Internationalis Contra Cancrum*, 9(3):531–41, 1953.

[2] J Benichou. A review of adjusted estimators of attributable risk. *Statistical Methods in Medical Research*, 10(3):195–216, 2001.

[3] MA Laaksonen, P Knekt, T Härkänen, E Virtala, and H Oja. Estimation of the population attributable fraction for mortality in a cohort study using a piecewise constant hazards model. *American Journal of Epidemiology*, 171(7):837–47, 2010.

[4] SO Samuelsen and GE Eide. Attributable fractions with survival data. *Statistics in Medicine*, 27(9):1447–67, 2008.

[5] YQ Chen, C Hu, and Y Wang. Attributable risk function in the proportional hazards model for censored time-to-event. *Biostatistics*, 7(4):515–29, 2006.

[6] L Chen, DY Lin, and D Zeng. Attributable fraction functions for censored event times. *Biometrika*, 97(3):713–26, 2010.

[7] C Cox, H Chu, and A Muñoz. Survival attributable to an exposure. *Statistics in Medicine*, 28(26):3276–93, 2009.

[8] A Sjölander and S Vansteelandt. Doubly robust estimation of attributable fractions in survival analysis. *Statistical Methods in Medical Research*, 2014. (In press).

[9] OS Miettinen. Proportion of disease caused or prevented by a given exposure, trait or intervention. *American Journal of Epidemiology*, 99(5):325–32, 1974.

[10] SD Walter. The estimation and interpretation of attributable risk in health research. *Biometrics*, 1(3):229–43, 1976.

[11] PM Gargiullo, RB Rothenberg, and HG Wilson. Confidence intervals, hypothesis tests, and sample sizes for the prevented fraction in cross-sectional studies. *Statistics in Medicine*, 14(1):51–72, 1995.

[12] S Greenland. Variance estimators for attributable fraction estimates consistent in both large strata and sparse data. *Statistics in Medicine*, 6(6):701–8, 1987.

[13] S Greenland and JM Robins. Conceptual problems in the definition and interpretation of attributable fractions. *American Journal of Epidemiology*, 128(6):1185–97, 1988.

[14] JM Robins and S Greenland. Estimability and estimation of excess and etiologic fractions. *Statistics in Medicine*, 8(7):845–59, 1989.

[15] KJ Rothman and S Greenland. *Modern Epidemiology*. Lippincott-Raven, Philadelphia, 1998.

[16] LA Cox. Probability of causation and the attributable proportion risk. *Risk Analysis*, 4(3):221–30, 1984.

[17] LA Cox. A new measure of attributable risk for public health applications. *Management Science*, 31(7):800–13, 1985.

[18] FA Seiler. Attributable risk, probability of causation, assigned shares, and uncertainty. *Environment International*, 12(6):635–41, 1986.

[19] P Cole and B MacMahon. Attributable risk percent in case-control studies. *British Journal of Preventive and Social Medicine*, 25(4):242–4, 1971.

[20] C Poole. A history of the population attributable fraction and related measures. *Annals of Epidemiology*, 25(3):147–54, 2015.

[21] S Greenland. Concepts and pitfalls in measuring and interpreting attributable fractions, prevented fractions, and causation probabilities. *Annals of Epidemiology*, 25(3):155–61, 2015.

[22] B MacMahon and TF Pugh. *Epidemiology: principles and methods*. Little, Brown, Boston, 1970.

[23] J Benichou. Methods of adjustment for estimating the attributable risk in case-control studies: a review. *Statistics in Medicine*, 10(11):1753–73, 1991.

[24] AS Whittemore. Estimating attributable risk from case-control studies. *American Journal of Epidemiology*, 117(1):76–85, 1983.

[25] P Bruzzi, SB Green, DP Byar, LA Brinton, and C Schairer. Estimating the population attributable risk for multiple risk factors using case-control data. *American Journal of Epidemiology*, 122(5):904–14, 1985.

[26] RC Rao. *Linear Statistical Inference and its Application.* Wiley, New York, 1965.

[27] SD Walter. The distribution of Levin's measure of attributable risk. *Biometrika*, 2(62):371–74, 1975.

[28] HM Leung and LL Kupper. Comparisons of confidence intervals for attributable risk. *Biometrics*, 37(2):293–302, 1981.

[29] AS Whittemore. Statistical methods for estimating attributable risk from retrospective data. *Statistics in Medicine*, 1(3):229–43, 1982.

[30] DR Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–220, 1972.

[31] D Spiegelman, E Hertzmark, and HC Wand. Point and interval estimates of partial population attributable risks in cohort studies: examples and software. *Cancer Causes Control*, 18(5):571–9, 2008.

[32] SJ Kuritz and JR Landis. Attributable risk ratio estimation from matched-pairs case-control data. *American Journal of Epidemiology*, 125(2):324–8, 1987.

[33] SJ Kuritz and JR Landis. Attributable risk estimation from matched case-control data. *Biometrics*, 44(2):355–67, 1988.

[34] SJ Kuritz and JR Landis. Summary attributable risk estimation from unmatched case-control data. *Statistics in Medicine*, 7(4):507–17, 1988.

[35] DG Kleinbaum, LL Kupper, and H Morgenstern. *Epidemiologic Research: Principles and Quantitative Methods.* Lifetime Learning Publications, Belmont, 1982.

[36] GA Satten and L Grummer-Strawn. Crosssectional study. In MH Gail and J Bénichou, editors, *Encyclopedia of Epidemiologic Methods*, pages 279–82. Wiley, Chichester, 2000.

[37] H Checkoway, N Pearce, and D Crawford-Brown. *Research Methods in Occupational Epidemiology.* Oxford University Press, New York, 1989.

[38] RE Tarone. On summary estimators of relative risk. *Journal of Chronic Diseases*, 34(9-10):463–8, 1981.

[39] JR Landis, TJ Sharp, SJ Kuritz, and G Koch. Mantel–Haenszel methods. In MH Gail and J Bénichou, editors, *Encyclopedia of Epidemiologic Methods*, pages 499–512. Wiley, Chichester, 2000.

[40] NE Breslow and NE Day. *Statistical Methods in Cancer Research. Vol. 2: The design and analysis of cohort studies.* International Agency for Research on Cancer, Lyon, 1987.

[41] N Mantel and W Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4):719–48, 1959.

[42] JJ Schlesselman. *Case-control Studies. Design, Conduct and Analysis.* Oxford University Press, New York, 1982.

[43] NE Breslow and NE Day. *Statistical Methods in Cancer Research.Vol. 1: The analysis of case–control studies.* International Agency for Research on Cancer, Lyon, 1980.

[44] A Ejigou. Estimation of attributable risk in the presence of confounding. *Biometrical Journal*, 21(2):155–65, 1979.

[45] SD Walter, CC Hsieh, and Q Liu. Effect of exposure misclassification on the mean squared error of population attributable risk and prevented fraction estimates. *Statistics in Medicine*, 26(26):4833–42, 2007.

[46] JP Klein and ML Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data.* Springer, New York, 2005.

[47] EL Kaplan and P Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–81, 1958.

[48] PK Andersen, Ø Borgan, RD Gill, and N Keiding. *Statistical Models Based on Counting Processes*. Springer, New York, 1993.

[49] M Greenwood. The natural duration of cancer. *Reports on Public Health and Medical Subjects*, (33):1–26, 1926.

[50] O Aalen. Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6(4):701–26, 1978.

[51] W Nelson. Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1(1):27–52, 1969.

[52] NE Breslow. Contribution to the discussion on the paper by DR Cox, Regression and life tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):216–17, 1972.

[53] N Breslow. Covariance analysis of censored survival data. *Biometrics*, 30(1):89–99, 1974.

[54] R Peto and J Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A (General)*, 135(2):185–206, 1972.

[55] EA Gehan. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52:203–23, 1965.

[56] RL Prentice. Linear rank tests with right censored data. *Biometrika*, 65(1):167–79, 1978.

[57] C Hill, C Com-Nougué, A Kramar, T Moreau, J O'Quigley, R Senoussi, and C Chastang. *Analyse Statistique des Données de Survie*. Médecine-Science Flammarion, Paris, 1996.

[58] JF Lawless. *Statistical Models and Methods for Lifetime Data*. Wiley, New York, 1982.

[59] DR Cox. Partial likelihood. *Biometrika*, 62(2):269–76, 1975.

[60] TR Fleming and DP Harrington. *Counting Processes and Survival Analysis*. Wiley, Hoboken, 2005.

[61] D Schoenfeld. Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*, 1(67):145–53, 1980.

[62] JW Vaupel, KG Manton, and E Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–54, 1979.

[63] TA Gooley, W Leisenring, J Crowley, and BE Storer. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in Medicine*, 18(6):695–706, 1999.

[64] JD Kalbfleisch and RL Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, Hoboken, 2002.

[65] JP Fine and RJ Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496, 1999.

[66] RL Prentice, JD Kalbfleisch, AV Peterson, N Flournoy, VT Farewell, and NE Breslow. The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):541–54, 1978.

[67] M Lunn and D McNeil. Applying Cox regression to competing risks. *Biometrics*, 51(2):524–32, 1995.

[68] H Putter, M Fiocco, and RB Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–430, 2007.

[69] MS Pepe and M Mori. Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in Medicine*, 12(8):737–51, 1993.

[70] A Allignol, A Latouche, J Yan, and JP Fine. A regression model for the conditional probability of a competing event: application to monoclonal gammopathy of unknown significance. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(1):135–42, 2011.

[71] JM Robins and DM Finkelstein. Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics*, 56(3):779–88, 2000.

[72] C Danieli, L Remontet, N Bossard, L Roche, and A Belot. Estimating net survival: the importance of allowing for informative censoring. *Statistics in Medicine*, 31(8):775–86, 2012.

[73] MA Laaksonen, T Härkänen, P Knekt, E Virtala, and H Oja. Estimation of population attributable fraction (PAF) for disease occurrence in a cohort study design. *Statistics in Medicine*, 29(7-8):860–74, 2010.

[74] S Murray and AA Tsiatis. Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics*, 52(1):137–51, 1996.

[75] L Chen. 'paf': Attributable fraction function for censored survival data, R package version 1.0, 2014. https://cran.r-project.org/web/packages/paf/index.html.

[76] X Bai, AA Tsiatis, and SM O'Brien. Doubly-robust estimators of treatment-specific survival distributions in observational studies with stratified sampling. *Biometrics*, 69(4):830–39, 2013.

[77] D Zeng, L Mao, and DY Lin. Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika*, 103(2):253–71, 2006.

[78] M Friedman. Piecewise exponential models for survival data with covariates. *Annals of Statistics*, 10(1):101–13, 1982.

[79] EL Korn, BI Graubard, and D Midthune. Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. *American Journal of Epidemiology*, 145(1):72–80, 1997.

[80] MA Laaksonen, E Virtala, P Knekt, H Oja, and T Härkänen. SAS macros for calculation of population attributable fraction in a cohort study design. *Journal of Statistical Software*, 47(7):1–25, 2011.

[81] M Stevenson, T Nunes, C Heuer, J Marshall, J Sanchez, R Thomton, J Reiczigel, J Robison-Cox, P Sebastiani, P Solymos, K Yoshida, G Jones, S Pirikahu, S Firestone, and R Kyle. 'epiR': Tools for the analysis of epidemiological data, R package version 0.9-79, 2016. https://cran.r-project.org/web/packages/epiR/index.html.

[82] L Schenck, E Atkinson, C Crowson, and T Therneau. 'Attribrisk': Population attributable risk, R package version 0.1, 2015. https://cran.r-project.org/web/packages/attribrisk/index.html.

[83] E Dahlqwist, J Zetterqvist, Y Pawitan, and A Sjölander. Model-based estimation of the attributable fraction for cross-sectional, case-control and cohort studies using the R package AF. *European Journal of Epidemiology*, 31(6):575–82, 2016.

[84] E Dahlqwist and A Sjölander. 'AF': Model-based estimation of confounder-adjusted attributable fractions, R package version 0.1.2, 2016. https://cran.r-project.org/web/packages/AF/index.html.

[85] A Fournier, S Mesrine, L Dossus, MC Boutron-Ruault, F Clavel-Chapelon, and N Chabbert-Buffet. Risk of breast cancer after stopping menopausal hormone therapy in the E3N cohort. *Breast Cancer Research and Treatment*, 145(2):535–43, 2014.

[86] L Dartois, G Fagherazzi, L Baglietto, MC Boutron-Ruault, S Delaloge, S Mesrine, and F Clavel-Chapelon. Proportion of premenopausal and postmenopausal breast cancers attributable to known risk factors: Estimates from the E3N-EPIC cohort. *International Journal of Cancer*, 138(10):2415–27, 2016.

[87] F Clavel-Chapelon, C Jadand, H Goulard, and C Guibout-Peigné. E3N, étude de cohorte sur les facteurs de risque de cancer auprès de femmes adhérentes de la MGEN. Description du protocole, des principales caractéristiques et de la population [E3N, a cohort study on cancer risk factors in MGEN women. Description of protocol, main characteristics and population]. *Bulletin du Cancer*, 83(12):1008–13, 1996.

[88] F Clavel-Chapelon, MJ van Liere, C Guibout, MY Niravong, H Goulard, C Le Corre, LA Hoang, J Amoyel, A Auquier, and E Duquesnel. E3N, a French cohort study

on cancer risk factors. E3N Group. Etude Epidémiologique auprès de femmes de l'Education Nationale. *European Journal of Cancer Prevention*, 6(5):473–8, 1997.

[89] A Fournier, F Berrino, E Riboli, V Avenel, and F Clavel-Chapelon. Breast cancer risk in relation to different types of hormone replacement therapy in the E3N-EPIC cohort. *International Journal of Cancer*, 114(3):448–54, 2005.

[90] ACM Thiébaut and J Bénichou. Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Statistics in Medicine*, 23(24):3803–20, 2004.

[91] CS Crowson, TM Therneau, and WM O'Fallon. Attributable risk estimation in cohort studies. Technical report, 2009. http://www.mayo.edu/research/documents/biostat-82pdf/doc-10027843.

[92] A Alpérovitch, T Kurth, M Bertrand, M-L Ancelin, C Helmer, S Debette, and C Tzourio. Primary prevention with lipid lowering drugs and long term risk of vascular events in older people: population based cohort study. *BMJ*, 350:h2335, 2015.

[93] 3C Study Group. Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population. *Neuroepidemiology*, 22(6):316–25, 2003.

[94] A Alpérovitch, JF Dartigues, K Ritchie, C Tzourio, B Mazoyer, P Amouyel, P Ducimetière, and le Groupe 3C. L'étude des Trois Cités : relation entre pathologie vasculaire et démence. *Revue Médicale de l'Assurance Maladie*, 2(37):117–24, 2006.

[95] SL West, BL Strom, and C Poole. Validity of pharmacoepidemiologic drug and diagnosis data. In *Pharmacoepidemiology*, pages 709–65. Wiley, Chichester, 2005.

[96] P Noize, F Bazin, C Dufouil, N Lechevallier-Michel, ML Ancelin, JF Dartigues, C Tzourio, N Moore, and A Fourrier-Réglat. Comparison of health insurance claims and patient interviews in assessing drug use: data from the Three-City (3C) Study. *Pharmacoepidemiology and Drug Safety*, 18(4):310–9, 2009.

[97] P Noize, F Bazin, A Pariente, C Dufouil, ML Ancelin, C Helmer, N Moore, and A Fourrier-Réglat. Validity of chronic drug exposure presumed from repeated patient interviews varied according to drug class. *Journal of Clinical Epidemiology*, 65(10):1061–8, 2012.

[98] P Barberger-Gateau, C Raffaitin, L Letenneur, C Berr, C Tzourio, JF Dartigues, and A Alpérovitch. Dietary patterns and risk of dementia: the Three-City cohort study. *Neurology*, 69(20):1921–30, 2007.

[99] HA. Guess. Exposure-time-varying hazard function ratios in case-control studies of drug effects. *Pharmacoepidemiology and Drug Safety*, 15(2):81–92, 2006.

[100] JNS Eisenberg, BL Lewis, TC Porco, AH Hubbard, and JM Colford. Bias due to secondary transmission in estimation of attributable risk from intervention trials. *Epidemiology*, 14(4):442–50, 2003.

[101] S Schneeweiss, T Stürmer, and M Maclure. Case-crossover and case-time-control designs as alternatives in pharmacoepidemiologic research. *Pharmacoepidemiology and Drug Safety*, 6 Suppl 3:S51–9, 1997.

[102] KS Kaye, AD Harris, M Samore, and Y Carmeli. The case-case-control study design: addressing the limitations of risk factor studies for antimicrobial resistance. *Infection Control and Hospital Epidemiology*, 26(4):346–51, 2005.

[103] C P Farrington, J Nash, and E Miller. Case series analysis of adverse reactions to vaccines: a comparative evaluation. *American Journal of Epidemiology*, 143(11):1165–73, 1996.

[104] MN Hocine and M Chavance. La méthode de la série de cas [The case series method]. *Revue d'Épidémiologie et de Santé Publique*, 58(6):435–40, 2010.

[105] HJ Whitaker, CP Farrington, B Spiessens, and P Musonda. Tutorial in biostatistics: the self-controlled case series method. *Statistics in Medicine*, 25(10):1768–97, 2006.

[106] HJ Whitaker, MN Hocine, and CP Farrington. The methodology of self-controlled case series studies. *Statistical Methods in Medical Research*, 18(1):7–26, 2009.

[107] P Farrington, S Pugh, A Colville, A Flower, J Nash, P Morgan-Capner, M Rush, and E Miller. A new method for active surveillance of adverse events from diphtheria/tetanus/pertussis and measles/mumps/rubella vaccines. *Lancet*, 345(8949):567–9, 1995.

[108] A Allorant. Estimation de la fraction de risque attribuable dans les séries de cas. Technical report, Mémoire de Master 2 Recherche en Santé Publique parcours Épidémiologie sous la direction d'Anne Thiébaut, Sylvie Escolano et Pascale Tubert-Bitter, Université Paris Sud, Paris, 2016.

Annexes

Annexe A

Implémentation des approches non paramétriques KM et KMP

L'estimation de la variance du RA pour les approches non paramétriques KM et KMP dépend des résidus martingales $M(t)$ et $M_0(t)$ et des expressions de $\eta_1(t)$ et de $\eta_2(t)$ (cf. chapitre 2). Nous détaillons dans cette partie les différentes approches utilisées pour calculer ces expressions mais aussi le code utilisé pour l'estimation du RA et de sa variance.

A.1 Expression théorique

— **Expressions de $M(t)$ et $M_0(t)$**

Pour se fixer les idées et bien coder les résidus martingales, dans le cas d'un modèle non paramétrique, nous avons construit la matrice suivante qui correspond à la

matrice d'observation :

$$\begin{array}{c} Z \quad T \quad Y(t_1) \quad Y(t_2) \quad \dots \quad Y(t_j) \quad \dots \quad Y(t_n) \\ \begin{pmatrix} z_1 & t_1 & 1 & 0 & \dots & 0 & \dots & 0 \\ z_2 & t_2 & 1 & 1 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ z_i & t_i & 1 & 1 & \dots & \mathbb{1}_{t_i \geq t_j} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ z_n & t_n & 1 & 1 & \dots & 1 & \dots & 1 \end{pmatrix} \end{array}$$

Dans cette matrice sont représentés, en ligne, les individus $i = 1, .., n$ et, en colonne, les temps observés en ordre croissant t_j, $j = 1, ..., n$. Pour un individu $i = 1, ..., n$, sa martingale au temps $t \geq 0$ est définie comme :

$$M_i(t) = N_i(t) - \int_0^t Y_i(u) d\hat{\Lambda}(u)$$

où $Y_i(t) = \mathbb{1}_{t_i \geq t}$ est l'indicatrice à risque et $\Lambda(t) = -\ln\{S(t)\}$.

Considérant t_i le délai d'observation du sujet i, deux cas se présentent :

1. Si $t_i \leq t$

$$\begin{aligned} M_i(t) &= N_i(t) - \int_0^t \mathbb{1}_{t_i \geq u} d\Lambda(u) \\ &= N_i(t) - \int_0^{t_i} 1 \times d\Lambda(u) - \int_{t_i}^t 0 \times d\Lambda(u) \\ &= N_i(t) - [\Lambda(u)]_0^{t_i} \\ &= N_i(t) - \Lambda(t_i) + \Lambda(0) \end{aligned}$$

2. Si $t_i > t$

$$\begin{aligned} M_i(t) &= N_i(t) - \int_0^t 1 \times d\Lambda(u) \\ &= N_i(t) - [\Lambda(u)]_0^t \\ &= N_i(t) - \Lambda(t) + \Lambda(0) \end{aligned}$$

En résumé :

$$M_i(t) = N_i(t) - Y_i(t)\left[\Lambda(t) - \Lambda(t_i)\right] - \Lambda(t_i) + \Lambda(0)$$

Nous avons fait de même avec l'équivalent de la martingale chez les non exposés :

$$M_{0i}(t) = \mathbb{1}_{Z_i=0}\{N_i(t) - Y_i(t)\left[\Lambda_0(t) - \Lambda_0(t_i)\right] - \Lambda_0(t_i) + \Lambda_0(0)\}$$

où $\Lambda_0(t) = -\ln\left[S_0(t)\right]$.

— **Expressions de $\eta_1(t)$ et $\eta_2(t)$**

Afin d'implémenter η_1 et η_2, nous avons dans un premier temps approché les proportions de sujets à risque au temps u

$$\mathbf{E}\left[Y(u)\right] = \mathbf{P}\left[Y(u) = 1\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{t_i \geq u} \text{ et}$$

$$\mathbf{E}\left[\mathbb{1}_{Z=0}Y(u)\right] = \mathbf{P}\left[Z = 0 \cap Y(u) = 1\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Z_i=0, t_i \geq u}$$

comme des moyennes sur l'ensemble des individus pour la première et sur l'ensemble des sujets non exposés pour la seconde. Nous avons, pour chaque individu i au temps t_j :

$$\eta_{1i}(t_j) = -S_0(t_j)\int_0^{t_j} \frac{dM_{0i}(u)}{\mathbf{E}\left[\mathbb{1}_{Z_i=0}Y_i(u)\right]}.$$

Nous avons calculé $dM_{0i}(t_j)$ comme suit :

$$dM_{0i}(t_1) = M_{0i}(t_1) \text{ et } \forall j \geq 2 \quad dM_{0i}(t_j) = M_{0i}(t_j) - M_{0i}(t_{j-1}).$$

L'expression de $\eta_{1i}(t_j)$ a été approchée comme produit de la fonction de survie estimée en t_j et de la somme cumulée jusqu'en t_j du rapport $\dfrac{dM_{0i}(u)}{\mathbf{E}\left[\mathbb{1}_{Z_i=0}Y_i(u)\right]}$.
Nous avons suivi le même raisonnement pour η_2.

A.2 Code R

A.2.1 Estimation du RA et de sa variance par l'approche KM

```
require(survival)
```

```
require(RcppArmadillo)
require(Rcpp)
sourceCpp("AR_KM_KM.cpp") # permet de compiler un code C++ dans R
# Le fichier AR_KM_KM.cpp contient le programme suivant
#include <RcppArmadillo.h>
// [[Rcpp::depends(RcppArmadillo)]]
using namespace Rcpp;
// [[Rcpp::export]]
arma::mat fonction_test1(const arma::vec& temps, const int n, const int m){
  arma::mat res = arma::mat(n,m);
  for (int i = 0; i < n; i++) {for (int j = 0; j < m; j++){
      res(i, j) = (temps[i] >= temps[j]) ? true : false;
      }}
  return res;
  }
// [[Rcpp::export]]
arma::mat fonction_test2(const arma::vec& temps, const arma::vec& status,
const int n, const int m){arma::mat res = arma::mat(n,m);
for (int i = 0; i < n; i++) {for (int j = 0; j < m; j++){
res(i, j) = (temps[i] <= temps[j] && status[i]==1) ? true : false;
    }}
  return res;
}
// [[Rcpp::export]]
arma::mat fonction_Mt_KM(const arma::mat& Nt, const arma::mat& yt,
const arma::vec& Lambdat, const double Lt0) {
arma::mat Mt = arma::mat(Nt.n_rows, Nt.n_cols);
for (int i = 0; i < Nt.n_rows; i++){for (int j = 0; j < Nt.n_cols; j++){
Mt(i, j) = Nt(i, j) - yt(i, j)*(Lambdat[j] - Lambdat[i]) - Lambdat[i] + Lt0;
```

```
    }}
  return Mt;
}

// [[Rcpp::export]]
arma::mat fonction_Mt0_KM(const arma::mat& Nt, const arma::mat& yt,
const arma::vec& Lambdat0, const double L0, const arma::vec& donneeZ){
arma::mat Mt0 = arma::mat(Nt.n_rows, Nt.n_cols);
double Z;
for (int i = 0; i < Nt.n_rows; i++){Z = (donneeZ[i] == 0) ? 1 : 0;
for (int j = 0; j < Nt.n_cols; j++){
Mt0(i, j) = Z * (Nt(i, j) - yt(i, j)*
(Lambdat0[j] - Lambdat0[i]) - Lambdat0[i] + L0);
    }}
  return Mt0;
}
// [[Rcpp::export]]
arma::mat fonction_eta2_KM(const arma::mat& dMt,
const arma::vec& Ymoyen) {
arma::mat eta = arma::mat(dMt.n_rows, dMt.n_cols);
for (int i = 0; i < dMt.n_rows; i++){eta(i, 0) = dMt(i, 0) / Ymoyen[0];
for (int j = 1; j < dMt.n_cols; j++){
eta(i, j) = eta(i, j-1) + dMt(i, j)/Ymoyen[j];
    }}
  return eta;
}
// [[Rcpp::export]]
arma::mat fonction_eta1_KM(const arma::mat& dMt0, const arma::vec& Y0moyen){
arma::mat eta = arma::mat(dMt0.n_rows, dMt0.n_cols);
```

```
  for (int i = 0; i < dMt0.n_rows; i++) {
    eta(i, 0) = dMt0(i, 0) / Y0moyen[0];
    for (int j = 1; j < dMt0.n_cols; j++) {
      eta(i, j) = eta(i, j-1) + dMt0(i, j) / Y0moyen[j];
    }}
  return eta;
}

// [[Rcpp::export]]
arma::mat fonction_eta_KM(const arma::mat& eta1b, const arma::mat& eta2b,
const arma::vec& S, const arma::vec& S0) {
arma::mat eta = arma::mat(eta1b.n_rows, eta1b.n_cols);
for (int j = 0; j < eta1b.n_cols; j++) {
eta.col(j) = (1 / (1 - S[j])) * (eta1b.col(j)- eta2b.col(j)*(1-S0[j])/(1-S[j]));
  }
  return eta;
}

# Calcul du RA et de sa variance
PafKaplanKaplan = function(donnee){
  donnee = donnee[order(donnee$temps), ]
  donnee <- transform(donnee, temps = round(temps, digits = 5))
  donnee0 <- donnee[which(donnee$Z==0),]
  fit = survfit(Surv(donnee$temps, donnee$status==1)~1, type = "kaplan-meier")
  donneeKM <- data.frame(temps=fit$time, S=fit$surv)
  stepS <- stepfun(fit$time, c(1, fit$surv))
  donnee <- merge(donnee, donneeKM, by="temps", all=TRUE)
  fit0 <- survfit(Surv(donnee0$temps, donnee0$status==1)~1, type = "kaplan-meier")
  donneeKM0 <- data.frame(temps=fit0$time, S0=fit0$surv)
```

```
donnee <- merge(donnee, donneeKM0, by="temps", all=TRUE)
stepS0 <- stepfun(fit0$time, c(1, fit0$surv))
donnee <- transform(donnee, S0 = stepS0(temps))
idS = which(donnee[,"S"]==1)
if(length(idS) > 0)   donnee = donnee[-idS,]
donnee$AR = (donnee$S0-donnee$S)/(1-donnee$S)
donnee$Lambdat = -log(donnee[,"S"])
donnee$Lambdat0 = -log(donnee[,"S0"])
Lt0 = -log(stepS(0))
L0 = -log(stepS0(0))
donnee$Ymoyen = sapply(donnee$temps, function(j){
  mean(donnee$temps >= j)
  })
donnee$Y0moyen = sapply(donnee$temps, function(j){
mean(donnee$temps >= j & donnee$Z==0)
  })
id = which(donnee[,"S"]==0 | donnee[,"S0"]==0 | donnee[,"Ymoyen"]==0 |
donnee[,"Y0moyen"]==0)
if (length(id) > 0) donnee = donnee[-id,]
nn = length(donnee$AR)
Nt = sapply(donnee$temps, function(j){n = (donnee$temps <= j & donnee$status==1)
  c(n)})
yt = sapply(donnee$temps,function(j) { y = (donnee$temps >= j)
  c(y)})
Mt = fonction_Mt_KM(Nt, yt, donnee$Lambdat, Lt0)
Mt0 = fonction_Mt0_KM(Nt, yt, donnee$Lambdat0, L0, donnee$Z)
dMt = cbind(Mt[,1],t(apply(Mt, MARGIN=1, diff)))
dMt0 = cbind(Mt0[,1],t(apply(Mt0, MARGIN=1, diff)))
eta2 = fonction_eta2_KM(dMt, donnee$Ymoyen)
```

```
  eta1 = fonction_eta1_KM(dMt0, donnee$Y0moyen)
  eta2b = - eta2 %*% diag(donnee[,"S"])
  eta1b = - eta1 %*% diag(donnee[,"S0"])
  eta = fonction_eta_KM(eta1b, eta2b, donnee$S, donnee$S0)
  donnee$SEAR = sqrt(apply(eta^2, MARGIN=2, sum))/nn
  result <- donnee[,c("temps", "S", "S0", "AR", "SEAR")]
  idAR <- which(result$AR==1)
  if(length(idAR)>0) result = result[-idAR,]
  result$SElogAR = result$SEAR/abs(result$AR-1)
  result$IC_inf = 1-(1-result$AR)*exp(1.96*result$SEAR/(1-result$AR))
  result$IC_sup = 1-(1-result$AR)*exp(-1.96*result$SEAR/(1-result$AR))
  return(result)
}
```

A.2.2 Estimation du RA et de sa variance par l'approche KMP

```
require(survival)
require(RcppArmadillo)
require(Rcpp)
sourceCpp("AR_KM_KMP.cpp")
# contenu du fichier AR_KM_KMP.cpp
#include <RcppArmadillo.h>
// [[Rcpp::depends(RcppArmadillo)]]
using namespace Rcpp;
// [[Rcpp::export]]
arma::mat fonction_test1(const arma::vec& temps,
                         const int n, const int m) {
  arma::mat res = arma::mat(n, m);
  for (int i = 0; i < n; i++) {for (int j = 0; j < m; j++) {
```

```
      res(i, j) = (temps[i] >= temps[j]) ? 1 : 0;
    }}
 return res;
}
// [[Rcpp::export]]
arma::mat fonction_test2(const arma::vec& temps,
const arma::vec& status, const int n, const int m){
arma::mat res = arma::mat(n,m);
for (int i = 0; i < n; i++) {for (int j = 0; j < m; j++) {
      res(i, j) = (temps[i] <= temps[j] && status[i]==1) ? true : false;
    }}
  return res;
}
// [[Rcpp::export]]
arma::mat fonction_Mt0_KMP(const arma::mat& Nt, const arma::mat& yt,
const arma::vec& Lambda0, const double L0, const arma::vec& donneeZ) {
arma::mat Mt0 = arma::mat(Nt.n_rows, Nt.n_cols);
double Z;
for (int i = 0; i < Nt.n_rows; i++){Z = (donneeZ[i] == 0) ? 1 : 0;
for (int j = 0; j < Nt.n_cols; j++) {
Mt0(i, j) = Z * (Nt(i, j) - yt(i, j)*(Lambda0[j] - Lambda0[i]) - Lambda0[i] + L0);
    }}
  return Mt0;
}

// [[Rcpp::export]]
arma::mat fonction_Mt1_KMP(const arma::mat& Nt, const arma::mat& yt,
const arma::vec& Lambda1, const double L1) {
arma::mat Mt1 = arma::mat(Nt.n_rows, Nt.n_cols);
```

```
for (int i = 0; i < Nt.n_rows; i++) {for (int j = 0; j < Nt.n_cols; j++) {
Mt1(i, j) = Nt(i, j) - yt(i, j)*(Lambda1[j] - Lambda1[i]) - Lambda1[i] + L1;
    }}
  return Mt1;
}
// [[Rcpp::export]]
arma::mat fonction_Mt00_KMP(const arma::mat& Nt, const arma::mat& yt,
const arma::vec& Lambda0, const double   L0){
arma::mat Mt00 = arma::mat(Nt.n_rows, Nt.n_cols);
for (int i = 0; i < Nt.n_rows; i++) {for (int j = 0; j < Nt.n_cols; j++) {
Mt00(i, j) = Nt(i, j) - yt(i, j)*(Lambda0[j] - Lambda0[i]) - Lambda0[i] + L0;
    }}
   return Mt00;
  }
// [[Rcpp::export]]
arma::mat fonction_dMt1_Y1_KMP(const arma::mat& dMt1,
const arma::vec& Y1moyen) {
arma::mat int1 = arma::mat(dMt1.n_rows, dMt1.n_cols);
for (int i = 0; i < dMt1.n_rows; i++){int1(i, 0) = dMt1(i, 0) / Y1moyen[0];
for (int j = 1; j < dMt1.n_cols; j++){
int1(i, j) = int1(i, j-1) + dMt1(i, j) / Y1moyen[j];
    }}
  return int1;
}
// [[Rcpp::export]]
arma::mat fonction_dMt0_Y0_KMP(const arma::mat& dMt0,
const arma::vec& Y0moyen){
arma::mat eta = arma::mat(dMt0.n_rows, dMt0.n_cols);
for (int i = 0; i < dMt0.n_rows; i++){eta(i, 0) = dMt0(i, 0) / Y0moyen[0];
```

```
for (int j = 1; j < dMt0.n_cols; j++){
eta(i, j) = eta(i, j-1) + dMt0(i, j) / Y0moyen[j];
    }}
    return eta;
}

// [[Rcpp::export]]
arma::mat fonction_dMt00_Y0_KMP(const arma::mat& dMt00, const arma::vec& Y0moyen){
arma::mat int2 = arma::mat(dMt00.n_rows, dMt00.n_cols);
for (int i = 0; i < dMt00.n_rows; i++) {int2(i, 0) = dMt00(i, 0) / Y0moyen[0];
for (int j = 1; j < dMt00.n_cols; j++) {
int2(i, j) = int2(i, j-1) + dMt00(i, j) / Y0moyen[j];
    }}
  return int2;
 }
// [[Rcpp::export]]
arma::mat fonction_eta2_KMP(const arma::mat& int0, const arma::mat& int1,
const arma::vec& S1, const arma::vec& S, const arma::vec& S0,
const double p_est, const arma::vec& donneeZ ) {
arma::mat eta2b = arma::mat(int0.n_rows, int0.n_cols);
int Z0;
int Z1;
for (int i=0; i < int0.n_rows; i++) {
Z0 = (donneeZ[i] == 0) ? 1 : 0;
Z1 = (donneeZ[i] == 1) ? 1 : 0;
for (int j = 0; j < int0.n_cols; j++) {
eta2b(i,j) = Z0*S0[j]+Z1*S1[j]-S[j]-(1-p_est)*Z0*S0[j]*int0(i,j)-
        p_est*Z1*S1[j]*int1(i,j);
    }}
```

```
  return eta2b;
}
// [[Rcpp::export]]
arma::mat fonction_eta_KMP(const arma::mat& eta1b, const arma::mat& eta2b,
const arma::vec& S, const arma::vec& S0) {
arma::mat eta = arma::mat(eta1b.n_rows, eta1b.n_cols);
for (int j = 0; j < eta1b.n_cols; j++) {
eta.col(j) = (1/(1 - S[j]))*(eta1b.col(j) - eta2b.col(j)*(1-S0[j])/(1-S[j]));
  }
  return eta;
}

PafKaplanKaplanPondere = function(donnee){
  donnee = donnee[order(donnee$temps), ]
  donnee <- transform(donnee, temps = round(temps, digits = 5))
  donnee0 <- donnee[which(donnee$Z==0),]
  donnee1  <- donnee[which(donnee$Z == 1),]
  p1 = sum(donnee[,"Z"])/length(donnee[,"Z"])
  fit0 <- survfit(Surv(donnee0$temps, donnee0$status==1)~1)
  donneeKM0 <- data.frame(fit0$time, fit0$surv)
  names(donneeKM0) <- c("temps", "S0")
  donnee <- merge(donnee, donneeKM0, by="temps", all=TRUE)
  stepS0 <- stepfun(fit0$time, c(1, fit0$surv))
  donnee <- transform(donnee, S0 = stepS0(temps))
  fit1 <- survfit(Surv(donnee1$temps, donnee1$status) ~ 1)
  donneeKM1 <- data.frame(fit1$time, fit1$surv)
  names(donneeKM1) <- c("temps","S1")
  donnee <- merge(donnee, donneeKM1, by = "temps", all = TRUE)
  stepS1 <- stepfun(fit1$time, c(1, fit1$surv))
```

```
donnee <- transform(donnee, S1 = stepS1(temps))
donnee$S = (1-p1)*donnee$S0 + p1*donnee$S1
idS <- which(donnee$S==1)
if(length(idS) > 0) donnee = donnee[-idS, ]
donnee$AR = (donnee$S0-donnee$S)/(1-donnee$S)
donnee$Lambda0 = -log(donnee$S0)
donnee$Lambda1 = -log(donnee$S1)
L0 = -log(stepS0(0))
L1 = -log(stepS1(0))
donnee$Y0moyen = sapply(donnee$temps, function(j){
  Y_moyen = mean(donnee$temps >= j & donnee$Z==0)
  c(Y_moyen)})
donnee$Y1moyen = sapply(donnee$temps, function(j){
  Y1 = mean(donnee$temps >= j & donnee$Z==1)
  c(Y1)})
id = which(donnee[,"S"]==0 | donnee[,"S0"]==0 | donnee[,"S1"]==0 |
    donnee[,"Y1moyen"]==0 | donnee[,"Y0moyen"]==0)
if (length(id) > 0) donnee = donnee[-id,]
nn = length(donnee$AR)
Nt = sapply(donnee$temps, function(j){n = (donnee$temps <= j & donnee$status==1)
  c(n)})

yt = sapply(donnee$temps,function(j) { y = (donnee$temps >= j)
  c(y)})
Mt0 = fonction_Mt0_KMP(Nt, yt, donnee$Lambda0, L0, donnee$Z)
Mt1 = fonction_Mt1_KMP(Nt, yt, donnee$Lambda1, L1)
Mt00 = fonction_Mt00_KMP(Nt, yt, donnee$Lambda0, L0)
dMt0 = cbind(Mt0[,1],t(apply(Mt0, MARGIN=1, diff)))
dMt1 = cbind(Mt1[,1],t(apply(Mt1, MARGIN=1, diff)))
```

```
  dMt00 = cbind(Mt00[,1],t(apply(Mt00, MARGIN=1, diff)))
  int0 = fonction_dMt00_Y0_KMP(dMt00, donnee$Y0moyen)
  int1 = fonction_dMt1_Y1_KMP(dMt1, donnee$Y1moyen)
  eta1 = fonction_dMt0_Y0_KMP(dMt0, donnee$Y0moyen)
  eta1b = - eta1 %*% diag(donnee[,"S0"])
  eta2b = fonction_eta2_KMP(int0, int1, donnee$S1, donnee$S, donnee$S0,
  p1, donnee$Z)
  eta = fonction_eta_KMP(eta1b, eta2b, donnee$S, donnee$S0)
  donnee$SEAR = sqrt(apply(eta^2, MARGIN=2, sum))/nn
  result <- donnee[,c("temps", "S", "S0", "AR", "SEAR")]
  idAR <- which(result$AR==1)
  if(length(idAR)>0) result = result[-idAR,]
  result$SElogAR = result$SEAR/abs(result$AR-1)
  result$IC_inf = 1-(1-result$AR)*exp(1.96*result$SEAR/(1-result$AR))
  result$IC_sup = 1-(1-result$AR)*exp(-1.96*result$SEAR/(1-result$AR))
  return(result)
}
```

Annexe B

Reproduction des résultats de simulation pour le risque attribuable défini à partir des fonctions de risque instantané

Dans cette partie, nous détaillons l'approche utilisée pour reproduire le plan de simulation de Chen *et al.* [5] pour le risque attribuable $\varphi(t)$ défini en fonction des risques instantanés, les auteurs nous ayant fourni leur programme d'estimation sous le logiciel `R`.

B.1 Génération des temps de censure

Nous avons considéré une censure C indépendante de la covariable Z en générant un délai avant censure selon une loi uniforme sur $[0, \tau]$ où τ est la durée maximale de l'étude. Cela revient à supposer que les sujets sont inclus de façon uniforme jusqu'à la fin de l'étude fixée à τ.

On sait que la probabilité d'avoir un événement est de la forme :

$$\mathbf{P}(T < C) = \int \int \mathbb{1}_{t<c} f_T(t) f_C(c) dt dc.$$

Or $F(t|z) = 1 - \exp[-\Lambda(t|z)]$, ce qui donne par dérivation la densité de $T|Z$:

$$f_{T|Z}(t|z) = \lambda(t|z)\exp[-\Lambda(t|z)] = \lambda(t|z)S(t|z) = -S'(t|z)$$

où $S'(t|z)$ désigne la dérivée par rapport à t de la fonction de survie conditionnelle $S(t|z) = \exp[-\Lambda(t[z)]$.

On en déduit la densité conjointe $f(t;z) = f_{T|Z}(t|z)f(z) = -S'(t|z)f(z)$.

Par ailleurs, Z suivant une loi de Bernoulli de paramètre p, sa densité est :

$$f(z) = p^z(1-p)^{(1-z)}\mathbb{1}_{0,1}(z).$$

Par conséquent, selon le théorème des probabilités totales,

$$f_T(t) = \sum_{z=0}^{1} f(t;z) = \sum_{z=0}^{1} -S'(t|z)f(z)$$

et

$$\begin{aligned}
\mathbf{P}(T<C) &= \int\int -\sum_{z=0}^{1} f(z)\mathbb{1}_{t<c}S'(t|z)f_C(c)dcdt \\
&= \sum_{z=0}^{1} f(z)\int\int -S'(t|z)\mathbb{1}_{t<c}f_C(c)dcdt \\
&= \sum_{z=0}^{1} f(z)\int\left\{\int_0^c -S'(t|z)dt\right\}f_C(c)dc \\
&= \sum_{z=0}^{1} f(z)\int[-S(t|z)]_0^c f_C(c)dc \\
&= \sum_{z=0}^{1} f(z)\int(1-S(c|z))f_C(c)dc.
\end{aligned}$$

Or, sous l'hypothèse d'une censure uniforme, $f_C(c) = \frac{1}{\tau}\mathbb{1}_{[0,\tau]}(c)$, donc

$$\begin{aligned}
\mathbf{P}(T<C) &= \sum_{z=0}^{1} f(z)\int(1-S(c|z))\frac{1}{\tau}\mathbb{1}_{[0,\tau]}(c)dc \\
&= \sum_{z=0}^{1} f(z)\frac{1}{\tau}\int_0^\tau(1-S(c|z))dc \\
&= \sum_{z=0}^{1} f(z)\frac{1}{\tau}\left[\tau - \int_0^\tau S(c|z)dc\right].
\end{aligned}$$

Posons $J = \int_0^\tau S(c|z)dc$. Partant d'un modèle à risques proportionnels avec un risque de base constant (modèle exponentiel),

$$\lambda(t|Z) = \lambda_0 \exp(\beta Z),$$

on en déduit la fonction de risque cumulé conditionnelle aux covariables:

$$\Lambda(t|Z) = \lambda_0 t \exp(\beta Z)$$

et la fonction de survie conditionnelle :

$$S(t|Z) = \exp\left[-\Lambda(t|Z)\right] = \exp\left[-\lambda_0 t \exp(\beta Z)\right].$$

D'où

$$\begin{aligned} J &= \int_0^\tau S(c|z)dc \\ &= \int_0^\tau \exp\left[-\lambda_0 c \exp(\beta z)\right] dc. \end{aligned}$$

Posons $x = -\lambda_0 c \exp(\beta z)$ donc $dx = -\lambda_0 \exp(\beta z)dc$ et

$$\begin{aligned} J &= \int_0^\tau \exp\left[-\lambda_0 c \exp(\beta z)\right] dc \\ &= \int_0^{-\lambda_0 \tau \exp(\beta z)} \exp(x) \frac{-dx}{\lambda_0 \exp(\beta z)} \\ &= \frac{-1}{\lambda_0 \exp(\beta z)} \int_0^{-\lambda_0 \tau \exp(\beta z)} \exp(x)dx \\ &= \frac{-1}{\lambda_0 \exp(\beta z)} \left[\exp(x)\right]_0^{-\lambda_0 \tau \exp(\beta z)} \\ &= \frac{-1}{\lambda_0 \exp(\beta z)} \left\{\exp\left[-\lambda_0 \tau \exp(\beta z)\right] - 1\right\}. \end{aligned}$$

On en déduit

$$\begin{aligned} \mathbf{P}(T < C) &= \sum_{z=0}^{1} f(z) \frac{1}{\tau} \{\tau - \int_0^\tau S(c|z)dc\} \\ &= \sum_{z=0}^{1} f(z) \frac{1}{\tau} \{\tau - J\} \\ &= \sum_{z=0}^{1} f(z) \frac{1}{\tau} \left\{\tau + \frac{1}{\lambda_0 \exp(\beta z)} \left\{\exp\left[-\lambda_0 \tau \exp(\beta z)\right] - 1\right\}\right\}. \end{aligned}$$

Si $\beta = 0$, cette relation devient :

$$\begin{aligned}\mathbf{P}(T < C) &= \sum_{z=0}^{1} f(z)\frac{1}{\tau}\left\{\tau + \frac{1}{\lambda_0}\left\{\exp\left[-\lambda_0\tau\right] - 1\right\}\right\} \\ &= \frac{1}{\tau}\left\{\tau + \frac{1}{\lambda_0}\left[\exp(-\lambda_0\tau) - 1\right]\right\} \\ &= 1 + \frac{1}{\lambda_0\tau}\left[\exp(-\lambda_0\tau) - 1\right].\end{aligned}$$

Supposons que $q = 1 - \mathbf{P}(T < C)$ (probabilité d'être censuré) et $x = \exp(-\lambda_0\tau)$, alors

$$\begin{aligned}1 - q &= 1 + \frac{1}{\lambda_0\tau}\left[\exp(-\lambda_0\tau) - 1\right] \\ -q &= \frac{1}{-\ln(x)}\left(x - 1\right) \\ \ln(x)q &= x - 1 \\ q\ln(x) - x + 1 &= 0.\end{aligned}$$

Cette équation n'admettant pas de solution analytique, nous l'avons résolue de façon numérique. En remplaçant x par son expression, nous avons

$$-q\lambda_0\tau - \exp(-\lambda_0\tau) + 1 = 0.$$

Dans le cas où $\beta = 0$, la date de fin d'étude ne dépend pas de la probabilité d'exposition. Pour un risque de base λ_0 et un pourcentage de censure q donnés, nous choisissons une valeur approchée de τ qui annule l'équation. Cette valeur sera fixée dans les simulations.

Dans le cas où $\beta = \ln(2)$, nous avons :

$$\begin{aligned}\mathbf{P}(T < C) &= \sum_{z=0}^{1} f(z)\frac{1}{\tau}\left\{\tau + \frac{1}{\lambda_0\exp(\beta z)}\left\{\exp\left[-\lambda_0\tau\exp(\beta z)\right] - 1\right\}\right\} \\ &= \frac{p}{\tau}\left\{\tau + \frac{1}{\lambda_0\exp(\beta)}\left\{\exp\left[-\lambda_0\tau\exp(\beta)\right] - 1\right\}\right\} + \frac{1-p}{\tau}\left\{\tau + \frac{1}{\lambda_0}\left[\exp\left(-\lambda_0\tau\right) - 1\right]\right\} \\ &= 1 + \frac{p}{\lambda_0\tau\exp(\beta)}\left\{\exp\left[-\lambda_0\tau\exp(\beta)\right] - 1\right\} + \frac{1-p}{\lambda_0\tau}\left[\exp\left(-\lambda_0\tau\right) - 1\right].\end{aligned}$$

Posons $x = \exp(-\lambda_0 \tau)$ et $q = 1 - P(T < C)$. On obtient alors :

$$\begin{aligned}
1 - q &= 1 + \frac{p}{\lambda_0 \tau \exp(\beta)} \left\{ \exp\left[-\lambda_0 \tau \exp(\beta) \right] - 1 \right\} + \frac{1-p}{\lambda_0 \tau} \left[\exp\left(-\lambda_0 \tau \right) - 1 \right] \\
-q &= \frac{p}{-2\ln(x)} \left(x^2 - 1 \right) + \frac{1-p}{-ln(x)} \left(x - 1 \right) \\
q &= \frac{p}{2\ln(x)} \left(x^2 - 1 \right) + \frac{1-p}{\ln(x)} \left(x - 1 \right) \\
2q\ln(x) &= p(x^2 - 1) + 2(1-p)(x-1) \\
2q\ln(x) &= (x-1)(px + p + 2 - 2p) \\
2q\ln(x) &= (x-1)(px - p + 2).
\end{aligned}$$

En remplaçant x par son expression, nous avons :

$$-2q\lambda_0\tau - \left[\exp(-\lambda_0 \tau) - 1 \right] \left[p \exp(-\lambda_0 \tau) - p + 2 \right] = 0.$$

L'équation ci-dessus a été résolue de manière numérique pour obtenir la date de fin d'étude pour une probabilité d'exposition p, un risque de base λ_0 et un pourcentage de censure q donnés.

B.2 Choix des paramètres pour l'étude de simulation sur le risque attribuable $\varphi(t)$

Nous partons du plan de simulation de Chen *et al.* [5] afin de reproduire le tableau 1 de leur article. Ainsi, nous supposons un risque de base constant (distribution exponentielle) de paramètre $\lambda_0 = 0{,}01$. Nous avons généré 1 000 jeux d'observations de 200 et 500 individus indépendants chacun. Le paramètre β, dans le modèle à risques proportionnels, est fixé à $\ln(2)$ et 0, soit un RR égal à 2 et 1 respectivement et la probabilité d'exposition à 0,25 et 0,50. La durée maximale de l'étude τ est alors calculée de manière numérique afin d'avoir approximativement 10 % et 30 % de censure. Le tableau B.1 résume les valeurs de τ utilisées pour les différentes études de simulation.

λ_0	p	β	% censure	τ
0,01	0,25	0	10 %	999,954
0,01	0,25	0	30 %	319,705
0,01	0,50	0	10 %	999,954
0,01	0,50	0	30 %	319,705
1,00	0,25	0	10 %	9,999
1,00	0,25	0	30 %	3,197
1,00	0,50	0	10 %	9,999
1,00	0,50	0	30 %	3,197
0,01	0,25	ln(2)	10 %	874,881
0,01	0,25	ln(2)	30 %	275,614
0,01	0,50	ln(2)	10 %	749,722
0,01	0,50	ln(2)	30 %	232,994
1,00	0,25	ln(2)	10 %	8,748
1,00	0,25	ln(2)	30 %	2,756
1,00	0,50	ln(2)	10 %	7,497
1,00	0,50	ln(2)	30 %	2,329

Tableau B.1: Valeurs approchées de τ utilisées dans les plans de simulation

Nous avons calculé le RA, les biais, les écarts-types empiriques et les écarts-types estimés moyens en deux temps $\hat{t}_1$ et $\hat{t}_2$ correspondant à 75 % et 50 % de survie respectivement.

Pour obtenir ces temps, nous avons utilisé la fonction de survie marginale théorique définie selon le théorème des probabilités totales :

$$\begin{aligned} S(t) &= \mathbf{P}(T \geq t) \\ &= \mathbf{P}(T \geq t|Z=0)\mathbf{P}(Z=0) + \mathbf{P}(T \geq t|Z=1)\mathbf{P}(Z=1) \\ &= (1-p)S(t|Z=0) + pS(t|Z=1), \end{aligned}$$

soit, pour le modèle exponentiel de risque de base λ_0,

$$S(t) = (1-p)\exp(-\lambda_0 t) + p\exp\left[-\lambda_0 t \exp(\beta)\right].$$

Pour chaque jeu de paramètres $(p;\beta;\lambda_0)$, nous calculons les temps t_1 et t_2 tels que $S(t_1) = 0{,}75$ et $S(t_2) = 0{,}50$.

Soit s la valeur (connue) du percentile, $s = 0{,}75$ pour t_1 et $0{,}50$ pour t_2. Dans le cas où $\beta = 0$, nous avons :

$$s = (1-p)\exp(-\lambda_0 t) + p\exp\left[-\lambda_0 t\right] = \exp(-\lambda_0 t).$$

Les temps t_1 et t_2 sont donc obtenus à partir des expressions suivantes :

$$t_1 = -\frac{1}{\lambda_0}\ln(0{,}75) \text{ et } t_2 = -\frac{1}{\lambda_0}\ln(0{,}50).$$

Dans le cas où $\beta = \ln(2)$, nous avons :

$$s = (1-p)\exp(-\lambda_0 t) + p\exp(-2\lambda_0 t).$$

En posant $x = \exp(-\lambda_0 t)$, on obtient une équation du second degré à résoudre :

$$s = px^2 + (1-p)x \quad \text{ou} \quad px^2 + (1-p)X - s = 0.$$

Le discriminant $\Delta = (1-p)^2 + 4ps$ étant positif, cette équation admet deux solutions

$$x_1 = \frac{-(1-p) - \sqrt{(1-p)^2 + 4ps}}{2p} \quad \text{et} \quad x_2 = \frac{-(1-p) + \sqrt{(1-p)^2 + 4ps}}{2p}.$$

Comme $x_1 < 0$, l'unique solution de cette équation est x_2.

On en déduit les temps t_1 et t_2 correspondant aux percentiles $s = 0{,}75$ et $s = 0{,}50$ respectivement :

$$t_1 = -\frac{1}{\lambda_0}\ln\left[\frac{\sqrt{p^2+p+1}-(1-p)}{2p}\right] \text{ et } t_2 = -\frac{1}{\lambda_0}\ln\left[\frac{\sqrt{p^2+1}-(1-p)}{2p}\right].$$

Le tableau B.2 présente les valeurs des temps t_1 et t_2 obtenues de manière analytique. Dans nos simulations, les temps estimés $\hat{t}_k$ correspondent aux temps observés dans les jeux de données simulés les plus proches par la gauche des temps théoriques, c'est-à-dire :

$$\hat{t}_k = \arg\min_{\hat{t}_k < t_k} \left|\hat{t}_k - t_k\right| \text{ avec } k = 1, 2.$$

Une autre manière d'obtenir $\hat{t}_1$ et $\hat{t}_2$ est d'estimer la fonction de survie empirique, selon l'approche non paramétrique de Kaplan-Meier par exemple, et d'en déduire les percentiles empiriques :

$$\hat{t}_1 = \inf\{t; \hat{S}(t) \leq 0{,}75\} \text{ et } \hat{t}_2 = \inf\{t; \hat{S}(t) \leq 0{,}50\}.$$

B.3 Résultats de simulation pour le risque attribuable $\varphi(t)$

Les tableaux B.3 et B.4 présentent les résultats de simulations pour le risque attribuable φ en des temps t_1 et t_2 proches des percentiles de la fonction de survie marginale théorique dans le cas où la censure est uniforme sur $[0, \tau]$ et τ est obtenu par résolution numérique.

Dans le cas où $\beta = 0$ (Tableau B.3), de façon générale, les biais moyens sont faibles et les taux de recouvrement compris entre 93,2 % et 96,0 %. Les écarts-types empiriques et les écarts-type estimés moyens sont très proches, ce qui suggère la validité du calcul de la variance. Dans le cas d'un risque de base à 0,01 et lorsque la taille de l'échantillon passe de 200 à 500 observations avec la même probabilité d'exposition et le même pourcentage de censure, les biais et les écarts-types deviennent plus faibles, ce qui tend à resserrer

λ_0	p	β	t_1	t_2
0,01	0,25	0	23,40935	57,70495
0,01	0,50	0	19,49502	48,12118
0,01	0,25	ln(2)	23,40935	57,70495
0,01	0,50	ln(2)	19,49502	48,12118
1,00	0,25	0	0,23409	0,57705
1,00	0,50	0	0,19495	0,48121
1,00	0,25	ln(2)	0,23409	0,57705
1,00	0,50	ln(2)	0,19495	0,48121

Tableau B.2: Valeurs de t_1 et t_2 obtenues de manière analytique pour une survie de 75 % et 50 % respectivement

les intervalles de confiance et à diminuer les taux de recouvrement. Lorsque la probabilité d'exposition augmente (de 0,25 à 0,50), les écarts-types estimés moyens augmentent également. Les résultats sont similaires lorsque le risque de base est égal à 1 comparé au cas où le risque de base est égal à 0,01.

Dans le cas où $\beta = \ln(2)$ (Tableau B.4), nous retrouvons les mêmes conclusions que dans le cas où $\beta = 0$.

Les bons résultats obtenus en termes de probabilité de couverture nous laissent confiants sur l'estimateur proposé.

Nous avons néanmoins obtenu quelques différences avec les résultats publiés de Chen *et al.* [5] notamment dans le cas où $\beta = \ln(2)$ (5 lignes sur les 32 que compte leur tableau 1). En particulier, les écarts-types estimés moyens que nous avons obtenus diffèrent de ceux publiés par Chen *et al.* [5]. Nous avons contacté les auteurs pour comprendre les divergences observées mais n'avons pas obtenu de réponse.

				$t_1 : S(t_1) = 0{,}75$				$t_2 : S(t_2) = 0{,}50$			
λ_0	p	% de censure	n	Biais	SSD	SEE	PC	Biais	SSD	SEE	PC
0,01	0,25	10 %	200	0,001209	0,044022	0,043435	0,954	0,005078	0,044729	0,044098	0,933
0,01	0,25	10 %	500	0,000208	0,028156	0,027304	0,941	0,002857	0,028951	0,027592	0,944
0,01	0,25	30 %	200	0,001096	0,049389	0,049128	0,953	0,006324	0,051759	0,050926	0,947
0,01	0,25	30 %	500	0,000607	0,031384	0,030923	0,943	0,003908	0,032318	0,031454	0,944
0,01	0,50	10 %	200	0,000701	0,077446	0,075439	0,947	0,005618	0,078737	0,076626	0,948
0,01	0,50	10 %	500	0,001639	0,048797	0,047436	0,945	0,004324	0,049295	0,047855	0,939
0,01	0,50	30 %	200	0,000275	0,087420	0,085287	0,946	0,007658	0,090163	0,087767	0,944
0,01	0,50	30 %	500	0,002075	0,054062	0,053707	0,945	0,005335	0,054742	0,054286	0,942
1,00	0,25	10 %	200	0,001203	0,044020	0,043435	0,954	0,005085	0,044727	0,044097	0,933
1,00	0,25	10 %	500	0,000207	0,028156	0,027304	0,941	0,002856	0,028951	0,027592	0,944
1,00	0,25	30 %	200	0,001097	0,049393	0,049129	0,953	0,006325	0,051764	0,050924	0,947
1,00	0,25	30 %	500	0,000606	0,031384	0,030923	0,943	0,003906	0,032315	0,031453	0,944
1,00	0,50	10 %	200	0,000698	0,077450	0,075439	0,947	0,005623	0,078741	0,076626	0,948
1,00	0,50	10 %	500	0,001634	0,048797	0,047436	0,945	0,004320	0,049293	0,047854	0,939
1,00	0,50	30 %	200	0,000271	0,087415	0,085288	0,946	0,007661	0,090152	0,087766	0,944
1,00	0,50	30 %	500	0,002082	0,054064	0,053708	0,945	0,005341	0,054743	0,054287	0,942

Tableau B.3: Reproduction de la partie supérieure du tableau 1 page 522 [5] : Résultats de simulations pour la fonction de risque attribuable φ, en des temps observés t_1 et t_2 proches des temps théoriques, sous le modèle exponentiel à risques proportionnels $\lambda(t|Z) = \lambda_0 \exp(\beta Z)$ avec $\beta = 0$ et une censure uniforme $[0, \tau]$ où τ est obtenu par résolution numérique

				$t_1 : S(t_1) = 0{,}75$				$t_2 : S(t_2) = 0{,}50$			
λ_0	p	% de censure	n	Biais	SSD	SEE	PC	biais	SSD	SEE	PC
0,01	0,25	10 %	200	0,003153	0,046839	0,046964	0,949	0,000426	0,033216	0,033135	0,944
0,01	0,25	10 %	500	0,000307	0,030107	0,029905	0,948	0,000710	**0,021847**	**0,020979**	0,945
0,01	0,25	30 %	200	0,001089	0,049516	0,050691	0,957	0,002491	**0,036385**	**0,036449**	0,943
0,01	0,25	30 %	500	0,000782	0,031833	0,032153	0,944	0,001391	**0,023365**	**0,022979**	0,946
0,01	0,50	10 %	200	0,003724	0,064111	0,064489	0,946	0,002287	**0,054547**	**0,054475**	0,946
0,01	0,50	10 %	500	0,000429	0,041946	0,040805	0,938	0,000828	0,035427	0,034367	0,947
0,01	0,50	30 %	200	0,001195	0,072019	0,071744	0,941	0,000590	0,061192	0,060770	0,95
0,01	0,50	30 %	500	0,001011	0,046197	0,045325	0,940	0,001151	0,038926	0,038271	0,947
1,00	0,25	10 %	200	0,003156	0,046838	0,046964	0,949	0,000424	0,033216	0,033135	0,944
1,00	0,25	10 %	500	0,000306	0,030108	0,029904	0,948	0,000708	0,021848	0,020979	0,945
1,00	0,25	30 %	200	0,001085	0,049518	0,050691	0,957	0,002509	**0,036349**	**0,036448**	0,943
1,00	0,25	30 %	500	0,000778	0,031834	0,032153	0,943	0,001385	0,023374	0,022978	0,946
1,00	0,50	10 %	200	0,003723	0,064111	0,064489	0,946	0,002287	0,054547	0,054475	0,947
1,00	0,50	10 %	500	0,000431	0,041947	0,040805	0,938	0,000831	0,035430	0,034367	0,947
1,00	0,50	30 %	200	0,001188	0,071968	0,071747	0,941	0,000605	0,061155	0,060769	0,949
1,00	0,50	30 %	500	0,001025	0,046194	0,045329	0,940	0,001159	0,038924	0,038274	0,948

Tableau B.4: Reproduction de la partie inférieure du tableau 1 page 522 [5] : Résultats de simulations pour la fonction de risque attribuable φ, en des temps observés t_1 et t_2 proches des temps théoriques, sous le modèle exponentiel à risques proportionnels $\lambda(t|Z) = \lambda_0 \exp(\beta Z)$ avec $\beta = \ln(2)$ et une censure uniforme $[0, \tau]$ où τ est obtenu par résolution numérique

Les chiffres en gras correspondent aux résultats qui diffèrent de ceux des auteurs.

Annexe C

Variance de la survie attribuable

Un estimateur naturel de $AS(t)$ est :

$$\widehat{AS}(t) = \frac{\hat{S}_0(t) - \hat{S}(t)}{\hat{S}_0(t)}$$

où $\hat{S}(t)$ et $\hat{S}_0(t)$ sont respectivement les estimateurs de la fonction de survie globale dans la population et la fonction de survie chez les non exposés.

En nous inspirant du raisonnement de Chen *et al.* pour l'estimateur du RA, nous avons pu obtenir une formulation explicite de l'écart-type de $\widehat{AS}(t)$.

Ainsi, il est possible d'écrire :

$$\begin{aligned}
\widehat{AS}(t) - AS(t) &= \frac{\hat{S}_0(t) - \hat{S}(t)}{\hat{S}_0(t)} - \frac{S_0(t) - S(t)}{S_0(t)} \\
&= \frac{S_0(t)\left[\hat{S}_0(t) - \hat{S}(t)\right] - \hat{S}_0(t)\left[S_0(t) - S(t)\right]}{\hat{S}_0(t)S_0(t)} \\
&= \frac{-S_0(t)\hat{S}(t) + \hat{S}_0(t)S(t)}{\hat{S}_0(t)S_0(t)} \\
&= \frac{S(t)}{\hat{S}_0(t)S_0(t)}\left[\hat{S}_0(t) - S_0(t)\right] + \frac{S(t)S_0(t)}{\hat{S}_0(t)S_0(t)} - \frac{S_0(t)\hat{S}(t)}{\hat{S}_0(t)S_0(t)} \\
&= \frac{S(t)}{\hat{S}_0(t)S_0(t)}\left[\hat{S}_0(t) - S_0(t)\right] - \frac{1}{\hat{S}_0(t)}\left[\hat{S}(t) - S(t)\right]
\end{aligned}$$

Notons P_n la mesure empirique sur les données observées et P la distribution théorique. Chen *et al.* ont montré que $\sqrt{n}\{\hat{S}(t) - S(t)\}$ et $\sqrt{n}\{\hat{S}_0(t) - S_0(t)\}$ sont asymptotiquement équivalents à $\sqrt{n}\,(P_n - P)\,\eta_2(t)$ et $\sqrt{n}\,(P_n - P)\,\eta_1(t)$ respectivement où les fonctions $\eta_1(t)$

et $\eta_2(t)$ dépendent des méthodes d'estimation de $S_0(t)$ et $S(t)$ respectivement. Ainsi, sous certaines conditions de régularité, $\sqrt{n}\{\widehat{AS}(t) - AS(t)\}$ est asymptotiquement équivalent à

$$\sqrt{n}\,(P_n - P)\,\frac{1}{S_0(t)}\left\{\frac{S(t)}{S_0(t)}\eta_1(t) - \eta_2(t)\right\}$$

Par conséquent, $\sqrt{n}\{\widehat{AS}(t) - AS(t)\}$ converge faiblement vers un processus Gaussien de moyenne nulle et de matrice de variance-covariance $\mathbf{E}\{\xi(t)\xi^T(s)\}$ entre les temps t et s avec :

$$\xi(t) = \frac{1}{S_0(t)}\left\{\frac{S(t)}{S_0(t)}\eta_1(t) - \eta_2(t)\right\}$$

Un estimateur consistant de la fonction de variance-covariance serait alors

$$\hat{\sigma}_{AS}(s,t) = \frac{1}{n}\sum_{i=1}^{n}\hat{\xi}_i(t)\hat{\xi}_i(s)^T$$

où

$$\hat{\xi}_i(t) = \frac{1}{\hat{S}_0(t)}\left\{\frac{\hat{S}(t)}{\hat{S}_0(t)}\eta_{1i}(t) - \eta_{2i}(t)\right\}$$

et $\hat{\eta}_{1i}(t)$ et $\hat{\eta}_{2i}(t)$ sont les estimations de $\eta_1(t)$ et $\eta_2(t)$ respectivement pour le i-ème sujet $i = 1, ..., n$ au temps t.

Annexe D

Article en revue au journal *BMC Medical Research Methodology*

BMC Medical Research Methodology

Comparison of methods for estimating the attributable risk in the context of survival analysis

--Manuscript Draft--

Manuscript Number:	BMRM-D-16-00277R3
Full Title:	Comparison of methods for estimating the attributable risk in the context of survival analysis
Article Type:	Research article
Section/Category:	Data analysis, statistics and modelling
Funding Information:	Recipient of PhD grant from the French Ministry of Research — Dr Malamine Gassama Agence Nationale de Sécurité du Médicament et des Produits de Santé — Dr Anne C.M. Thiébaut
Abstract:	Background: The attributable risk (AR) measures the proportion of disease cases that can be attributed to an exposure in the population. Several definitions and estimation methods have been proposed for survival data. Methods: Using simulations, we compared four methods for estimating AR defined in terms of survival functions: two nonparametric methods based on Kaplan-Meier's estimator, one semiparametric based on Cox's model, and one parametric based on the piecewise constant hazards model, as well as one simpler method based on estimated exposure prevalence at baseline and Cox's model hazard ratio. We considered a fixed binary exposure with varying exposure probabilities and strengths of association, and generated event times from a proportional hazards model with constant or monotonic (decreasing or increasing) Weibull baseline hazard, as well as from a nonproportional hazards model. We simulated 1,000 independent samples of size 1,000 or 10,000. The five methods were compared in terms of mean bias, mean estimated standard error, empirical standard deviation and 95% confidence interval coverage probability at four equally spaced time points. Results: Under proportional hazards, all five methods yielded unbiased results regardless of sample size. Nonparametric methods displayed greater variability than other approaches. All methods showed satisfactory coverage except for nonparametric methods at the end of follow-up for a sample size of 1,000 especially. With nonproportional hazards, nonparametric methods yielded similar results to those under proportional hazards, whereas semiparametric, parametric and the simpler approaches that all relied on the proportional hazards assumption performed poorly. These methods were applied to estimate the AR of breast cancer due to menopausal hormone therapy (MHT) in 38,359 women of the E3N cohort. Conclusion: In practice, our study suggests to use the semiparametric or parametric approaches to estimate AR as a function of time in cohort studies if the proportional hazards assumption appears appropriate.
Corresponding Author:	Anne C.M. Thiébaut, PhD INSERM Montigny-le-Bretonneux, FRANCE
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	INSERM
Corresponding Author's Secondary Institution:	
First Author:	Malamine Gassama, PhD
First Author Secondary Information:	
Order of Authors:	Malamine Gassama, PhD Jacques Bénichou, MD, PhD Laureen Dartois, PhD

	Anne C.M. Thiébaut, PhD
Order of Authors Secondary Information:	
Response to Reviewers:	The authors' response letter has been included as a supplementary file

Comparison of methods for estimating the attributable risk in the context of survival analysis

Malamine Gassama, Biostatistics, Biomathematics, Pharmacoepidemiology and Infectious Diseases (B2PHI), Inserm, UVSQ, Institut Pasteur, Université Paris-Saclay, Paris, France
malamine.gassama@pasteur.fr

Jacques Bénichou, Inserm, U 1219, University of Rouen, France; Department of Biostatistics, Rouen University Hospital, Rouen, France
jacques.benichou@chu-rouen.fr

Laureen Dartois, CESP, Fac. de médecine - Univ. Paris-Sud, Fac. de médecine - UVSQ, INSERM, Université Paris-Saclay, Villejuif, France; Gustave Roussy, Villejuif, France
laureen.dartois@gustaveroussy.fr

Anne Thiébaut, Biostatistics, Biomathematics, Pharmacoepidemiology and Infectious Diseases (B2PHI), Inserm, UVSQ, Institut Pasteur, Université Paris-Saclay, Paris, France
anne.thiebaut@inserm.fr

Short title: Attributable risk estimation for survival data

Correspondence to:
Anne Thiébaut
Inserm, UMR 1181 (B2PHI)
Institut Pasteur
Bâtiment Laveran
25 rue du Docteur Roux
75724 Paris Cedex 15, France
Tel.: +33 (0)1 40 61 39 81
Fax: +33 (0)1 45 68 82 04
anne.thiebaut@inserm.fr

RESEARCH

Comparison of methods for estimating the attributable risk in the context of survival analysis

Malamine Gassama[1], Jacques Bénichou[2,3], Laureen Dartois[4,5] and Anne C.M. Thiébaut[1*]

*Correspondence:
anne.thiebaut@inserm.fr
[1] Biostatistics, Biomathematics, Pharmacoepidemiology and Infectious Diseases (B2PHI), Inserm, UVSQ, Institut Pasteur, Université Paris-Saclay, 25 rue du Dr. Roux, 75724, Paris Cedex 15, France
Full list of author information is available at the end of the article

Abstract

Background: The attributable risk (AR) measures the proportion of disease cases that can be attributed to an exposure in the population. Several definitions and estimation methods have been proposed for survival data.

Methods: Using simulations, we compared four methods for estimating AR defined in terms of survival functions: two nonparametric methods based on Kaplan-Meier's estimator, one semiparametric based on Cox's model, and one parametric based on the piecewise constant hazards model, as well as one simpler method based on estimated exposure prevalence at baseline and Cox's model hazard ratio. We considered a fixed binary exposure with varying exposure probabilities and strengths of association, and generated event times from a proportional hazards model with constant or monotonic (decreasing or increasing) Weibull baseline hazard, as well as from a nonproportional hazards model. We simulated 1,000 independent samples of size 1,000 or 10,000. The five methods were compared in terms of mean bias, mean estimated standard error, empirical standard deviation and 95% confidence interval coverage probability at four equally spaced time points.

Results: Under proportional hazards, all five methods yielded unbiased results regardless of sample size. Nonparametric methods displayed greater variability than other approaches. All methods showed satisfactory coverage except for nonparametric methods at the end of follow-up for a sample size of 1,000 especially. With nonproportional hazards, nonparametric methods yielded similar results to those under proportional hazards, whereas semiparametric, parametric and the simpler approaches that all relied on the proportional hazards assumption performed poorly. These methods were applied to estimate the AR of breast cancer due to menopausal hormone therapy (MHT) in 38,359 women of the E3N cohort.

Conclusion: In practice, our study suggests to use the semiparametric or parametric approaches to estimate AR as a function of time in cohort studies if the proportional hazards assumption appears appropriate.

Keywords: Attributable risk; Weighted Kaplan-Meier estimator; Piecewise constant hazards model; Cox model; Cohort studies; Breast cancer

Background

In epidemiology, it is important not only to assess the association between one exposure and the occurrence of health events, but also to quantify the impact of this exposure on the occurrence of these events. This is done by estimating the attributable risk (AR) or the proportion of cases associated with this exposure in the population. This estimation takes into account not only the strength of the link between exposure and disease but also the importance (prevalence) of exposure in the population [1]. It expresses the proportion of disease cases that can be attributed to exposure [2], that is to say, under certain conditions, the proportion of potentially preventable cases by eliminating exposure. The AR is defined as:

$$AR = \frac{\mathbf{P}(D) - \mathbf{P}(D|\bar{E})}{\mathbf{P}(D)}, \tag{1}$$

where $\mathbf{P}(D)$ is the probability of disease (incidence) in the population, which includes exposed E and unexposed $\bar{E}$ subjects, and $\mathbf{P}(D|\bar{E})$ is the hypothetical probability of disease in the same population but with all exposure eliminated.

The AR can be estimated from different types of studies including case-control studies for which many estimation methods exist (as reviewed in [3]), but it is rarely estimated from cohort studies. In the context of cohort studies and time-to-event outcomes, AR measures can be defined as functions of time [4-9] although a single AR estimate has been proposed alternatively [10].

Recent developments for estimating AR as a function of time from cohort studies in the survival analysis context have not so far led to a consensus definition. Several definitions have been proposed depending on whether authors interpret disease incidences $\mathbf{P}(D)$ and $\mathbf{P}(D|\bar{E})$ in equation (1) as cumulative distribution functions (CDFs) [6-9] or as instantaneous hazard functions [4, 5]. The two definitions converge only for rare diseases or low exposure prevalence [4]. Here we focus on the first definition of AR based on CDFs which looks more consistent with the standard

AR definition and appears to be the most used in the literature. Several methods of estimation have been proposed for the AR defined in this case, including non-parametric approaches based on Kaplan-Meier's estimator of the survival function [7], a semiparametric approach based on Cox's proportional hazards model [7] and a fully parametric approach assuming a piecewise constant hazards model [8]. Some evaluations were made for the nonparametric and semiparametric approaches [7] but, to the best of our knowledge, the performances of these various approaches have not been systematically compared.

The aim of this paper was to compare available methods for estimating AR when defined using CDFs. In the sections to follow, we first review the corresponding estimation methods so far published in the statistical literature. Simulations were conducted to assess the performance of the proposed AR estimators. The methods were then applied to data on menopausal hormone therapy (MHT) and breast cancer from the E3N women cohort (*Étude Épidémiologique auprès de Femmes de la Mutuelle Générale de l'Éducation Nationale*) [11]. For the purpose of our illustration, we considered 38,359 participants who were postmenopausal and free of cancer when they completed a self-administered questionnaire on their past use of any MHT in January 1992. In total, 17,185 (44.8%) women had ever used MHT at baseline and were considered exposed thereafter. By June 2008 (for a maximal 16.4 years and mean 14.0 years of follow-up), 2,228 invasive breast cancers had been diagnosed (1,106 in unexposed women). A recent work on the E3N cohort estimated a 14.5% postmenopausal breast cancer risk attributable to MHT use after 15 years of follow-up [12]. We estimated AR as a CDF-based function of time at four time points using nonparametric, semiparametric and parametric approaches, as well as the single overall AR measure proposed by Spiegelman *et al.* [10].

Methods

Review of estimation methods

When interpreting the incidence of disease $\mathbf{P}(D)$ as the event probability until some time t, the AR is defined as follows [4, 6, 7]:

$$A(t) = \frac{\mathbf{P}(T \leq t) - \mathbf{P}(T \leq t|Z^*)}{\mathbf{P}(T \leq t)}$$

where T denotes the time to disease or event time, Z a p-vector of risk factors and Z^* the p-vector of their chosen target values in order to quantify the potential impact of modifying the current distribution of Z to Z^*. Since, in most applications, Z^* is defined by setting one of the components of Z to its baseline (unexposed) level, we use notation $Z = 0$ instead of Z^* in the following. Using the survival functions $S(t) = \mathbf{P}(T > t)$ and $S_0(t) = S(T > t|Z = 0)$, the AR for time-to-event outcomes can be written as follows [7, 9]:

$$A(t) = \frac{S_0(t) - S(t)}{1 - S(t)} = 1 - \frac{1 - S_0(t)}{1 - S(t)}. \quad (2)$$

A natural estimate of $A(t)$ is obtained by replacing the survival functions $S_0(.)$ and $S(.)$ by their respective estimators $\hat{S}_0(.)$ and $\hat{S}(.)$. Various estimators $\hat{S}_0(.)$ and $\hat{S}(.)$ have been proposed, as detailed in the following subsubsections.

Nonparametric approaches

Chen *et al.* [7] considered several approaches for estimating survival functions $S_0(.)$ and $S(.)$ depending on covariate type: nonparametric when all p covariates are categorical and independent of time, otherwise semiparametric. The former case applies to a single categorical covariate or several covariates forming $K + 1$ categories.

When all p covariates are categorical and independent of time and under the assumption that censoring is independent of the covariates, Chen *et al.* [7] suggested estimating both $S_0(.)$ and $S(.)$ by the Kaplan-Meier method [13].

When all p covariates are categorical and independent of time but the assumption of covariate-independent censoring does not hold, Chen *et al.* [7] suggested estimating $S(.)$ by the weighed Kaplan-Meier (WKM) estimator [14] and $S_0(.)$ by the Kaplan-Meier method. For a p-vector Z of covariates with $K+1$ categories, the WKM estimator is defined as:

$$\hat{S}(t) = \frac{1}{n}\sum_{k=0}^{K} n_k \hat{S}_k(t)$$

where $\hat{S}_k(t)$ is the Kaplan-Meier estimator among those with covariate profile $k = 0, 1, 2, \ldots, K$ and n_k is the number of subjects with covariate profile k so that $\sum_{k=0}^{K} n_k$ equals n, the total number of subjects.

In all cases, the estimation of the variance of $\hat{A}(t)$ is based on the expression of $\{\hat{A}(t) - A(t)\}$ as a linear combination of $\{\hat{S}_0(t) - S_0(t)\}$ and $\{\hat{S}(t) - S(t)\}$ and relies on counting process results [7].

Semiparametric approach

For a more general type of covariates Z, i.e., when covariates are continuous, time-dependent or with too large a number of profile categories for nonparametric approaches, Chen *et al.* [7] considered using semiparametric instead of nonparametric methods to estimate $S_0(.)$ and $S(.)$. Of these, the Cox proportional hazards model [15] is one of the most familiar. It assumes that, at any time t, the hazard function $\lambda(t|Z)$ is the product of a nonparametric baseline hazard $\lambda_0(t)$ and a parametric function of the p-vector of covariates Z (or $Z(t)$ in the case of time-dependent covariates) and the p-vector of corresponding parameters β. The parametric function is usually taken to be the exponential function, such that $\lambda(t|Z) = \lambda_0(t)\exp(\beta^T Z)$.

In this case,

$$\hat{S}_0(t) = \exp\Big[-\hat{\Lambda}_0(t)\Big] \text{ and } \hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n}\exp\left[-\int_0^t \exp\{\hat{\beta}^T z_i(u)\}d\hat{\Lambda}_0(u)\right]$$

where $\hat{\Lambda}_0(.)$ is the Breslow estimator [16] of the baseline cumulative risk $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ and $\hat{\beta}$ is the maximum partial likelihood estimator.

The expression of the variance of $\hat{A}(t)$ follows the same general principles as for the nonparametric approaches above [7].

Parametric approach

Laaksonen *et al.* [8] proposed a parametric estimator based on a proportional hazards model with piecewise constant hazards (PCH). In this approach, follow-up time is partitioned into J prespecified intervals $(0 = a_0, a_1], (a_1, a_2], \ldots, (a_{j-1}, a_j], \ldots, (a_{J-1}, a_J]$, and the survival function at time t is estimated assuming a constant baseline hazard $\hat{\lambda}_{0j} = \exp(\hat{\alpha}_j)$ in each j-th interval $(a_{j-1}, a_j]$, $j = 1, 2, \ldots, J$ as follows:

$$\hat{S}_{PCH}(t|Z_i) = \exp\left\{-\sum_{j=1}^{J}\exp(\hat{\alpha}_j + \hat{\beta}^T Z_i)\delta_j(t)\right\}$$

where $\delta_j(t)$ defines the length of follow-up in the j-th interval:

$$\delta_j(t) = \begin{cases} 0 & \text{if } t \leq a_{j-1}, \\ t - a_{j-1} & \text{if } a_{j-1} < t \leq a_j, \\ a_j - a_{j-1} & \text{if } t > a_j. \end{cases}$$

The so-called population attributable fraction (PAF) estimator [8] is then defined using the following parametric estimators:

$$\hat{S}_0(t) = \frac{1}{n}\sum_{i=1}^{n} \hat{S}_{PCH}(t|Z_i = 0) \text{ and } \hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} \hat{S}_{PCH}(t|Z_i)$$

The model parameter estimates $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_J)$ and $\hat{\beta}$ are obtained by maximum likelihood estimation. The variance of $\hat{A}(t)$ is estimated using the delta method [8].

Global approaches over the whole follow-up period

Alternatively to the definition of the AR as a function of time, Spiegelman *et al.* [10] proposed to estimate a single value in cohort studies:

$$AR = \frac{\sum_{k=0}^{K} q_k(RR_k - 1)}{1 + \sum_{k=0}^{K} q_k(RR_k - 1)}$$

where RR_k and q_k, $k = 0, \ldots, K$, are the relative risk and prevalence in the target population for the kth combination of risk factors.

Upon using Cox's proportional hazards model, the overall AR can be estimated using estimated hazard ratio (HR) for relative risk and person-years for exposure prevalence in the cohort. The asymptotic variance is estimated using the multivariate delta method [10].

In the case of an unadjusted, bivariate exposure variable, the formula by Spiegelman *et al.* [10] simplifies into

$$AR = \frac{q(RR - 1)}{1 + q(RR - 1)} \tag{3}$$

where q denotes the exposure prevalence and RR the relative risk of exposed relative to nonexposed subjects. This formula resembles the well-known formula used by

epidemiologists [1, 2] where q is estimated by the proportion of exposed subjects at baseline (instead of exposed person-years over the whole follow-up).

Simulations

In this work, we considered a single, binary covariate Z representing exposure with $Z = 0$ and 1 for unexposed and exposed subjects respectively, simulated as a Bernoulli random variable with probability of exposure (q) set to 0.25, 0.50 and 0.75. To compare the different approaches for estimating AR, we considered either proportional or nonproportional hazards between the exposed and the unexposed.

For proportional hazards, we used instantaneous hazard functions of the form $\lambda(t|Z) = \lambda_0(t)\exp(\beta Z)$ where β denotes the regression parameter set to $\ln(2)$ or 0, and $\lambda_0(t)$ the baseline hazard function taken from a Weibull distribution with shape parameter γ and scale parameter θ: $\lambda_0(t) = \gamma\theta^{-\gamma}t^{\gamma-1}$, and generated event times from $(1/\theta)\left[-\ln(U)/\exp(\beta Z)\right]^{1/\gamma}$ with U an uniform on $(0,1)$. We explored situations where the baseline hazard was constant ($\gamma = 1$) or dependent on time, either increasing ($\gamma = 4/3$) or decreasing ($\gamma = 3/4$) with time. The scale parameter θ was chosen as a function of the shape parameter γ so as to obtain median survival time equal to 15 years for unexposed subjects in all scenarios. We calculated survival functions $S_0(.)$ and $S(.)$ as $\exp\{-(t/\theta)^\gamma\}$ and $(1-q)\exp\left[-(t/\theta)^\gamma\right] + q\exp\left[-(t/\theta)^\gamma\exp(\beta)\right]$ respectively and derived theoretical values of AR as a function of time from equation (2). For the global AR derived from the simpler approach, theoretical values were obtained as $q[\exp(\beta)-1]/\{1+q[\exp(\beta)-1]\}$.

For nonproportional hazards, we generated event times from $G^{-1}[-\ln(U)]/[\lambda_0\exp(\beta Z)]$ assuming a cumulative hazard function of the form $\Lambda(t|Z) = G\left[\lambda_0 t\exp(\beta Z)\right]$ where G is the logarithmic transformation $G(t) = \ln(1+2t)/2$ [7]. Setting $\lambda_0 = 0.1$ yielded a median survival time for unexposed subjects equal to 15 years as in the proportional hazards case. Setting the regression coefficient β to $\ln(2)$, the hazard ratio between the exposed and the unexposed decreased from 2 toward 1 over

time. We calculated survival functions $S_0(.)$ and $S(.)$ as $\exp\{-\ln(1+2\lambda_0 t)/2\}$ and $(1-q)\exp\{-\ln(1+2\lambda_0 t)/2\}+q\exp\{-\ln(1+2\lambda_0 t\exp(\beta))/2\}$ respectively and derived theoretical values of AR as a function of time from equation (2).

We generated censoring times independent of the covariate Z and event times T from a uniform distribution on $[0,\tau]$, with τ the maximal follow-up time of the study set at 20 years. Depending on scenarios, we obtained censoring percentages around 47-68% (ranges across simulations from 42% up to 73%).

We generated 1,000 data sets of n = 1,000 or 10,000 independent observations and calculated estimators $\hat{A}(.)$ of the AR as a function of time and their associated variances using the four approaches: two non-parametric approaches corresponding to the case where $S_0(.)$ and $S(.)$ are both estimated by the Kaplan-Meier method (KM) and to the case where $S_0(.)$ and $S(.)$ are estimated by the Kaplan-Meier and the weighted Kaplan-Meier methods, respectively (WKM) [7], one semiparametric approach using Cox's proportional hazards model (COX) [7], and one parametric approach corresponding to the case where survival functions are estimated assuming piecewise constant hazards (PCH) [8] considering four intervals of 5-year width. In the case where no event was generated in any five-year interval, the simulated dataset was discarded and replaced by a new one. We also considered the simpler approach based on equation (3) to estimate a global AR.

Results of the time-dependent approaches are presented at times $t=\tau/4$, $\tau/2$, $3\tau/4$ and τ (respectively, 5, 10, 15 and 20 years). For the nonparametric and semiparametric approaches, estimates were obtained at times actually observed in the dataset so we considered values taken at the closest preceding time point. While nonparametric estimations are based on data available until the time of interest, semiparametric and parametric methods use data available over the whole follow-up period. To allow for a fairer comparison under the proportional hazards assumption, we also computed semiparametric and parametric estimators after censoring

observation times at either $\tau/4$ or $\tau/2$. The parametric approach was then based on one or two interval(s) of 5-year width respectively.

For all five approaches, results displayed are the average absolute bias relative to the theoretical value $A(.)$, the Sampling Standard Deviation of $\hat{A}(.)$ (SSD), the average Standard Error Estimator of $A(.)$ (SEE) and the coverage probability (CP) of the 95% confidence interval (95%CI) of $A(.)$. Although authors [7, 8] have suggested to use the complementary logarithmic transformation $\ln\{1 - A(.)\}$ to improve coverage probabilities in case of small sample size, this did not notably improve coverage probabilities in our results (data not shown) so results presented are for the untransformed $A(.)$.

Simulations were performed using R release 3.0.1. We coded the nonparametric methods using R software and tested the validity of our code by comparing our simulation results with those of the authors using the same parameters [7]. For the semiparametric method [7], we used the R package `paf` developed by Chen [17]. For the parametric method [8], we used SAS release 9.3 and a set of macros developed by Laaksonen *et al.* [18]. For the global approach by Spiegelman *et al.* [10], we used the `%par` SAS macro developed by the authors.

Results

Simulations

We first considered the case of proportional hazards between the exposed and the unexposed with $\beta = \ln(2)$ and probability of exposure equal to 0.50, starting with a constant baseline hazard. With a sample size of 1,000 observations and for the four time-dependent approaches (Table 1, left-hand side), there was more upward bias at the end of follow-up τ, especially with the KM method and the WKM method (to a lesser extent), but AR estimators for all methods and time points were virtually unbiased (relative bias $< 2.5\%$). Variance estimators accurately reflected the true variation and the 95%CIs had proper coverage probabilities, except in τ for the

two nonparametric methods, where the variance was somewhat underestimated, yielding lower than nominal coverage. Parametric and semiparametric estimators were more precise than nonparametric estimators, particularly at times $\tau/4$ and τ. Estimators of parameter β were unbiased for the semiparametric and parametric approaches (relative bias $< 0.7\%$, data not shown).

When considering samples of size 10,000 (Table 1, right hand side), with approaches giving an estimate of the AR function of time, bias decreased in magnitude compared to a sample size of 1,000 observations (relative bias $< 0.7\%$ for AR in all methods and $< 0.04\%$ for β in the semiparametric and parametric approaches). As expected, precision increased markedly for all methods, by a factor of about $\sqrt{10}$. Moreover, SEEs and SSDs were in closer agreement even at time τ with nonparametric methods and all coverage probabilities fell within the 0.940 to 0.960 range.

Similar observations held when considering a decreasing baseline hazard (Table 2). When $\gamma = 3/4$, biases were close to those observed with $\gamma = 1$ except for a moderate increase for the parametric approach and both sample sizes (relative bias $< 2.4\%$). Nevertheless coverage probabilities remained satisfactory for this method and the others, except again at the end of follow-up τ for the two nonparametric methods and $n = 1{,}000$ (0.915 and 0.906 for KM and WKM respectively).

Under an increasing baseline hazard (Table 3), coverage probabilities at τ of the two nonparametric estimators worsened with $n = 1{,}000$ (0.891 and 0.898 for KM and WKM approaches respectively) as a result of increased biases compared to constant and decreasing baseline hazards. Results were otherwise satisfactory and biases for the parametric method were comparable with those obtained under constant baseline hazard.

With a lower or greater prevalence of exposure (25% or 75% exposed), coverage probabilities in τ for the nonparametric approaches improved but sometimes remained lower than the nominal value despite a sample size of 10,000 (Supple-

mentary Tables A1 and A2 for $\gamma = 1$ and $\beta = \ln(2)$). The same general picture held with other values of γ (data not shown), except for the parametric approach which showed slightly insufficient (93%) coverage at times $< \tau$ for $\gamma = 3/4$ and both exposure probabilities 0.25 and 0.75.

Under the same parameters but $\beta = 0$ (Supplementary Table A3 for $\gamma = 1$ and 50% exposed), results were similar to those with $\beta = \ln(2)$ except for slightly improved coverage probabilities in τ for the nonparametric approaches and a sample size of 1,000.

Under all scenarios with proportional hazards (Tables 1, 2, 3, A1, A2 and A3), estimators of global AR from the simpler approach were virtually unbiased with satisfactory coverage probabilities. The estimated single values were generally greater than those of time-dependent approaches at any point in time.

When follow-up was stopped at $\tau/4$ or $\tau/2$, under proportional hazards (data not shown), estimates for the two nonparametric methods were of course identical to those obtained at the same time points with a complete follow-up. SEEs for the semiparametric and parametric methods increased, getting closer to those of nonparametric methods with censoring at $\tau/2$ and even closer with censoring at $\tau/4$. Coverage probabilities remained satisfactory except for the parametric method under decreasing baseline hazard ($\gamma = 3/4$) where they tended to be lower than the nominal value e.g., 0.935 and 0.918 at $\tau/4$ for censoring at $\tau/4$, $\beta = \ln(2)$ and 50% exposed, and for $n = 1{,}000$ and $n = 10{,}000$ respectively.

Finally, when considering nonproportional hazards between the exposed and the unexposed (Table 4, for $\beta = \ln(2)$ and 50% exposed), nonparametric methods yielded similar results to those under proportional hazards. However, the semiparametric and parametric approaches that both relied on the proportional hazards assumption performed poorly. With a sample size of 1,000 observations (Table 4, left-hand side), estimates using the semiparametric approach were biased (relative bias between 7.9 and 32.6%) with poor coverage probabilities except at $\tau/2$. The

parametric approach resulted in even more severe biases (relative bias between 14.6 and 81.6%) and poorer coverage probabilities. With $n = 10{,}000$, bias remained high and became similar with the semiparametric and parametric approaches (between 7.1 and 31.2% and between 8.3 and 32% respectively), and coverage deteriorated further as a result of tighter 95%CIs (Table 4, right hand side). With a lower or greater prevalence of exposure, coverage probabilities with the parametric approach improved at all times but generally remained less than 93% (data not shown).

Data example

As in our simulations, we used time-on-study rather than attained age as the time-scale after checking that both yielded similar results. Fitting a Weibull distribution to the observed survival data and considering incident invasive breast cancer as the event of interest (i.e., considering time to breast cancer occurrence), the shape (γ) and scale (θ) parameters were estimated as 1.2 and 178.2 respectively and the corresponding estimated Weibull survival function almost coincided with nonparametric Kaplan-Meier estimate (data not shown). The assumption of proportional hazards between women ever-exposed and those never-exposed to any MHT at baseline seemed appropriate (Schoenfeld residual test, $p = 0.7$), with an estimated hazard ratio (HR) at 1.22 (95%CI, 1.13 to 1.33) for MHT exposure from the Cox model.

The AR estimates from nonparametric approaches KM and WKM were almost identical (Figure 1). They tended to increase until 12 years of follow-up (e.g., for the KM approach, from 5.5% (95%CI, −2.7 to 13.6%) after four years to 12.0% (95%CI, 7.8 to 16.2%) after 12 years of follow-up), then to decrease and converge to semiparametric and parametric estimates at the end of follow-up with an estimated 9.2% AR (95%CI, 5.4 to 13.0%) after 16 years. In comparison, estimates using the semiparametric and parametric approaches slightly decreased monotonically over time from 9.0% (95%CI, 5.3 to 12.8%) to 8.8% (95%CI, 5.1 to 12.4%) and from 8.9% (95%CI, 5.2 to 12.6%) to 8.7% (95%CI, 5.0 to 12.3%) respectively. Thus, after

16 years of follow-up, the proportion of invasive breast cancer cases attributable to MHT exposure was close to 9% whatever the method used. Estimates using non parametric approaches were far less precise at earlier times and displayed wider 95%CIs (even including 0 at time 4 years) than semiparametric and parametric approaches in the first half of the follow-up: e.g., at time 8 years, AR was estimated as 8.9% (95%CI, 3.5 to 14.4%) and 9.0% (95%CI, 5.2 to 12.7%) from the KM and Cox approaches, respectively. Adjusting for age at baseline, either as a continuous covariate in the semiparametric approach or as a dichotomous covariate in all approaches, hardly modified these estimates (data not shown).

Finally, using the method proposed by Spiegelman *et al.* [10], we found that 9.2% (95%CI, 5.4 to 13.0%) of cases who developed invasive breast cancer at various times in the cohort follow-up were attributable to MHT exposure. Using the simpler approach with the proportion of exposed subjects at inclusion, we obtained a close, slightly smaller estimate at 9.1% (95%CI, 5.3 to 12.8%).

Discussion

Comparing different methods of AR estimation when disease incidence is interpreted as a CDF [7, 8], we observed that AR estimators were essentially unbiased for all approaches when we generated event times from a proportional hazards model. Empirical and estimated variances were close, with proper coverage probabilities except at the end of follow-up for the nonparametric methods and a smaller sample size. When considering a non-constant baseline hazard, estimates using the parametric approach were robust despite misspecification of the baseline hazard. For nonparametric approaches, biases tended to increase at the end of follow-up (time τ) when the baseline hazard increased with time. With the simpler approach, results were satisfactory. However, under nonproportional hazards, estimates using the semiparametric and parametric approaches were biased with poor coverage probabilities.

To our knowledge, this is the first simulation study comparing nonparametric, semiparametric, parametric methods of AR estimation as a function of time as well as a simpler, gobal approach for a diversity of scenarios (proportional or nonproportional hazards, constant or nonconstant baseline hazard, varying exposure probabilities, strengths of association and sample sizes) in the survival analysis context. Chen *et al.* [7] reported simulations for the Kaplan-Meier, weighted Kaplan-Meier and transformation models when event times were generated from proportional or nonproportional hazards models with regression parameter $\beta = 1$, 40% probability of exposure and a sample size of 1,000 observations. Like them, we found that, under the assumption of independent censoring, results with KM and WKM approaches were very close. Differences between the two nonparametric approaches were apparent when censoring was dependent on covariates [7], which we did not evaluate in this study.

Also in line with Chen *et al.* [7], when we generated event times from a proportional hazards model, we found that nonparametric and semiparametric estimates were all unbiased, nonparametric estimates had larger variances than semiparametric estimates and estimated variances accurately reflected the true variance except in τ for the nonparametric approaches and a sample size of 1,000 observations. Nonparametric approaches tended to perform better (respectively worse) when exposure prevalence was lower (respectively higher) which could be expected from the possibly unstable and inefficient Kaplan-Meier estimator of survival among the unexposed when the proportion of those is small [7]. This general picture held in our simulations whether event times were generated with constant, decreasing or increasing baseline hazard. We note, however, that, when we considered a larger sample size, the discrepancies between estimated and empirical variances tended to diminish, with most often satisfactory coverage probabilities in τ.

For nonproportional hazards, we generated event times using a transformation model considered by Chen *et al.* [7] and found consistent results for the nonpara-

metric approaches, similar to those in the case of proportional hazards. However, while Chen *et al.* [7] applied the same non proportional hazards model for both data generation and analysis (AR estimation), we generated data under non proportional hazards and estimated AR from (misspecified) Cox's proportional hazards model. This explains the impaired performance we observed when the proportional hazards assumption was violated in contrast with the satisfactory results obtained by Chen *et al.* [7]. Sjölander and Vansteelandt [9] recently proposed an alternative semiparametric estimator of AR also based on Cox's proportional hazards model that proved robust to various model misspecifications. However these authors did not evaluate deviations from the proportional hazards assumption.

Like Chen *et al.* [7] in their simulation and example analysis, we observed greater imprecision of the nonparametric estimators at the start of follow-up, which could explain possible early negative AR values. This imprecision could be expected because the estimation of the survival function relies upon the information available until the time of interest and not many events have yet occurred by then. This differs from the semiparametric and parametric methods which take advantage of the estimation of parameter β being performed over the entire follow-up. Consistently, we found larger variances for the semiparametric and parametric approaches with shorter lengths of follow-up.

Another novelty of this work was the evaluation of the parametric approach to AR estimation proposed by Laaksonen *et al.* [8] using simulations and its comparison with nonparametric and semiparametric approaches. Generally under proportional hazards, we found close agreement between the semiparametric and parametric approaches, in our simulations as well as in the example. Of note, the parametric approach seemed robust despite misspecification of baseline hazard, *i.e.*, when we considered decreasing or increasing (instead of constant) baseline hazard and proportional hazards. However, like the semiparametric approach based on Cox's model, the parametric approach was sensitive to the proportional hazards assump-

tion and performed poorly in our simulations when this assumption was violated. We also evaluated the simpler, global approach and our results were satisfactory under proportional hazards.

As noted by several authors [4, 7], simpler approaches based on equation (1) or equivalent formulas [1, 2] are generally defined for binary outcomes with time-independent risk factors. Consequently, they prove to be inadequate for cohort studies with censored time-to-event outcomes and possibly time-dependent covariates. In contrast, the nonparametric, semiparametric and parametric approaches we considered here have been specifically developped for censored time-to-event outcomes and produce AR estimate as a function of time, thus allowing the AR to be time-varying. A major limitation of the simpler approach in the context of cohort studies is that it only takes account of the proportion of exposed subjects at the beginning of follow-up. The proportion of exposed subjects indeed decreases as follow-up time increases (because exposed subjects fail earlier than nonexposed subjects) [6]. This explains why our AR estimates from the simpler approach were generally greater than those from time-dependent approaches and further underlines why approaches estimating AR as a function of time are an improvement on the simpler approach in the context of survival analysis.

In our study, we used the definition of AR based on CDFs because it is a natural extension of the standard AR definition (equation 1) for time-to-event outcomes [6-9] and it is equivalent to the standard definition when time t is the end of follow-up in cohort studies [4]. In addition several estimation methods have been proposed for the CDF-based AR definition in cohort studies and the survival analysis context in contrast to the alternative definition based on instantaneous hazard functions [4, 5] for which only one method of estimation based on Cox's proportional hazards model has been published [4].

In cohort studies where exposed individuals are over-sampled relative to the exposure prevalence in the population, AR will correctly reflect the impact of exposure

in the cohort, but the impact at the population level will be overestimated. The marginal survival function $S(t)$ should be corrected in order to alleviate this upward bias on AR estimates. The AR (and its estimates) being a function of time, various representations of AR estimates can be used. We used a graphical representation of the whole time function in our example and produced estimates at four equally spaced times in our simulations. Alternatively, a single overall estimate could be obtained by averaging out the time function of AR estimates or by using the alternative approach by Spiegelman *et al.* [10] as in our example.

We chose our simulation parameters to resemble real epidemiologic cohorts. These often include a few thousands participants followed for several years. For a smaller sample size ($n = 500$), whether we used the logarithmic transformation or not, we observed findings generally similar to those presented with a sample size of 1,000 observations. This was true with the notable exception of the less than nominal coverage probability for the semiparametric approach at time τ for constant ($\gamma = 1$) and decreasing ($\gamma = 3/4$) baseline hazards, and at times $\tau/4$ and τ for increasing ($\gamma = 4/3$) baseline hazard (data not shown).

In our application, the proportional hazards assumption seemed appropriate, as well as a Weibull distribution for event times with an increasing baseline hazard and shape parameter halfway between the values $\gamma = 1$ and $4/3$ considered in our simulation study. Exposure frequency was also close to our simulated 0.5 probability of exposure. However, as in many epidemiologic cohorts, the censoring rate was much greater in our example (94.2%) than in our simulations. The resulting imprecision may explain the nonparametric AR estimates apparently increasing until three quarters of total follow-up but compatible with the more expected decreasing trend. Chen *et al.* [7] observed the same finding in their application on a shorter length of follow-up.

Using the approach described by Spiegelman *et al.* [10], the overall AR estimate for ever use of MHT at baseline was 9.2% in the E3N cohort. This estimate was close to

those obtained at the end of follow-up with the nonparametric methods and at the start of follow-up with the parametric and semiparametric approaches. In a recent publication, Dartois *et al.* [12] reported a higher AR estimate of 14.5% (95%CI, 9.2 to 19.6%) for recent MHT use and postmenopausal invasive breast cancer from the E3N cohort data, using the approach proposed by Spiegelman *et al.* [10] and a more refined, adjusted analysis with MHT exposure as a time-dependent covariate.

This study has some limitations. First, we did not evaluate AR estimates adjusted for covariates. Adjustment for multiple variables is common practice in epidemiology, especially age which can also be used as the underlying time-variable [19]. In our example, using analyses unadjusted or parametrically adjusted for age, there was a statistically significant association between baseline MHT ever use and breast cancer risk, in line with findings from more complex models with age as the timescale and adjustment for other covariates in the original study [11]. Although adjustment for covariates is available in packages for semiparametric and parametric approaches [7, 8], there are constraints in nonparametric approaches as the number of covariates must be limited and adjustment for continuous variables is not possible. Moreover, available packages for estimating the AR would need to be adapted to allow left truncation resulting from using age as the timescale. Second, in our example, we only considered women who had ever received MHT at baseline as exposed whereas exposure can vary during follow-up. Other methodological studies are needed to take into account the exposure time dependency for estimating AR as a function of time [9]. Finally, we have ignored the competing risk of death and cancers of other sites (11.2% of our 94.2% censored observations) which might also bias our estimate of breast cancer risk attributable to MHT [20].

Conclusions

The AR estimators from the four methods had satisfactory performance under the proportional hazards assumption. Estimators using semiparametric and parametric

approaches were not robust in case of nonproportional hazards. Lack of precision could be an issue for nonparametric methods at the beginning of the follow-up time in cohorts of relatively low sample size. In practice, if the proportional hazards assumption seems appropriate, the semiparametric or parametric approaches should be used.

List of abbreviations

AR: Attributable Risk

CDF: Cumulative Distribution Function

CI: Confidence Interval

COX: Cox's proportional hazards model

CP: Coverage Probability

E3N: *Étude Épidémiologique auprès de Femmes de la Mutuelle Générale de l'Éducation Nationale*

HR: Hazard Ratio

KM: Kaplan-Meier

MHT: Menopausal Hormone Therapy

PCH: Piecewise Constant Hazards

RR: Relative Risk

SEE: Standard Error Estimator

SSD: Sampling Standard Deviation

WKM: Weighted Kaplan-Meier

Declarations

Ethics approval and consent to participate

The E3N cohort received ethical approval from the French National Commission for Computed Data and Individual Freedom ('Commission Nationale de l'Informatique et des Libertés', CNIL) under the reference CNIL 186 and all participants in the study provided informed consent.

Consent to publish

Not applicable.

Availability of data and materials

The E3N dataset analyzed during the current study was available from the E3N study team but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the E3N study team upon reasonable request and permission of E3N principal investigator.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by the French Medicines Agency ('Agence Nationale de Sécurité du Médicament et des produits de santé', ANSM). MG is a recipient of PhD grant from the French Ministry of Research.

Authors' contributions

MG conducted the simulation study and prepared the first draft of the manuscript. JB and ACMT devised the analytical strategy and co-drafted the manuscript. LD contributed to cohort data analysis. All authors contributed to the preparation of the final manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors wish to thank Pascale Tubert-Bitter for her constructive comments, Mohammed Sedki for statistical advice and Agnès Fournier for kindly sharing the E3N data. The authors are also grateful to all participants, practitioners and study staff of the E3N study. The E3N cohort is conducted with the financial support of 'Mutuelle Générale de l'Éducation Nationale' (MGEN); the European Community; 'Ligue nationale contre le Cancer'; 'Institut Gustave-Roussy'; 'Institut National de la Santé et de la Recherche Médicale' (Inserm); and 'Fondation de France'.

Author details

[1] Biostatistics, Biomathematics, Pharmacoepidemiology and Infectious Diseases (B2PHI), Inserm, UVSQ, Institut Pasteur, Université Paris-Saclay, 25 rue du Dr. Roux, 75724, Paris Cedex 15, France. [2] Inserm, U 1219, University of Rouen, 1 rue de Germont, 76031, Rouen Cedex, France. [3]Department of Biostatistics, Rouen University Hospital, 1 rue de Germont, 76031, Rouen Cedex, France. [4]CESP, Fac. de médecine - Univ. Paris-Sud, Fac. de médecine - UVSQ, INSERM, Université Paris-Saclay, 114 rue Edouard Vaillant, 94805, Villejuif Cedex, France. [5]Gustave Roussy, 114 rue Edouard Vaillant, 94805, Villejuif Cedex, France.

References

1. Cole P, MacMahon B. Attributable risk percent in case-control studies. *British Journal of Preventive Social Medicine* 1971; 25(4): 242–244.
2. Levin ML. The occurrence of lung cancer in man. *Acta - Unio Internationalis Contra Cancrum* 1953; 9(3): 531–541.
3. Bénichou J. A review of adjusted estimators of attributable risk. *Statistical Methods in Medical Research* 2001; 10(3): 195–216.
4. Chen YQ, Hu C, Wang Y. Attributable risk function in the proportional hazards model for censored time-to-event. *Biostatistics* 2006; 7(4): 515–529. DOI: 10.1093/biostatistics/kxj023.
5. Samuelsen SO, Eide GE. Attributable fractions with survival data. *Statistics in Medicine* 2008; 27(9): 1447–1467. DOI: 10.1002/sim.3022.
6. Cox C, Chu H, Muñoz A. Survival attributable to an exposure. *Statistics in Medicine* 2009; 28(26): 3276–3293. DOI: 10.1002/sim.3705.
7. Chen L, Lin DY, Zeng D. Attributable fraction functions for censored event times. *Biometrika* 2010; 97(3): 713–726. DOI: 10.1093/biomet/asq023.
8. Laaksonen MA, Knekt P, Härkänen T, Virtala E, Oja H. Estimation of the population attributable fraction for mortality in a cohort study using a piecewise constant hazards model. *American Journal of Epidemiology* 2010; 171(7): 837–847. DOI: 10.1093/aje/kwp457.
9. Sjölander A, Vansteelandt S. Doubly robust estimation of attributable fractions in survival analysis. *Statistical Methods in Medical Research* 2014 (in press). DOI: 10.1177/0962280214564003.
10. Spiegelman D, Hertzmark E, Wand HC. Point and interval estimates of partial population attributable risks in cohort studies: examples and software. *Cancer Causes & Control* 2008; 18(5): 571–579. DOI: 10.1007/s10552-006-0090-y.

11. Fournier A, Mesrine S, Dossus L, Boutron-Ruault MC, Clavel-Chapelon F, Chabbert-Buffet N. Risk of breast cancer after stopping menopausal hormone therapy in the E3N cohort. *Breast Cancer Research and Treatment* 2014; 145(2): 535–543. DOI: 10.1007/s10549-014-2934-6.
12. Dartois L, Fagherazzi G, Baglietto L, Boutron-Ruault MC, Delaloge S, Mesrine S, Clavel-Chapelon F. Proportion of premenopausal and postmenopausal breast cancers attributable to known risk factors: Estimates from the E3N-EPIC cohort. *International Journal of Cancer* 2016; 138(10):2415–27 DOI: 10.1002/ijc.29987.
13. Kaplan EL, Meier P. Non parametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; 53(282): 457–481.
14. Murray S, Tsiatis AA. Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics* 1996; 52(1): 137–151.
15. Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society Series B* 1972; 34(2): 187–220.
16. Breslow, N. E. Discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society Series B* 1972; 34(2): 216–217.
17. Chen L. 'paf': Attributable fraction function for censored survival data, R package version 1.0, 2014. http://cran.r-project.org/web/packages/paf/index.html [accessed 2 June 2014]
18. Laaksonen MA, Virtala E, Knekt P, Oja H, Härkänen T. SAS macros for calculation of population attributable fraction in a cohort study design. *Journal of Statistical Software* 2011; 43(7): 1–25.
19. Thiébaut ACM, Bénichou J. Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Statistics in Medicine* 2004; 23(24): 3803–3820.
20. Laaksonen MA, Härkänen T, Knekt P, Virtala E, Oja H. Estimation of population attributable fraction (PAF) for disease occurrence in a cohort study design. *Statistics in Medicine* 2010; 29(7-8): 860–874. DOI: 10.1002/sim.3792.

Figures

Figure 1 Estimation of the risk of invasive breast cancer attributable to ever use of menopausal hormone therapy at baseline as a time function, E3N cohort, 1992-2008. The dark solid and dark dashed curves pertain to the point estimates by KM and COX, respectively, the dark circles to the point estimates by the 4-year interval PCH; the light solid and light dashed curves, as well as the light circles, show the corresponding 95% confidence intervals. The WKM curves are not displayed because they almost coincided with the KM curves at the chosen scale.

Tables

Additional Files

Table 1 Simulation results for the estimation of attributable risk $A(.)$ under proportional hazards, constant baseline hazard ($\gamma = 1$) with regression parameter $\beta = \ln(2)$ and probability of exposure $q = 0.5$

Estimation method	Time	$A(t)$	$n = 1{,}000$ Bias	SEE	SSD	CP	$n = 10{,}000$ Bias	SEE	SSD	CP
KM	$\tau/4$	0.284	0.001584	0.052440	0.052591	0.949	−0.000011	0.016622	0.016349	0.944
	$\tau/2$	0.240	0.001496	0.039210	0.039099	0.948	0.000235	0.012434	0.012420	0.944
	$3\tau/4$	0.200	0.001100	0.035666	0.035948	0.946	−0.000333	0.011353	0.011354	0.949
	τ	0.166	0.004047	0.043238	0.053015	0.912	0.001025	0.017251	0.019598	0.943
WKM	$\tau/4$	0.284	0.001594	0.052516	0.052483	0.949	0.000003	0.016613	0.016357	0.946
	$\tau/2$	0.240	0.001541	0.039144	0.038926	0.950	0.000285	0.012401	0.012398	0.946
	$3\tau/4$	0.200	0.001093	0.035402	0.035479	0.953	−0.000286	0.011283	0.011297	0.952
	τ	0.166	0.002922	0.040635	0.048602	0.902	0.000497	0.016646	0.018245	0.942
COX	$\tau/4$	0.284	0.000977	0.038843	0.038208	0.958	−0.000136	0.012292	0.012206	0.956
	$\tau/2$	0.240	0.001108	0.033847	0.033524	0.951	0.000006	0.010700	0.010616	0.958
	$3\tau/4$	0.200	0.001031	0.029264	0.028893	0.958	−0.000081	0.009237	0.009253	0.954
	τ	0.166	0.002577	0.027146	0.027753	0.946	0.000148	0.008965	0.009087	0.950
PCH	$\tau/4$	0.284	0.001356	0.038338	0.038248	0.952	−0.000086	0.012120	0.012209	0.953
	$\tau/2$	0.240	0.001372	0.033380	0.033529	0.948	0.000034	0.010543	0.010608	0.952
	$3\tau/4$	0.200	0.001113	0.028804	0.028870	0.957	−0.000081	0.009088	0.009263	0.952
	τ	0.166	0.001564	0.025811	0.025420	0.961	−0.000154	0.008105	0.008153	0.952
Simpler	–	0.333	0.000826	0.043356	0.043147	0.952	−0.000209	0.013715	0.013776	0.955

KM: nonparametric approach based on Kaplan-Meier estimation for $S(t)$; WKM: nonparametric approach based on weighted Kaplan-Meier estimation for $S(t)$; COX: semiparametric approach; PCH: parametric approach using a piecewise constant hazards model; Simpler: simpler approach based on proportion of exposed subjects; Bias: sampling mean of the difference between $\hat{A}(t)$ and $A(t)$; SEE: sampling mean of standard error estimate of $A(t)$; SSD: sampling standard deviation of $\hat{A}(t)$; CP: coverage probability of the 95% Wald confidence interval.

Table 2 Simulation results for the estimation of attributable risk $A(.)$ under proportional hazards, decreasing baseline hazard ($\gamma = 3/4$) with regression parameter $\beta = \ln(2)$ and probability of exposure $q = 0.5$

Estimation method	Time	$A(t)$	$n = 1{,}000$ Bias	SEE	SSD	CP	$n = 10{,}000$ Bias	SEE	SSD	CP
KM	$\tau/4$	0.269	0.001799	0.044659	0.045486	0.940	0.000129	0.014162	0.014200	0.946
	$\tau/2$	0.231	0.001217	0.036054	0.036037	0.943	0.000351	0.011437	0.011547	0.946
	$3\tau/4$	0.200	0.001164	0.034218	0.034637	0.948	−0.000204	0.010895	0.010746	0.956
	τ	0.176	0.003532	0.041550	0.047835	0.915	0.000299	0.016351	0.019086	0.948
WKM	$\tau/4$	0.269	0.001832	0.044713	0.045359	0.942	0.000131	0.014153	0.014197	0.946
	$\tau/2$	0.231	0.001283	0.035999	0.035858	0.947	0.000368	0.011408	0.011509	0.947
	$3\tau/4$	0.200	0.001132	0.034004	0.034272	0.950	−0.000193	0.010838	0.010716	0.956
	τ	0.176	0.002628	0.039647	0.045615	0.906	0.000116	0.015851	0.017720	0.947
COX	$\tau/4$	0.269	0.000957	0.036029	0.035611	0.955	0.000107	0.011401	0.011229	0.955
	$\tau/2$	0.231	0.001067	0.031741	0.031499	0.954	0.000129	0.010031	0.009949	0.953
	$3\tau/4$	0.200	0.000972	0.028300	0.028071	0.962	0.000060	0.008937	0.008899	0.949
	τ	0.176	0.002177	0.026818	0.027274	0.955	0.000168	0.008790	0.008771	0.956
PCH	$\tau/4$	0.269	0.003717	0.035027	0.035896	0.940	0.002630	0.011076	0.011300	0.939
	$\tau/2$	0.231	0.002926	0.030819	0.031734	0.945	0.001853	0.009736	0.009995	0.936
	$3\tau/4$	0.200	0.002124	0.027440	0.028260	0.949	0.001247	0.008666	0.008949	0.940
	τ	0.176	0.001883	0.025457	0.025679	0.958	0.000621	0.008014	0.008240	0.946
Simpler	–	0.333	0.000814	0.041900	0.041749	0.952	0.000050	0.013257	0.013257	0.947

KM: nonparametric approach based on Kaplan-Meier estimation for $S(t)$; WKM: nonparametric approach based on weighted Kaplan-Meier estimation for $S(t)$; COX: semiparametric approach; PCH: parametric approach using a piecewise constant hazards model; Simpler: simpler approach based on proportion of exposed subjects; Bias: sampling mean of the difference between $\hat{A}(t)$ and $A(t)$; SEE: sampling mean of standard error estimate of $A(t)$; SSD: sampling standard deviation of $\hat{A}(t)$; CP: coverage probability of the 95% Wald confidence interval.

Table 3 Simulation results for the estimation of attributable risk $A(.)$ under proportional hazards, increasing baseline hazard ($\gamma = 4/3$) with regression parameter $\beta = \ln(2)$ and probability of exposure $q = 0.5$

Estimation			$n = 1{,}000$				$n = 10{,}000$			
method	Time	$A(t)$	Bias	SEE	SSD	CP	Bias	SEE	SSD	CP
KM	$\tau/4$	0.299	0.000814	0.064311	0.064377	0.947	−0.000024	0.020388	0.020204	0.956
	$\tau/2$	0.250	0.002020	0.043388	0.043169	0.952	0.000210	0.013761	0.013651	0.944
	$3\tau/4$	0.200	0.001174	0.037152	0.037027	0.955	−0.000469	0.011824	0.011798	0.960
	τ	0.153	0.007382	0.043968	0.054032	0.891	0.000554	0.018140	0.021081	0.939
WKM	$\tau/4$	0.299	0.000805	0.064427	0.064296	0.950	−0.000010	0.020380	0.020196	0.954
	$\tau/2$	0.250	0.002055	0.043322	0.042973	0.949	0.000272	0.013722	0.013643	0.947
	$3\tau/4$	0.200	0.001193	0.036838	0.036463	0.962	−0.000410	0.011739	0.011741	0.958
	τ	0.153	0.005596	0.040652	0.048586	0.898	0.000055	0.017280	0.019095	0.935
COX	$\tau/4$	0.299	0.001207	0.041863	0.040891	0.960	−0.000209	0.013250	0.013076	0.962
	$\tau/2$	0.250	0.001321	0.036377	0.035580	0.954	−0.000062	0.011499	0.011341	0.958
	$3\tau/4$	0.200	0.001300	0.030350	0.029672	0.956	−0.000121	0.009572	0.009502	0.965
	τ	0.153	0.002791	0.027165	0.028199	0.945	−0.000309	0.009206	0.010402	0.945
PCH	$\tau/4$	0.299	−0.000084	0.041594	0.040674	0.961	−0.001831	0.013151	0.013022	0.957
	$\tau/2$	0.250	0.000876	0.036176	0.035464	0.956	−0.000759	0.011424	0.011313	0.958
	$3\tau/4$	0.200	0.001462	0.030163	0.029655	0.959	−0.000051	0.009509	0.009485	0.961
	τ	0.153	0.002572	0.025716	0.024704	0.961	0.000622	0.008058	0.007962	0.945
Simpler	–	0.333	0.001129	0.044983	0.044481	0.955	−0.000242	0.014226	0.014195	0.957

KM: nonparametric approach based on Kaplan-Meier estimation for $S(t)$; WKM: nonparametric approach based on weighted Kaplan-Meier estimation for $S(t)$; COX: semiparametric approach; PCH: parametric approach using a piecewise constant hazards model; Simpler: simpler approach based on proportion of exposed subjects; Bias: sampling mean of the difference between $\hat{A}(t)$ and $A(t)$; SEE: sampling mean of standard error estimate of $A(t)$; SSD: sampling standard deviation of $\hat{A}(t)$; CP: coverage probability of the 95% Wald confidence interval.

Table 4 Simulation results for the estimation of attributable risk $A(.)$ under nonproportional hazards with regression parameter $\beta = \ln(2)$ and probability of exposure $q = 0.5$

Estimation			$n = 1{,}000$				$n = 10{,}000$			
method	Time	$A(t)$	Bias	SEE	SSD	CP	Bias	SEE	SSD	CP
KM	$\tau/4$	0.181	0.001124	0.045053	0.045787	0.954	0.000289	0.014277	0.014126	0.949
	$\tau/2$	0.133	0.001330	0.037581	0.037647	0.953	−0.000029	0.011915	0.012154	0.935
	$3\tau/4$	0.109	0.001211	0.036543	0.036593	0.953	−0.000301	0.011618	0.011608	0.952
	τ	0.093	0.002743	0.043713	0.051764	0.933	−0.000888	0.016362	0.019957	0.950
WKM	$\tau/4$	0.181	0.001138	0.045090	0.045739	0.954	0.000291	0.014274	0.014130	0.949
	$\tau/2$	0.133	0.001347	0.037587	0.037593	0.956	−0.000024	0.011911	0.012151	0.938
	$3\tau/4$	0.109	0.001165	0.036511	0.036518	0.952	−0.000293	0.011612	0.011607	0.956
	τ	0.093	0.001685	0.042617	0.049261	0.920	−0.000708	0.016157	0.019107	0.946
COX	$\tau/4$	0.181	−0.018761	0.037521	0.037543	0.933	−0.019843	0.011869	0.011939	0.621
	$\tau/2$	0.133	0.010548	0.033500	0.033580	0.941	0.009504	0.010588	0.010676	0.847
	$3\tau/4$	0.109	0.023376	0.030960	0.031017	0.879	0.022314	0.009775	0.009879	0.368
	τ	0.093	0.030360	0.029427	0.029588	0.830	0.029168	0.009323	0.009456	0.127
PCH	$\tau/4$	0.181	0.026479	0.048525	0.049191	0.908	−0.017516	0.011688	0.012080	0.672
	$\tau/2$	0.133	0.057418	0.044915	0.045594	0.738	0.011082	0.010391	0.010768	0.806
	$3\tau/4$	0.109	0.070045	0.042342	0.043042	0.607	0.023478	0.009571	0.009936	0.313
	τ	0.093	0.075924	0.040403	0.041050	0.525	0.029848	0.009011	0.009360	0.098

KM: nonparametric approach based on Kaplan-Meier estimation for $S(t)$; WKM: nonparametric approach based on weighted Kaplan-Meier estimation for $S(t)$; COX: semiparametric approach; PCH: parametric approach using a piecewise constant hazards model; Bias: sampling mean of the difference between $\hat{A}(t)$ and $A(t)$; SEE: sampling mean of standard error estimate of $A(t)$; SSD: sampling standard deviation of $\hat{A}(t)$; CP: coverage probability of the 95% Wald confidence interval.

Figure

Click here to download Figure
Figure_AttributableRisk_Gassama.pdf

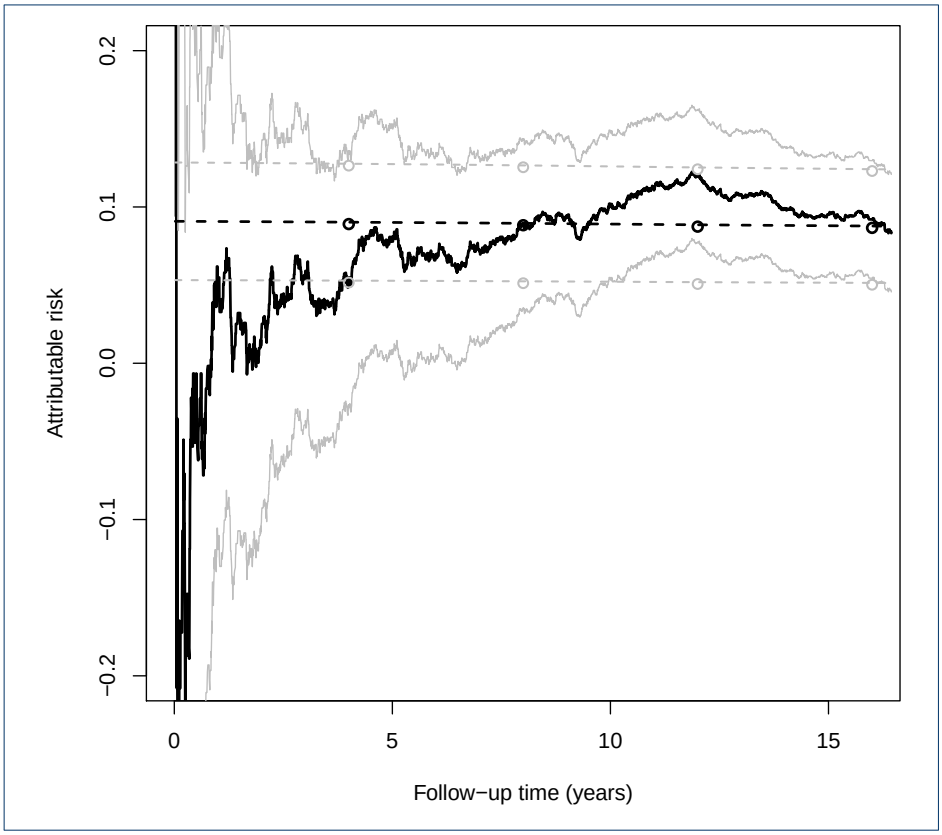


Table A1 Simulation results for the estimation of attributable risk $A(.)$ under proportional hazards, constant baseline hazard ($\gamma = 1$) with regression parameter $\beta = \ln(2)$ and probability of exposure $q = 0.25$

Estimation method	Time	$A(t)$	$n = 1{,}000$				$n = 10{,}000$			
			Bias	SEE	SSD	CP	Bias	SEE	SSD	CP
KM	$\tau/4$	0.166	0.000767	0.037566	0.037107	0.957	−0.000344	0.011889	0.011883	0.954
	$\tau/2$	0.136	0.000318	0.026562	0.026090	0.954	−0.000062	0.008412	0.008505	0.948
	$3\tau/4$	0.111	0.000345	0.022928	0.022308	0.958	−0.000162	0.007283	0.007317	0.951
	τ	0.090	0.000608	0.027222	0.033360	0.921	0.000311	0.010786	0.012074	0.951
WKM	$\tau/4$	0.166	0.000772	0.037668	0.036910	0.959	−0.000311	0.011871	0.011871	0.954
	$\tau/2$	0.136	0.000310	0.026405	0.025910	0.961	−0.000018	0.008350	0.008425	0.947
	$3\tau/4$	0.111	0.000490	0.022436	0.022317	0.954	−0.000149	0.007161	0.007210	0.955
	τ	0.090	0.000011	0.023394	0.027320	0.908	0.000035	0.009801	0.010984	0.926
COX	$\tau/4$	0.166	0.000446	0.028454	0.027789	0.962	−0.000238	0.009000	0.009034	0.950
	$\tau/2$	0.136	0.000327	0.022835	0.022332	0.957	−0.000132	0.007217	0.007226	0.952
	$3\tau/4$	0.111	0.000112	0.018279	0.017625	0.962	−0.000202	0.005767	0.005767	0.953
	τ	0.090	0.000947	0.015938	0.016030	0.947	−0.000031	0.005241	0.005310	0.950
PCH	$\tau/4$	0.166	0.000672	0.027455	0.027806	0.956	−0.000210	0.008677	0.009038	0.938
	$\tau/2$	0.136	0.000474	0.021919	0.022349	0.950	−0.000116	0.006924	0.007227	0.940
	$3\tau/4$	0.111	0.000145	0.017434	0.017630	0.953	−0.000208	0.005502	0.005761	0.944
	τ	0.090	0.000344	0.014600	0.014513	0.952	−0.000189	0.004577	0.004667	0.949
Simpler	–	0.200	0.000623	0.033982	0.034253	0.955	−0.000295	0.010748	0.011128	0.939

Table A2 Simulation results for the estimation of attributable risk $A(.)$ under proportional hazards, constant baseline hazard ($\gamma = 1$) with regression parameter $\beta = \ln(2)$ and probability of exposure $q = 0.75$

Estimation method	Time	$A(t)$	$n = 1{,}000$				$n = 10{,}000$			
			Bias	SEE	SSD	CP	Bias	SEE	SSD	CP
KM	$\tau/4$	0.373	−0.002134	0.075473	0.075167	0.956	0.000578	0.023935	0.024026	0.953
	$\tau/2$	0.321	−0.001817	0.058878	0.058752	0.944	0.000143	0.018721	0.018706	0.960
	$3\tau/4$	0.273	−0.001380	0.055677	0.058463	0.933	0.000299	0.017818	0.017566	0.950
	τ	0.229	0.002988	0.064332	0.079263	0.880	−0.000864	0.026996	0.033048	0.923
WKM	$\tau/4$	0.373	−0.002145	0.075537	0.075176	0.956	0.000561	0.023933	0.024012	0.953
	$\tau/2$	0.321	−0.001797	0.058865	0.058784	0.943	0.000131	0.018709	0.018715	0.959
	$3\tau/4$	0.273	−0.001410	0.055508	0.058293	0.933	0.000328	0.017795	0.017551	0.948
	τ	0.229	0.002194	0.063226	0.077751	0.879	−0.000705	0.026799	0.032880	0.919
COX	$\tau/4$	0.373	−0.002641	0.054982	0.055028	0.948	0.000035	0.017371	0.017306	0.955
	$\tau/2$	0.321	−0.001861	0.050909	0.050975	0.947	0.000099	0.016102	0.016095	0.952
	$3\tau/4$	0.273	−0.001044	0.046584	0.046373	0.952	0.000180	0.014736	0.014743	0.949
	τ	0.229	0.000817	0.043942	0.045675	0.944	−0.000077	0.014623	0.015048	0.951
PCH	$\tau/4$	0.373	−0.002187	0.054875	0.055066	0.949	0.000104	0.017318	0.017314	0.952
	$\tau/2$	0.321	−0.001530	0.050815	0.050969	0.946	0.000130	0.016052	0.016088	0.951
	$3\tau/4$	0.273	−0.000930	0.046465	0.046375	0.951	0.000192	0.014684	0.014709	0.948
	τ	0.229	−0.000619	0.042921	0.043412	0.944	0.000102	0.013576	0.013704	0.948
Simpler	–	0.429	−0.003536	0.058323	0.058297	0.951	−0.000106	0.018358	0.018324	0.952

Table A3 Simulation results for the estimation of attributable risk $A(.)$ under proportional hazards, constant baseline hazard ($\gamma = 1$) with regression parameter $\beta = 0$ and probability of exposure $q = 0.5$

Estimation			$n = 1{,}000$				$n = 10{,}000$			
method	Time	$A(t)$	Bias	SEE	SSD	CP	Bias	SEE	SSD	CP
KM	$\tau/4$	0.000	0.001588	0.066720	0.066661	0.951	0.000298	0.021095	0.020829	0.959
	$\tau/2$	0.000	0.001831	0.049184	0.049848	0.945	0.000730	0.015565	0.015568	0.940
	$3\tau/4$	0.000	0.000775	0.044459	0.044906	0.948	−0.000044	0.014117	0.014006	0.950
	τ	0.000	0.003632	0.054707	0.063368	0.921	0.001753	0.021781	0.024252	0.955
WKM	$\tau/4$	0.000	0.001584	0.066767	0.066675	0.951	0.000300	0.021097	0.020840	0.959
	$\tau/2$	0.000	0.001849	0.049203	0.049846	0.946	0.000731	0.015568	0.015582	0.940
	$3\tau/4$	0.000	0.000670	0.044428	0.044884	0.948	−0.000039	0.014115	0.014006	0.951
	τ	0.000	0.002958	0.052653	0.060895	0.905	0.001605	0.021354	0.024770	0.948
COX	$\tau/4$	0.000	0.001693	0.047709	0.047293	0.961	0.000334	0.015079	0.014993	0.958
	$\tau/2$	0.000	0.001521	0.042265	0.041903	0.963	0.000294	0.013349	0.013276	0.958
	$3\tau/4$	0.000	0.001314	0.037280	0.036941	0.963	0.000261	0.011758	0.011698	0.958
	τ	0.000	0.001276	0.032920	0.032690	0.962	0.000222	0.010319	0.010275	0.958
PCH	$\tau/4$	0.000	0.001755	0.047745	0.047311	0.961	0.000335	0.015080	0.014992	0.959
	$\tau/2$	0.000	0.001581	0.042294	0.041918	0.962	0.000295	0.013349	0.013274	0.959
	$3\tau/4$	0.000	0.001378	0.037290	0.036950	0.962	0.000262	0.011757	0.011694	0.959
	τ	0.000	0.001248	0.032792	0.032573	0.963	0.000221	0.010312	0.010254	0.959
Simpler	–	0.000	0.043998	0.053626	0.033774	0.961	0.013205	0.016996	0.009928	0.957

Annexe E

Article soumis dans *Biometrical Journal*

Use of the prevented fraction to estimate the proportion of stroke cases that could be avoided by using lipid lowering drugs in the French Three-City cohort

Journal:	*Biometrical Journal*
Manuscript ID	bimj.201600261
Wiley - Manuscript type:	Research Paper
Date Submitted by the Author:	15-Dec-2016
Complete List of Authors:	Gassama, Malamine; Biostatistics, Biomathematics, Pharmacoepidemiology and Infectious Diseases (B2PHI), INSERM, UVSQ, Institut Pasteur, Université Paris-Saclay, Thiébaut, Anne; INSERM, UMR 1181 B2PHI; INSERM, UMR1181 PhEMI Tzourio, Christophe; Univ. Bordeaux, Inserm, Population Health Research Center, UMR1219 Benichou, Jacques; Univ. Rouen, Rouen University Hospital, Inserm, UMR1219, Department of Biostatistics
Keywords:	Cox model, Piecewise constant hazards model, Prevented fraction, Statins, Stroke

Biometrical Journal **00** (2016) , ZZZZZZ / **DOI** 10.1002/bimj.201600000

Use of the prevented fraction to estimate the proportion of stroke cases that could be avoided by using lipid lowering drugs in the French Three-City cohort

Running title: Use of the prevented fraction in cohort studies

Malamine Gassama[1], Anne C.M. Thiébaut[1,*], Christophe Tzourio[2], and Jacques Bénichou[3]

[1] Biostatistics, Biomathematics, Pharmacoepidemiology and Infectious Diseases (B2PHI), Inserm, UVSQ, Institut Pasteur, Université Paris-Saclay, 25 rue du Dr Roux, 75724 Paris cedex 15, France.

[2] Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, team HEALTHY, UMR1219, 146 rue Léo Saignat, 33076 Bordeaux cedex, France

[3] Inserm, U1219, University of Rouen, Department of Biostatistics, Rouen University Hospital, 1 rue de Germont, 76031 Rouen cedex, France.

Correspondence to:
Anne Thiébaut
Institut Pasteur, PhEMI
Bâtiment Laveran
25 rue du Docteur Roux
75724 Paris cedex 15, France

The prevented fraction (PF) measures the proportion of disease cases that could be avoided in the presence of a protective exposure in the population. To our knowledge, it has never been considered for survival data in cohort studies. We estimated the PF of stroke for lipid-lowering drugs (LLDs) as a function of time from the Three-City cohort study. We derived PF estimates and associated 95% confidence intervals (95%CIs) from the corresponding attributable risk estimated at four equally distributed time points, using a semiparametric method based on Cox's model and a parametric method based on the piecewise constant hazards model. We also estimated the PF while taking into account coronary heart disease and death from other causes as competing risks using the parametric method. The estimated proportion of stroke cases avoidable by LLD use decreased with follow-up time from 9.99% (95%CI, 4.31% to 15.33%) after 2.5 years to 9.77% (95%CI, 4.16% to 15.05%) after 10 years using the semiparametric method. Similar values were obtained using the parametric method. Accounting for coronary heart disease and death from other causes as competing risks, estimates of avoidable stroke cases were virtually unchanged. The PF can thus be estimated to evaluate the impact of beneficial drugs in observational cohort studies while taking into account confounding factors and competing risks.

Keywords: Cox model; Piecewise constant hazards model; Prevented fraction; Statins; Stroke

[*]Corresponding author: e-mail: anne.thiebaut@inserm.fr, Phone: +33 (0)1 40 61 39 81, Fax: +33 (0)1 45 68 82 04

1 Introduction

In public health, an important goal is to estimate the impact at the population level of an association between an exposure and a disease event. When the exposure is associated with increased disease risk, the attributable risk (AR) can be used to estimate the proportion of cases that can be attributed to an exposure in the population. It was introduced by Levin (1953) and it is a function of the probability of disease, $\mathrm{P}(D)$, in the population, which includes exposed and unexposed subjects, and the hypothetical probability of disease in the same population in the absence of exposure, $\mathrm{P}(D|\bar{E})$. When the exposure is associated with reduced risk, the prevented fraction (PF) can be used instead. It was introduced by Miettinen (1974) and measures the proportion of disease cases that could be avoided in the presence of a protective exposure in the population:

$$PF = \frac{\mathrm{P}(D|\bar{E}) - \mathrm{P}(D)}{\mathrm{P}(D|\bar{E})}. \qquad (1)$$

The PF and the AR are related through (Walter, 1976):

$$1 - PF = \frac{1}{1 - AR}. \qquad (2)$$

Recent developments for estimating the AR from cohort studies with censored time-to-event data that allow for its time-varying nature have been proposed (Chen et al., 2006; Samuelsen and Eide, 2008; Cox et al., 2009; Chen et al., 2010; Laaksonen et al., 2010; Sjölander and Vansteelandt, 2014). Several programs are now publicly available which estimate the AR and associated variance as functions of time (Chen et al., 2010; Laaksonen et al., 2010; Dahlqwist et al., 2016). To our knowledge, however, the definition and estimation of the PF have never been considered for cohort studies in the survival analysis context. In this work, we estimated the AR and deduced the PF using equation (2). For the purpose of illustration, we used recently published cohort data on treatment with lipid lowering drugs (LLDs) and stroke incidence from the Three-City (3C) cohort study (Alpérovitch et al., 2015) to estimate the proportion of stroke cases that could be avoided using LLDs with several approaches.

2 Population and Methods

2.1 Study population

The 3C study is a longitudinal cohort study that aims at evaluating the relation between vascular diseases and dementia in persons aged 65 years and older. The details of the protocol were published elsewhere (3C Study Group, 2003). The 3C study received ethical approval from the University Hospital of Kremlin-Bicêtre, France, and all participants signed an informed consent form. From the 9,294 participants who were randomly selected from the electoral rolls of three large French cities (Bordeaux, Dijon and Montpellier) between March 1999 and March 2001, we selected 7,484 participants using the same inclusion criteria as in the original analysis (Alpérovitch et al., 2015). Among them, 2,048 (27.4%) were using LLDs (statins or fibrates only) at baseline (inclusion in the cohort) and are thereafter considered to be exposed. We focused here on the occurrence of stroke as a first ever vascular event (261 incident cases for a maximum length of follow-up of 10 years) and considered 1,490 competing events (due to coronary heart disease or death from other causes). Our study was motivated by recently published findings of a statistically significant association of stroke with LLD use (Alpérovitch et al., 2015).

Biometrical Journal **00** (2016) 00

2.2 Statistical analysis

Following previous work on the AR (Cox et al., 2009; Chen et al., 2010; Laaksonen et al., 2010; Sjölander and Vansteelandt, 2014), disease probabilities in equation (1) can be interpreted as cumulative distribution functions so that the PF can be defined in terms of the marginal survival function in the population, $S(t)$, and the survival function assuming all subjects had been unexposed, $S_0(t)$, at time $t > 0$:

$$PF(t) = \frac{S(t) - S_0(t)}{1 - S_0(t)}. \qquad (3)$$

We considered two approaches: a semiparametric method based on Cox's proportional hazards model (Chen et al., 2010; Dahlqwist et al., 2016) and a fully parametric method based on a piecewise constant hazards model (Laaksonen et al., 2010). For the latter approach, we partitioned the follow-up time into four intervals of 2.5-year width each. Moreover, because LLD exposure and other risk factors of stroke could also be related to coronary events and death from other causes, we considered an extension of the parametric approach based on cause-specific hazards to take into account coronary heart disease and death from other causes as competing risks (Laaksonen et al., 2010).

To mimic the published analysis (Alpérovitch et al., 2015), we considered two models for covariate adjustment: a simple model adjusted for sex, study center (Bordeaux, Dijon, Montpellier) and age (in deciles), with time-on-study as the time scale (model 1); and a more complex model (model 2) with adjustment for the same variables as in model 1 and for the following additional potential confounding factors: diabetes (yes, no), body mass index (<25, 25-29, ≥30 kg/m^2), smoking status (never, past, current smoker), alcohol consumption (never, past, current drinker), hypertension (yes, no), arrhythmia (yes, no), use of antithrombotic drugs (yes, no), triglyceride concentration (in tertiles), and low density lipoprotein to high density lipoprotein ratio (in tertiles).

We first evaluated the association between LLD use and stroke risk. For this, we used the nonparametric Kaplan-Meier method to estimate the proportion of subjects free of stroke over time, separately among exposed and unexposed subjects at baseline and used the log rank test to compare these two survival distributions. We estimated adjusted hazard ratios (HRs) of developing stroke for individuals exposed to LLDs relative to those unexposed and corresponding 95% confidence intervals (95%CIs) using both semiparametric and parametric approaches. In all semiparametric and parametric regression models, we tested the proportional hazards assumption by examining interactions between exposure and follow-up time divided into four intervals.

Second, we estimated the PF of stroke associated with baseline LLD use over follow-up time. We obtained AR estimates and associated variances as functions of time t, then derived $PF(t)$ as $-AR(t) / [1 - AR(t)]$ from equation (2) and applied the delta method to estimate the variance of PF as (Gargiullo et al., 1995):

$$\widehat{\mathrm{Var}}\left[\widehat{PF}(t)\right] = \frac{\widehat{\mathrm{Var}}[\widehat{AR}(t)]}{\left\{1 - \widehat{AR}(t)\right\}^4}.$$

To construct confidence intervals for $PF(t)$, we used the complementary logarithmic transformation $\ln\{1 - PF(t)\} = -\ln\{1 - AR(t)\}$ which is consistent with recommendations for the AR (Chen et al., 2010; Laaksonen et al., 2010). Results are presented at four equally distributed time points (after 2.5, 5, 7.5 and 10 years of follow-up). Overall, 68, 149, 213 and 261 stroke cases were diagnosed after 2.5, 5, 7.5 and 10 years of follow-up respectively in the

3C cohort study. For the semiparametric approach, estimates were obtained at times actually observed in the dataset so we considered values taken at the closest preceding time point.

All calculations were performed using R (version 3.2.2, R Core Team, Vienna, Austria) and SAS (version 9.3, SAS Institute Inc., Cary, NC, USA). For the semiparametric approach, we used the R package `paf` developed by Chen (2014). Results are not presented for the R package `AF` of Dahlqwist and Sjölander (2016) because point estimates were virtually identical and variances close to those obtained with Chen's `paf` package. For the parametric approach, we used a set of macros developed by Laaksonen et al. (2011).

3 Results

Figure 1 shows that subjects ever exposed to LLDs had a statistically significantly lower risk of stroke compared to those never exposed (log rank test, $p = 0.004$).

From model 1, HRs of subjects ever exposed compared to those never exposed to LLDs at baseline estimated using the Cox proportional hazards model (HR, 0.682; 95%CI, 0.501 to 0.928) and a piecewise constant hazards model (HR, 0.693; 95%CI, 0.509 to 0.943) were close numerically and both statistically significantly less than 1 (Table 1). Adjustment for more covariates in model 2 resulted in somewhat stronger associations between LLD exposure and stroke occurrence (HR, 0.630; 95%CI, 0.457 to 0.868 and HR, 0.640; 95%CI, 0.464 to 0.883, respectively). The proportional hazards assumption seemed appropriate for the Cox and the parametric models 1 and 2 (likelihood ratio test, $p > 0.8$ in all models).

When using model 1, with the semiparametric approach, the proportion of stroke cases avoidable thanks to LLD use slightly decreased with follow-up time from 8.18% (95%CI, 2.56% to 13.47%) after 2.5 years to 8.03% (95%CI, 2.47% to 13.27%) after 10 years (Table 1). For the fully parametric approach, the PF decreased with follow-up time from 7.89% (95%CI, 2.20% to 13.26%) after 2.5 years to 7.77% (95%CI, 2.13% to 13.08%) after 10 years (Table 1). With model 2, PF estimates were higher, in line with the stronger association between LLD use and stroke risk observed with more complete adjustment (Table 1), and also decreased with follow-up time from 9.99% (95%CI, 4.31% to 15.33%) after 2.5 years to 9.77% (95%CI, 4.16% to 15.05%) after 10 years using the semiparametric approach and from 9.67% (95%CI, 3.84% to 15.15%) to 9.49% (95%CI, 3.72% to 14.91%) with the parametric approach.

Finally, when considering coronary heart disease and death from other causes as competing risks with the parametric approach, estimates of stroke PFs were very close to those obtained ignoring censoring due to competing risks for both adjustment models (Table 1).

4 Discussion

Following a former report of a statistically significant one third reduction in stroke risk in users of LLDs compared to nonusers at baseline in the 3C population-based cohort of older people (Alpérovitch et al., 2015), we estimated that up to 10% of stroke cases could be prevented by using LLDs. This proportion is far from negligible owing to the incidence and severity of stroke in the elderly population.

To our knowledge, this is the first study proposing PF estimation in the context of cohort studies and censored time-to-event outcomes. Our study complements the original study where estimated HRs of stroke in LLD users compared to nonusers were reported

Biometrical Journal **00** (2016) 00

(Alpérovitch et al., 2015), by yielding PF estimates that take into account those HR estimates, as well as the initial 27.4% prevalence of exposure to LLDs. So far, measures of PF considered for cohort studies have assumed equal follow-up times for all subjects, thus ignoring censoring, loss of follow-up and competing risks (Walter et al., 2007). A major limitation of those measures in the context of cohort studies is that they only take account of the proportion of exposed subjects at the beginning of follow-up. The proportion of exposed subjects indeed increases as follow-up time increases because nonexposed subjects fail (*i.e.*, experience the disease or event of interest) earlier than exposed subjects in case of a beneficial exposure (Cox et al., 2009). Taking into account when risk is assessed is of importance in a longitudinal setting because of these variations of the prevalence of the risk factor during follow-up.

As for the AR functions in other examples with detrimental exposures and constant HR > 1, the PF also decreased with follow-up time in the presence of a protective exposure. In our example data, however, we observed limited variations of the PF estimates whether using the semiparametric or fully parametric approaches. More variation could be expected with disease outcomes with higher incidence rates.

In this work, we used two methods to estimate the PF as a function of time when disease probabilities are interpreted as cumulative distribution functions, that is when the PF can be expressed in terms of survival functions. The first approach was a semiparametric method based on Cox's proportional hazards model, the second approach was a parametric method based on the piecewise constant hazards (exponential) model. Nonparametric methods based on Kaplan-Meier's estimates of survival functions have also been suggested to estimate the AR (Chen et al., 2010) and could be used to estimate the PF. However, such approaches would be limited by the nature (categorical only) and number of covariates that could be taken into consideration. Hence, we feel these methods not suited for an application to the 3C study or studies of similar type with many potential confounders. An alternative definition of the PF as a function of time could be obtained upon interpreting disease probabilities in equation (1) as instantaneous hazard functions as has been considered for the AR (Chen et al., 2006; Samuelsen and Eide, 2008), but this definition for AR conceptually differs from the usual, not time-dependent measure and has received less attention in the literature.

In this study, we derived PF variances by applying the delta method to the estimated AR and associated variance which were obtained from available SAS and R programs. Alternatively, these programs can be adapted to calculate the PF directly from equation (3) as well as associated variance following the original approaches developed for the AR. For the semiparametric approach, Chen et al. (2010) relied on asymptotic theory while Sjölander and Vansteelandt (2011) combined a sandwich formula and the delta method. For the parametric approach, Laaksonen et al. (2010) obtained variance estimates using the delta method. These approaches yielded very close PF estimates and CIs to those derived from the AR in our data (data not shown).

Several authors (Chen et al., 2010; Laaksonen et al., 2010) have noted that the normal approximation for the sampling distribution of estimated AR may not be accurate in case of small sample sizes and recommended using the complementary logarithmic transformation $\ln\{1 - AR(t)\}$. The same could be true for PF estimates. For this reason, we considered the complementary logarithmic transformation $\ln\{1 - PF(t)\} = -\ln\{1 - AR(t)\}$.

As observed in our results, when using coronary heart disease and death of other causes as competing risks for the parametric approach, the PF estimates were close to those obtained when ignoring competing risks. We observed no statistically significant association between LLD use and competing events in our application (HR, 0.988; 95%CI, 0.878 to 1.113 and HR, 0.943; 95%CI, 0.834 to 1.067 when using models 1 and 2 respectively with the parametric

approach). The outcome of interest in our study was stroke and ignoring competing risks due to coronary heart disease and death from other causes means that the estimates of PF obtained only apply under the assumption that no one dies or develops coronary heart disease as a first cardiovascular event during the follow-up (Laaksonen, et al., 2010). The bias caused by ignoring censoring due to coronary heart disease and death from other causes was small in our application but could become larger if the association between LLD use and competing events were stronger and follow-up were longer (Laaksonen et al., 2010).

As for AR estimates, the interpretation of the PF as the proportion of cases that could be avoided thanks to exposure to the beneficial factor requires three conditions: unbiased estimates, causal relationship between exposure and disease, and an unchanged distribution of other risk factors as a result of introducing exposure (Walter, 1976). In this study, we found a relatively small PF of stroke associated with LLD use, which could be underestimated relative to subjects naïve to statin or fibrate treatment. Indeed, elderly subjects in the 3C cohort tended to receive multiple treatments while having a healthier lifestyle, which could contribute to reduce their vascular risk (Alpérovitch et al., 2015). Moreover, the prescription of statins and fibrates in the general population is not intended for the primary prevention of stroke specifically but has been demonstrated in clinical trials to reduce the incidence of cardiovascular and cerebrovascular events in general in people with a history of cardiovascular disease (Alpérovitch et al., 2015). Therefore, the PF estimate in our study should be interpreted with caution.

This study has some limitations. First, in their analysis, Alpérovitch et al. (2015) used age as the timescale in Cox regression models to estimate HRs from the 3C cohort data. They found HR estimates of 0.71 (95%CI, 0.53 to 0.95) and 0.66 (95%CI, 0.49 to 0.90) from the simpler and fully-adjusted models, respectively. Presently, available packages for estimating AR as a function of time do not allow left truncation resulting from using age as the timescale. For this reason, in our study, we used time-on-study as the timescale and parametrically adjusted our models for age instead, finding HR estimates close to the original publication based on the same data (0.68 and 0.63 with Cox's models 1 and 2, respectively). Second, we did not consider competing risks for the semiparametric approach. Indeed, the authors who proposed this approach did not address this topic (Chen et al., 2010; Sjölander and Vansteelandt, 2014) but it would be interesting to further develop the semiparametric approach in order to be able to account for competing risks with it as well. Finally, we cannot exclude that our PF estimate may be distorted by biases inherent to observational epidemiological studies of drug effects, in particular indication bias. However, these biases have been ruled out as the sole explanation for the statistically significantly lower risk of stroke in LLD users compared to nonusers in the original study (Alpérovitch et al., 2015) and we believe that investigating these biases was beyond the scope of the present study.

In conclusion, our study shows that the PF can be estimated to evaluate the potential beneficial impact of drugs in observational cohort studies while taking account of potential confounding factors and competing risks. The use of LLDs was associated with approximately 10% fewer cases of stroke in the 3C cohort elderly population.

Biometrical Journal **00** (2016) 00

Acknowledgements

This research used data from the Three-City (3C) study which is conducted under an agreement between “Institut National de la Santé et de la Recherche Médicale” (Inserm) and “Université Victor Segalen-Bordeaux 2”. The authors wish to thank Catherine Helmer for kindly sharing the 3C data. The 3C study supports are listed on the Study website (http://www.three-city-study.com/). This research was supported by the French Medicines Agency (“Agence Nationale de Sécurité du Médicament et des produits de santé”, ANSM). Malamine Gassama was a recipient of a PhD grant from the French Ministry of Research.

Conflict of Interest

The authors have declared no conflict of interest.

References

3C Study Group (2003). Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population. *Neuroepidemiology* **22**, 316–325.

Alpérovitch, A., Kurth, T., Bertrand, M., Ancelin, M.-L., Helmer, C., Debette, S., and Tzourio, C. (2015). Primary prevention with lipid lowering drugs and long term risk of vascular events in older people: population based cohort study. *BMJ* **350**, h2335.

Chen, L. (2014). "`paf`": Attributable fraction function for censored survival data, R package version 1.0 https://cran.r-project.org/web/packages/paf/index.html.

Chen, L., Lin, D., and Zeng, D. (2010). Attributable fraction functions for censored event times. *Biometrika* **97**, 713–726.

Chen, Y., Hu, C., and Wang, Y. (2006). Attributable risk function in the proportional hazards model for censored time-to-event. *Biostatistics* **7**, 515–529.

Cox, C., Chu, H., and Muñoz, A. (2009). Survival attributable to an exposure. *Statistics in Medicine* **28**, 3276–3293.

Dahlqwist, E., and Sjölander, A. (2016). "`AF`": Model-based estimation of confounder-adjusted attributable fractions, R package version 0.1.2 https://cran.r-project.org/web/packages/AF/index.html.

Dahlqwist, E., Zetterqvist, J., Pawitan, Y., and Sjölander, A. (2016). Model-based estimation of the attributable fraction for cross-sectional, case-control and cohort studies using the R package AF. *European Journal of Epidemiology* **31**, 575–582.

Gargiullo, P., Rothenberg, R., and Wilson, H. (1995). Confidence intervals, hypothesis tests, and sample sizes for the prevented fraction in cross-sectional studies. *Statistics in Medicine* **14**, 51–72.

Laaksonen, M., Härkänen, T., Knekt, P., Virtala, E., and Oja, H. (2010). Estimation of population attributable fraction (PAF) for disease occurrence in a cohort study design. *Statistics in Medicine* **29**, 860–874.

Laaksonen, M., Knekt, P., Härkänen, T., Virtala, E., and Oja, H. (2010). Estimation of the population attributable fraction for mortality in a cohort study using a piecewise constant hazards model. *American Journal of Epidemiology* **171**, 837–847.

Laaksonen, M., Virtala, E., Knekt, P., Oja, H., and Härkänen, T. (2011). SAS macros for calculation of population attributable fraction in a cohort study design. *Journal of Statistical Software* **47**, 1–25.

Levin, M. (1953). The occurrence of lung cancer in man. *Acta - Unio Internationalis Contra Cancrum* **9**, 531–41.

Miettinen, O. (1974). Proportion of disease caused or prevented by a given exposure, trait or intervention. *American Journal of Epidemiology* **99**, 325–332.

Samuelsen, S., and Eide, G. (2008). Attributable fractions with survival data. *Statistics in Medicine* **27**, 1447–1467.

Sjölander, A., and Vansteelandt, S. (2011). Doubly robust estimation of attributable fractions. *Biostatistics* **12**, 112–121.

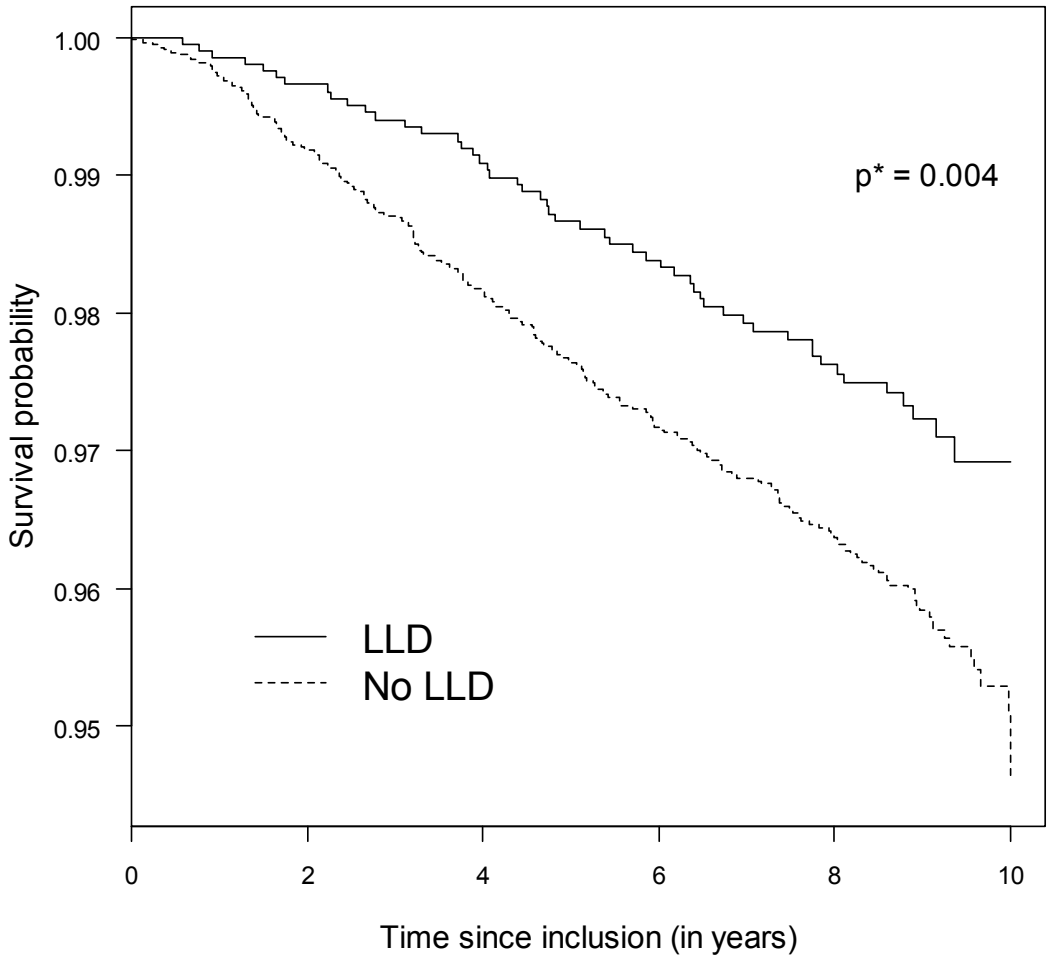
Biometrical Journal **00** (2016) 00

Sjölander, A., and Vansteelandt, S. (2014). Doubly robust estimation of attributable fractions in survival analysis. *Statistical Methods in Medical Research*. (in press)

Walter, S. (1976). The estimation and interpretation of attributable risk in health research. *Biometrics* **1**, 229–243.

Walter, S., Hsieh, C., and Liu, Q. (2007). Effect of exposure misclassification on the mean squared error of population attributable risk and prevented fraction estimates. *Statistics in Medicine* **26**, 4833–4842.

Figure 1: Estimated Kaplan-Meier curves of survival without stroke according to ever or never use of lipid lowering drugs at baseline, 3C study, 1999-2011



LLD: lipid lowering drugs
*p-value of the log rank test

Biometrical Journal **00** (2016) 00

Table 1: Estimation of the prevented fraction of stroke associated with baseline ever use of lipid lowering drugs, at times 2.5, 5, 7.5 and 10 years, 3C study, 1999-2011

Estimation	Model 1			Model 2		
method	HR (95%CI)	PF	95%CI	HR (95%CI)	PF	95%CI
COX	0.682 (0.501 - 0.928)	0.0818	0.0256 - 0.1347	0.630 (0.457 - 0.868)	0.0999	0.0431 - 0.1533
		0.0814	0.0253 - 0.1342		0.0993	0.0426 - 0.1525
		0.0810	0.0251 - 0.1336		0.0987	0.0423 - 0.1518
		0.0803	0.0247 - 0.1327		0.0977	0.0416 - 0.1505
PCH	0.693 (0.509 - 0.943)	0.0789	0.0220 - 0.1326	0.640 (0.464 - 0.883)	0.0967	0.0384 - 0.1515
		0.0785	0.0217 - 0.1320		0.0961	0.0380 - 0.1508
		0.0782	0.0215 - 0.1315		0.0956	0.0377 - 0.1500
		0.0777	0.0213 - 0.1308		0.0949	0.0372 - 0.1491
PCH*	0.693 (0.509 – 0.943)	0.0791	0.0220 - 0.1328	0.640 (0.464 – 0.883)	0.0968	0.0382 - 0.1517
		0.0789	0.0218 - 0.1327		0.0963	0.0376 - 0.1513
		0.0789	0.0216 - 0.1328		0.0958	0.0370 - 0.1511
		0.0789	0.0213 - 0.1330		0.0953	0.0362 - 0.1508

Model 1: adjusted for sex, study center (Bordeaux, Dijon, Montpellier) and age (deciles);
Model 2: adjusted as model 1 plus diabetes (yes, no), body mass index (<25, 25-29, ≥30 kg/m^2), smoking status (never, past, current smoker), alcohol consumption (never, past, current drinker), hypertension (yes, no), cardiac rhythm disorder (yes, no), antithrombotic therapy (yes, no), triglycerides (in tertiles), and low density lipoprotein to high density lipoprotein ratio (in tertiles)

95%CI: 95% confidence interval of prevented fraction; COX: semiparametric approach using Cox's proportional hazards model; HR: adjusted hazard ratio for stroke; PCH: parametric approach using a piecewise constant hazards model; PF: estimated prevented fraction

*Models 1 and 2 with parametric approach and coronary heart disease and death from other causes as competing risks

